

## **UC Merced**

### **Proceedings of the Annual Meeting of the Cognitive Science Society**

#### **Title**

Cognitive Science Society title

#### **Permalink**

<https://escholarship.org/uc/item/2f5769m9>

#### **Journal**

Proceedings of the Annual Meeting of the Cognitive Science Society, 22(22)

#### **Authors**

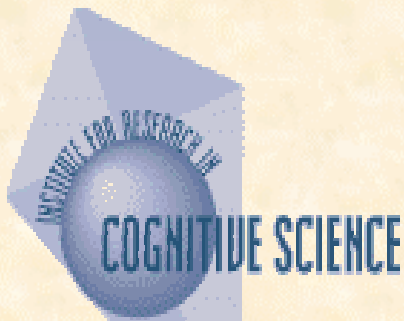
Gleitman, Lila R.

Joshi, Aravind K.

#### **Publication Date**

2000

Peer reviewed



Proceedings of the  
Twenty-Second Annual  
Conference of the  
**Cognitive  
Science  
Society**

August 13 - 15, 2000

Institute for Research in  
Cognitive Science

University of Pennsylvania

Editors: Lila R. Gleitman and Aravind K. Joshi

**Proceedings of the  
Twenty-Second Annual Conference  
of the  
Cognitive Science Society**

Lila R. Gleitman and Aravind K. Joshi  
Editors

August 13-15, 2000  
Institute for Research in Cognitive Science  
University of Pennsylvania  
Philadelphia, PA

## Foreword

This volume contains the final drafts of the papers and posters selected for presentation at the 22<sup>nd</sup> Annual Meeting of the Cognitive Science Society in Philadelphia, PA, August 13-15, 2000. A glance through the Table of Contents shows that this meeting reflects the eclectic and forward-looking character of Cognitive Science as it comes into its own at the turn of the Millennium. “CogSci2000” embodies a set of scholarly contributions that range from psychological and neuroscientific experimentation through computational modeling to philosophical musings about the foundations of the field. In our view, the coherence and range of these contributions characterize the emergence of Cognitive Science from the status of an interdisciplinary field to a full-blown discipline with its own formalisms, methods, techniques, and overarching goals: to find a computational theory of Intelligence.

Having just offered a description of cognitive science that has the smell of a definition, we hasten to back off from that designation: Hardly any scientific discipline, not even speaking of the humanistic ones, succumbs readily to such definition. We offer the phrase as one that suggests and encompasses many of the themes that underlie the specific research questions taken up in this collection and that recur elsewhere in the cognitive science literature.

Coherent as our field seems when described in one grand phrase, it is not easy to subdivide it into proper parts. Categorization is itself one of the most vexed unsolved issues in our field (and thus the topic of several sessions at this meeting). Yet as a matter of practicality, as the hosts of this conference, we were forced to try, if only to cobble together paper and poster sessions that would have maximum organization and minimal overlap, allowing the conference attendees to choose among parallel sessions with a minimum of handwringing and regret. We hope that the attendees will be satisfied that the sessions are productive, lively, informative, and, in general, a lot of fun.

We turn now to acknowledging and thanking the many people, groups, and institutions that made this conference as successful as it was.

The COGNITIVE SCIENCE SOCIETY BOARD, for inviting us to host CogSci2000 and for providing the framework, backup and advice from prior conference organizers.

The PROGRAM COMMITTEE that stewarded the review process, farming out the papers, personally reviewing many of them, nagging the reviewers until they sent their reports, and finally making their recommendations to us. There were 319 submissions requiring about 1000 reviews. 95 paper submissions (out of 248) and 130 posters were accepted. The poster presentations include those paper submissions that were converted to posters based on the recommendations of the program committee.

The VOLUNTEER REVIEWERS from the cogsci community who carried out the difficult task of reviewing quite wonderfully. This allowed us to provide at least two, in most cases three, reviews to each author whether or not the submission was accepted for presentation. Although these reviews were perforce short and sometimes there were partial mismatches between reviewer and reviewee, our impression is that the process – compressed into a few short weeks – worked very well. There were a few complaints, to be sure, so when appropriate we obtained yet more reviews at the last minute. We also received very many gratifying letters from authors both successful and unsuccessful expressing their overall satisfaction with the process.

Our POSTDOCTORAL FELLOWS and ADVANCED GRADUATE STUDENTS who were repeatedly pressed into service for a host of tasks.

TOBY MINTZ and IONE FINE for organizing the Tutorials, BONNIE WEBBER for the solicitation and selection of symposia, and MARK STEEDMAN for assembling the Panel on Education.

Our PROVOST ROBERT BARCHI for his useful opening welcome and remarks.

Our PLENARY SPEAKERS, JAMES ALLEN and RANDY GALLISTEL, for their talks which in addition to the scientific content set the intellectual tone for these meetings.

Our very special thanks go to TRISHA YANNUZZI who in essence “is” CogSci2000 from the beginning to the end in all administrative aspects including the automation of the very efficient and smoothly run submission and review process. We wish we had the wordsmithery to express our gratitude to her, but we must settle for a resounding and heartfelt THANK YOU to TRISHA and to her small, but incredibly efficient staff, ANN BIES, LAUREL SWEENEY, JENNIFER MACDOUGALL, NICOLE BOLDEN and LEE LEIBER (IRCS Technical Staff).

For financial support: NATIONAL INSTITUTES OF HEALTH, MICROSOFT CORPORATION, the COGNITIVE SCIENCE SOCIETY, and the INSTITUTE FOR RESEARCH IN COGNITIVE SCIENCE at the UNIVERSITY OF PENNSYLVANIA.

And, of course, the AUTHORS of paper and poster submissions without whom there would be no CogSci2000. We thank them and applaud them for their efforts in the preparation of these submissions and placing them before their peers, thereby furthering knowledge in the field of Cognitive Science and inspiring its practitioners – certainly including us.

Welcome to CogSci2000!

Lila Gleitman and Aravind Joshi  
Hosts, CogSci2000  
Co-Directors, Institute for Research in Cognitive Science  
University of Pennsylvania

## Table of Contents

Foreword from the Conference Co-Chairs.....	iii
Conference Page.....	xvi
Cognitive Science Society.....	xvii
Tutorial Program.....	xviii
Symposia.....	xix
Special Session on Undergraduate Education in Cognitive Science.....	xx
Reviewers for the Twenty-Second Annual Conference.....	xxi

### Invited Speakers

<i>Spoken Language Systems and Human Communication</i>	
James F. Allen.....	1
<i>The Symbolic Foundations of Conditioned Behavior</i>	
Randy Gallistel.....	1

### Education Panel

<i>Special Session on Undergraduate Education in Cognitive Science</i>	
Mark Steedman.....	2

### Symposia

<i>Perspectives on Conceptual Change</i>	
David Kaufman.....	5
<i>The Role of the Cerebellum in Cognition and Affect</i>	
Natika Newton.....	10
<i>Bayesian Approaches to Cognitive Modeling</i>	
Michael C. Mozer and Joshua Tenenbaum.....	16
<i>The Nature of Human Errors: An Emerging Interdisciplinary Perspective</i>	
Jiajie Zhang.....	17

### Papers

<i>Induction of Causal Chains</i>	
Woo-kyoung Ahn and Martin J. Dennis.....	19
<i>Evaluating the Effectiveness of a Cognitive Tutor for Fundamental Physics Concepts</i>	
Patricia L. Albacete and Kurt A. VanLehn.....	25
<i>Memory in Chains: Modeling Primacy and Recency Effects in Memory for Order</i>	
Erik M. Altmann.....	31
<i>Computational Explorations of the Irrelevant Sound Effect in Serial Short-Term Memory</i>	
C. Philip Beaman.....	37
<i>Sex, Syntax, and Semantics</i>	
Lera Boroditsky and Lauren A. Schmidt.....	42
<i>In Search of the Minority Default: The Case of Arabic Plurals</i>	
Sami Boudelaa and M. Gareth Gaskell.....	48
<i>Representing Categories in Artificial Neural Networks Using Perceptually-Derived Feature Networks</i>	
Robert B. Branstrom.....	54
<i>A Sensorimotor Map of Visual Space</i>	
Bruce Bridgeman.....	60

<i>Unity of Consciousness: What it is and Where it is Found</i>	
Andrew Brook.....	65
<i>Are Retrievals from Long-Term Memory Interruptible?</i>	
Michael D. Byrne.....	71
<i>Hemispheric Specialization During Episodic Memory Encoding in the Human Hippocampus and MTL</i>	
Daniel J. Casasanto, William D. S. Killgore, Guila Glosser, Joseph A. Maldjian, and John A. Detre.....	77
<i>A Connectionist Single-Mechanism Account of Rule-Like Behavior in Infancy</i>	
Morten H. Christiansen, Christopher M. Conway, and Suzanne Curtin.....	83
<i>Committing to an Ontology: A Connectionist Account</i>	
Eliana Colunga and Linda B. Smith.....	89
<i>An Exemplar Model of Classification in Single and Combined Categories</i>	
Fintan Costello.....	95
<i>A Six-Unit Network is All You Need to Discover Happiness</i>	
Matthew N. Dailey, Garrison W. Cottrell, and Ralph Adolphs.....	101
<i>Addressing the Learnability of Verb Subcategorizations with Bayesian Inference</i>	
Mike Dowman.....	107
<i>Hidden Markov Models for Coding Story Recall Data</i>	
Michael A. Durbin, Jason Earwood, and Richard M. Golden.....	113
<i>A Rarity Heuristic for Hypothesis Testing</i>	
Aidan Feeney, Jonathan St. B. T. Evans, and Simon Venn.....	119
<i>Modeling Orientation Effects in Symmetry Detection: The Role of Visual Structure</i>	
Ronald W. Ferguson.....	125
<i>Visual Learning for a Mid Level Pattern Discrimination Task</i>	
I. Fine and Robert A. Jacobs.....	131
<i>Decoding Syntactic Parameters: The Superparser as Oracle</i>	
Janet Dean Fodor and Virginia Teller.....	136
<i>Lexical Contact During Speech Perception: A Connectionist Model</i>	
Eric Forbell and Eric Chown.....	142
<i>The Resemblance of One-Year-Old Infants to their Fathers: Refuting Christenfeld &amp; Hill (1995)</i>	
Robert M. French, Serge Brédart, Johanne Huart, and Christophe Labiouse.....	148
<i>Memory versus Perceptual-Motor Tradeoffs in a Blocks World Task</i>	
Wai-Tat Fu and Wayne D. Gray.....	154
<i>Babies, Variables, and Relational Correlations</i>	
Michael Gasser and Eliana Colunga.....	160
<i>Resource-Adaptive Selection of Strategies in Learning from Worked-Out Examples</i>	
Peter Gerjets, Katharina Scheiter, and Werner H. Tack.....	166
<i>A Neural Network Model of Concept-Influenced Segmentation</i>	
Robert L. Goldstone.....	172
<i>The Determinants of Basic-Level Performance</i>	
Frédéric Gosselin and Philippe G. Schyns.....	178
<i>Latent Semantic Analysis Captures Causal, Goal-Oriented, and Taxonomic Structures</i>	
Arthur Graesser, Ashish Karnavat, Victoria Pomeroy, and Katja Wiemer-Hastings.....	184
<i>Rational Assessments of Covariation and Causality</i>	
Gustaf Gredebäck, Anders Winman, and Peter Juslin.....	190
<i>Function-Follows-Form Transformations in Scientific Problem Solving</i>	
Todd W. Griffith, Nancy J. Nersessian, and Ashok Goel.....	196
<i>Teacakes, Trains, Taxicabs and Toxins: A Bayesian Account of Predicting the Future</i>	
Thomas L. Griffiths and Joshua B. Tenenbaum.....	202

<i>Vagueness in Context</i>	
Steven Gross.....	208
<i>Simulating Causal Models: The Way to Structural Sensitivity</i>	
York Haggmayer and Michael R. Waldmann.....	214
<i>The Problem with Logic in the Logical Problem of Language Acquisition</i>	
Petra Hendriks.....	220
<i>Memory-Based Problem Solving and Schema Induction in Go</i>	
Alex Heneveld, Alan Bundy, Michael Ramscar, and Julian Richardson.....	226
<i>Toward an Integrated Account of Reflexive and Reflective Reasoning</i>	
John E. Hummel and Jesse M. Choplin.....	232
<i>Constituent Structure in Mathematical Expressions</i>	
Anthony R. Jansen, Kim Marriott, and Greg W. Yelland.....	238
<i>Algorithm, Heuristic or Exemplar: Process and Representation in Multiple-Cue Judgment</i>	
Sari Jones, Peter Juslin, Henrik Olsson, and Anders Winman.....	244
<i>Influences on Attribute Selection in Redescriptions: A Corpus Study</i>	
Pamela W. Jordan.....	250
<i>Verb Meanings, Object Affordances, and the Incremental Restriction of Reference</i>	
Edward Kako and John C. Trueswell.....	256
<i>Learning the Use of Discourse Markers in Tutorial Dialogue for an Intelligent Tutoring System</i>	
Jung Hee Kim, Michael Glass, Reva Freedman, and Martha W. Evens.....	262
<i>Are Structural Principles Useful for Automatic Disambiguation?</i>	
Alexandra Kinyon.....	268
<i>Dynamic Extension of Episode Representation in Analogy-Making in AMBR</i>	
Boicho Kokinov and Alexander Petrov.....	274
<i>Controlled Exploration of Alternative Mechanisms in Cognitive Modeling</i>	
Rita Kovordányi.....	280
<i>Modeling Infant Learning via Symbolic Structural Alignment</i>	
Sven E. Kuehne, Dedre Gentner, and Kenneth D. Forbus.....	286
<i>An Optimality-Theoretic Model of Acquisition of Tense and Agreement in French</i>	
Géraldine Legendre, Paul Hagstrom, Marina Todorova, and Anne Vainikka.....	292
<i>Infinite RAAM: A Principled Connectionist Basis for Grammatical Competence</i>	
Simon Levy, Ofer Melnik, and Jordan Pollack.....	298
<i>The Acquisition of Lexical and Grammatical Aspect in a Self-Organizing Feature-Map Model</i>	
Ping Li.....	304
<i>Irregularization: The Interaction of Item Frequency and Phonological Interference in Regular Past Tense Production</i>	
Christopher J. Long and Amit Almor.....	310
<i>A Computational Level Theory of Similarity</i>	
Bradley C. Love.....	316
<i>Viewpoint Dependent Facial Expression Recognition: Japanese Noh Masks and the Human Face</i>	
Michael J. Lyons, Andre Plante, Miyuki Kamachi, Shigeru Akamatsu, Ruth Campbell, and Mike Coleman.....	322
<i>Is Lexical Retrieval in Speech Production like Recall or Recognition? The Effects of Word Frequency and Neighbourhood Size</i>	
Siobhan B. G. MacAndrew, Trevor A. Harley, and Sheila Colgan.....	328
<i>A Generative Connectionist Model of the Development of Rule Use in Children</i>	
Stuart Marcovitch and Philip David Zelazo.....	334
<i>Transfer along a Continuum: Differentiation or Association?</i>	
I. P. L. McLaren and M. Suret.....	340
<i>Regularity and Irregularity in French Inflectional Morphology</i>	
Fanny Meunier and William Marslen-Wilson.....	346



<i>Neighborhood and Position Effects Interact in Naming Latency</i>	
Jeanne C. Milostan, Garrison W. Cottrell, and Victor Ferreira.....	352
<i>Individual Differences in Exemplar-Based Interference During Instructed Category Learning</i>	
David C. Noelle and Garrison W. Cottrell .....	358
<i>Deep Learning in Virtual Reality: How to Teach Children That the Earth is Round</i>	
Stellan Ohlsson, Thomas G. Moher, and Andrew Johnson.....	364
<i>ANCHOR: A Memory-Based Model of Category Rating</i>	
Alexander A. Petrov and John R. Anderson .....	369
<i>The Area Activation Model of Saccadic Selectivity in Visual Search</i>	
Marc Pomplun, Eyal M. Reingold, Jiye Shen, and Diane E. Williams .....	375
<i>The Use of a High-Dimensional, "Environmental" Context Space to Model Retrieval in Analogy and Similarity-Based Transfer</i>	
Michael Ramscar and Daniel Yarlett .....	381
<i>Organising Principles in Lexical Representation: Evidence from Polish</i>	
Agnieszka Reid and William Marslen-Wilson.....	387
<i>From Studying Examples to Solving Problems: Fading Worked-Out Solution Steps Helps Learning</i>	
Alexander Renkl, Robert K. Atkinson, and Uwe H. Maier.....	393
<i>Measuring Verb Similarity</i>	
Philip Resnik and Mona Diab .....	399
<i>The Advantages and Disadvantages of Semantic Ambiguity</i>	
Jennifer Rodd, Gareth Gaskell, and William Marslen-Wilson .....	405
<i>Attention to Action: Securing Task-Relevant Control in Spoken Word Production</i>	
Ardi Roelofs .....	411
<i>Non-Linguistic Constraints on the Acquisition of Phrase Structure</i>	
Jenny R. Saffran .....	417
<i>Attentional Biases in Artificial Noun Learning Tasks: Generalizations Across the Structure of Already-Learned Nouns</i>	
Larissa K. Samuelson.....	423
<i>Phrases as Carriers of Coherence Relations</i>	
Holger Schauer and Udo Hahn.....	429
<i>Syntactic Priming in German Sentence Production</i>	
Christoph Scheepers and Martin Corley .....	435
<i>Hypertext Navigation and Conflicting Goal Intentions: Using Log Files to Study Distraction and Volitional Protection in Learning and Problem Solving</i>	
Katharina Scheiter, Peter Gerjets, and Elke Heise .....	441
<i>Clarifying Word Meanings in Computer-Administered Survey Interviews</i>	
Michael F. Schober, Frederick G. Conrad, and Jonathan E. Bloom .....	447
<i>Seeking Coherent Explanations – A Fusion of Structured Connectionism, Temporal Synchrony, and Evidential Reasoning</i>	
Lokendra Shastri and Carter Wendelken.....	453
<i>Infant Familiarization to Artificial Sentences: Rule-Like Behavior Without Explicit Rules and Variables</i>	
Thomas R. Shultz and Alan C. Bale.....	459
<i>Simulation of Self-Affirmation Phenomena in Cognitive Dissonance</i>	
Thomas R. Shultz and Mark R. Lepper.....	464
<i>Linguistic Labels and the Development of Inductive Inference</i>	
Vladimir M. Sloutsky and Ya-Fen Lo.....	469
<i>Problem Representation in Experts and Novices: Part 2. Underlying Processing Mechanisms</i>	
Vladimir M. Sloutsky and Aaron S. Yarlas.....	475

<i>Prosodic Choice: Effects of Speaker Awareness and Referential Context</i>	
Jesse Snedeker, Lila Gleitman, Michael Felberbaum, Nicora Placa, and John Trueswell...	481
<i>Eye Movements During Comprehension of Spoken Scene Descriptions</i>	
Michael J. Spivey, Melinda J. Tyler, Daniel C. Richardson, and Ezekiel E. Young .....	487
<i>Heterogeneous Reasoning in Learning to Model</i>	
Keith Stenning and Melissa Sommerfeld .....	493
<i>Inducing Hybrid Models of Task Learning from Visualmotor Data</i>	
Devika Subramanian .....	499
<i>Ungrammatical Influences: Evidence for Dynamical Language Processing</i>	
Whitney Tabor and Bruno Galantucci.....	505
<i>Mapping the Syntax/Semantics Coastline</i>	
Whitney Tabor and Sean Hutchins.....	511
<i>Word Learning as Bayesian Inference</i>	
Joshua B. Tenenbaum and Fei Xu.....	517
<i>Aspectual Coercion and the Online Computation of Sentential Aspect</i>	
Marina Todorova, Kathy Straub, William Badecker, and Robert Frank.....	523
<i>Talking through Graphics: An Empirical Study of the Sequential Integration of Modalities</i>	
Ichiro Umata, Atsushi Shimojima and Yasuhiro Katagiri.....	529
<i>The Dynamics of Simple Prediction: Judging Reachability</i>	
Iris van Rooij, Raoul M. Bongers, and W. F. G. (Pim) Haselager .....	535
<i>Goal Specificity and Learning with a Multimedia Program</i>	
Regina Vollmeyer, Bruce D. Burns, and Falko Rheinberg .....	541
<i>Human Belief Revision and the Order Effect</i>	
Hongbin Wang, Jiajie Zhang, and Todd R. Johnson.....	547
<i>Situating GOMS Models Within Complex, Sociotechnical Systems</i>	
Robert L. West and Gabriella Nagy .....	553
<i>Motivation in Insight versus Incremental Problem Solving</i>	
Mareike Wieth and Bruce D. Burns .....	559
<i>Making Inferences and Classifications Using Categories that are not Linearly Separable</i>	
Takashi Yamauchi and Arthur B. Markman .....	565
<i>Structure-Mapping Theory and Lexico-Semantic Information</i>	
Daniel Yarlett and Michael Ramscar .....	571
<i>Selective Advantages of Syntactic Language – A Model Study</i>	
Willem H. Zuidema and Paulien Hogeweg.....	577

## Posters

<i>A Natural Bias for the Basic Level?</i>	
Annie Archambault, Frédéric Gosselin, and Philippe G. Schyns.....	585
<i>Subgoal Learning and the Effect of Conceptual vs. Computational Equations on Transfer</i>	
Robert K. Atkinson and Richard Catrambone.....	591
<i>A Comparative Study of Unsupervised Grapheme-Phoneme Alignment Methods</i>	
Timothy Baldwin and Hozumi Tanaka .....	597
<i>Discovering and Describing Category Differences: What Makes a Discovered Difference Insightful?</i>	
Stephen D. Bay and Michael J. Pazzani.....	603
<i>Harmonia Loosely Praestabilita: Discovering Adequate Inductive Strategies</i>	
Hilan Bensusan and Christophe Giraud-Carrier.....	609

<i>Path and Manner Priming: Verb Production and Event Recognition</i> Dorrit Billman, Angela Swilley, and Meredyth Krych .....	615
<i>Reasoning from Shared Structure</i> Sergey Victor Blok and Dedre Gentner .....	621
<i>Problem Solving: Phenomena in Search of a Thesis</i> Bruce D. Burns and Regina Vollmeyer .....	627
<i>The Representational Effect in Complex Systems: A Distributed Representation Approach</i> Johnny Chuah, Jiajie Zhang, and Todd R. Johnson .....	633
<i>Precursors to Number: Making the Most of Continuous Amount</i> Peter Drake, Kelly Mix, and Melissa Clearfield .....	639
<i>Subjacency Constraints without Universal Grammar: Evidence from Artificial Language Learning and Connectionist Modeling</i> Michelle R. Ellefson and Morten H. Christiansen .....	645
<i>Reflective Introspective Reasoning through CBR</i> Susan Eileen Fox .....	651
<i>The Chinese Room: Just Say "No!"</i> Robert M. French .....	657
<i>The Influence of Source and Cost of Information Access on Correct and Errorful Interactive Behavior</i> Wayne D. Gray and Wai-Tat Fu .....	663
<i>Practices of Questioning and Explaining in Learning to Model</i> James G. Greeno, Melissa C. Sommerfeld, and Muffie Wiebe .....	669
<i>Work at the Interface between Representing and Represented Worlds in Middle School Mathematics Design Projects</i> Rogers Hall .....	675
<i>Four Letters Good, Six Letters Better: Exploring the Exterior Letters Effect with a Split Architecture</i> John Hicks, Jon Oberlander, and Richard Shillcock .....	681
<i>Eye-Tracking and Conceptual Combination</i> Dietmar Janetzko .....	687
<i>The Role of Mental Imagery in Understanding Unknown Idioms</i> Armina Janyan and Elena Andonova .....	693
<i>Does Collaborative Learning Lead to the Construction of Common Knowledge?</i> Heisawn Jeong and Michelene T. H. Chi .....	699
<i>Learning from a Computer Workplace Simulation</i> Heisawn Jeong, Roger Taylor, and Michelene T. H. Chi .....	705
<i>A Dynamical Model of Insightful Memory Retrieval</i> Koji Jimura, Hisaaki Komazaki, Takashi Matsuoka, Masanori Nakagawa, and Takashi Kusumi .....	711
<i>Declarative and Procedural Learning in Alphabetic Retrieval</i> Todd R. Johnson, Hongbin Wang, and Jiajie Zhang .....	717
<i>A Process Model of Children's Early Verb Use</i> Gary Jones, Fernand Gobet, and Julian M. Pine .....	723
<i>Handedness and Heterogeneity in Cognitive Science</i> Gregory V. Jones and Maryanne Martin .....	729
<i>Using Referential Communication to Study Mental Models</i> Julia Kalmanson and Arthur B. Markman .....	735
<i>Was Apatosaurus a Vegan? Dinosaur Knowledge Rocks When Learning about Evolution</i> David R. Kaufman, Michael Ranney, Eric Lewis, Anna Thanukos, and Sarah Brem .....	741
<i>Evaluating Competition-Based Models of Word Order</i> Frank Keller .....	747

<i>The Précis of Project Nemo, Phase 2: Levels of Expertise</i> Susan S. Kirschenbaum and Wayne D. Gray .....	753
<i>Visual and Spatial Representations in Relational Reasoning</i> Markus Knauff and P. N. Johnson-Laird .....	759
<i>Constraints of Embodiment on Action Coordination</i> Günther Knoblich and J. Scott Jordan.....	764
<i>SEQL: Category Learning as Progressive Abstraction Using Structure Mapping</i> Sven E. Kuehne, Kenneth D. Forbus, Dedre Gentner, and Bryan Quinn.....	770
<i>Learning-Based Constraints on Schemata</i> Peter C. R. Lane, Fernand Gobet, and Peter C-H. Cheng .....	776
<i>Retrospective Effects in Human Causality Judgment</i> M. E. Le Pelley, D. L. Cutler, and I. P. L. McLaren .....	782
<i>A Constructivist Model of Robot Perception and Performance</i> Joseph A. Lewis and George F. Luger .....	788
<i>Point-Light Displays Illuminate the Abstract Nature of Children's Motion Verb Representations</i> Jing Liu, Roberta M. Golinkoff, Kim Piper, He Len Chung, Kathy Hirsh-Pasek, Christopher H. Ramey, and Bennett I. Bertenthal.....	794
<i>Learning at Different Levels of Abstraction</i> Bradley C. Love .....	800
<i>The Direct Route: Mediated Priming in Semantic Space</i> Will Lowe and Scott McDonald.....	806
<i>Zen in the Art of Language Acquisition: Statistical Learning and the Less is More Hypothesis</i> David Ludden and Prahlad Gupta .....	812
<i>Two Views are Better than One: Epistemic Actions May Prime</i> Paul P. Maglio and Michael J. Wenger.....	818
<i>Preschool Children's Use of Category Information to Interpret Negations</i> Bradley J. Morris.....	823
<i>An Inquiry into the Function of Implicit Knowledge and its Role in Problem Solving</i> Timothy J. Nokes and Stellan Ohlsson .....	829
<i>Hypothesis-Testing Method in a Community of Psychologists</i> Takeshi Okada and Takashi Shimokido.....	835
<i>Fast and Frugal Use of Cue Direction in States of Limited Knowledge</i> Magnus Persson and Peter Juslin .....	841
<i>Effects of Presentation Format on Memory for Order</i> Margaret J. Peterson and Erik M. Altmann.....	847
<i>Teaching and Supporting the Use of Qualitative and Quantitative Concepts in Classical Mechanics</i> Rolf Ploetzner and Sieghard Beller.....	853
<i>The Implications of Cognitive Science for the Significance of Experimentation in Science Teaching</i> Athanasios Raftopoulos and Constantinos P. Constantinou .....	859
<i>Evidence for the Processing of Re-representations during the Mapping of Externally Represented Analogies</i> Michael Ramscar.....	865
<i>Searching for Alternatives in Spatial Reasoning: Local Transformations and Beyond</i> Reinhold Rauh, Cornelius Hagen, Christoph Schlieder, Gerhard Strube, and Markus Knauff .....	871
<i>Memory for Continually Changing Information: A Task Analysis and Model of the Keeping Track Task</i> Wolfgang Schoppek .....	877

<i>Motivating Base-Rate Sensitivity (Sometimes): Testing Predictions of the RCCL Framework</i> Christian Schunn and Thuy L. Ngo.....	883
<i>Now They See the Point: Improving Science Reasoning through Making Predictions</i> Christian D. Schunn and Christine J. O'Malley.....	889
<i>Dueling Theories: Thought Experiments in Cognitive Science</i> Sam Scott.....	895
<i>Temporal Progression of the Cortical Potential Distribution for the AEP P300 Component in Mild Traumatic Brain Injury</i> Robert D. Sidman, Lan Ke, and Martin R. Ford .....	901
<i>What are Fallacies Good for? Representational Speed-Up in Propositional Reasoning</i> Vladimir M. Sloutsky .....	907
<i>The Primacy of One-to-One Generalization in Young Children's Induction</i> Vladimir M. Sloutsky and Ya-Fen Lo.....	912
<i>Simulating Conditional Reasoning Containing Negations: A Computer Model and Human Data</i> Jacques Sougné .....	918
<i>A Dynamic Field Model of Location Memory</i> John P. Spencer and Gregor Schöner .....	924
<i>A Simple Categorisation Model of Anaphor Resolution</i> Andrew J. Stewart and Frederic Gosselin .....	930
<i>Category Induction for Ordinary Facts</i> Roman Taraban and Matt Hayes.....	936
<i>Learning and Generalizing New Concepts</i> Jean-Pierre Thibaut .....	942
<i>Rules versus Statistics in Biconditional Grammar Learning: A Simulation Based on Shanks et al. (1997)</i> Bert Timmermans and Axel Cleeremans .....	947
<i>Representational Scaffolding during Scientific Inquiry: Interpretive and Expressive Use of Inscriptions in Classroom Learning</i> Eva Erdosne Toth.....	953
<i>From Dipsy-Doodles to Streaming Motions: Changes in Representation in the Analysis of Visual Scientific Data</i> Susan B. Trickett, Wai-Tat Fu, Christian D. Schunn, and J. Gregory Trafton.....	959
<i>Blobs, Dipsy-Doodles and Other Funky Things: Framework Anomalies in Exploratory Data Analysis</i> Susan B. Trickett, J. Gregory Trafton, and Christian D. Schunn .....	965
<i>Reaction Times and Predictions in Sequence Learning: A Comparison</i> Ingmar Visser, Maartje E. J. Raijmakers, and Peter C. M. Molenaar .....	971
<i>A Constructivist Dual-Representation Model of Verb Inflection</i> Gert Westermann.....	977
<i>Contextually Representing Abstract Concepts with Abstract Structures</i> Katja Wiemer-Hastings and Arthur C. Graesser .....	983
<i>Adding Syntactic Information to LSA</i> Peter Wiemer-Hastings .....	989
<i>Categorization and the Ratio Rule</i> A. J. Wills, Mark Suret, and I. P. L. McLaren .....	994
<i>Strategies and Tactics in Sentential Reasoning</i> Yingrui Yang, Jean-Baptiste van der Henst, and P. N. Johnson-Laird .....	1000
<i>Problem Representation in Experts and Novices: Part 1. Differences in the Content of Representation</i> Aaron S. Yarlas and Vladimir M. Sloutsky.....	1006

## Member Abstracts

<i>Accentuation of Category Differences: Revisiting a Classic Study</i>	
Janet K. Andrews and Kenneth R. Livingston .....	1015
<i>Training and Transfer of Foreign Word Identification at Three Speeds</i>	
Anita R. Bowles and Alice F. Healy .....	1016
<i>Temporal Tuning in the Acquisition of Cognitive Skill</i>	
Richard A. Carlson and Lisa M. Stevenson .....	1017
<i>Hemispheric Effects in Fusiform Gyrus across Face Encoding Tasks</i>	
Daniel J. Casasanto and John A. Detre .....	1018
<i>Detecting Animals in Point-Light Displays</i>	
Leslie Cohen, Thomas F. Shipley, Eve Marshark, Kathy Taht, and Denise Aster .....	1019
<i>Familiarity and Categorical Inference</i>	
David Collister and Barbara Tversky .....	1020
<i>How Words Get Special</i>	
Eliana Colunga and Linda B. Smith .....	1021
<i>Information Processing Speed During Functional Neuroimaging of Sentence Comprehension</i>	
Ayanna Cooke, Christian DeVita, David Alsop, James Gee, John Detre, and Murray Grossman.....	1022
<i>Experimentally Uncovering Hidden Strata in English Phonology</i>	
Lisa Davidson.....	1023
<i>Language after Hemispherectomy: Effects of Seizure Control</i>	
Stella de Bode, Susan Curtiss, and Gary W. Mathern.....	1024
<i>Compositional Functions in Nominal Combination</i>	
Zachary Estes .....	1025
<i>Eye Movements in Human Face Learning and Recognition</i>	
Richard J. Falk, Andrew Hollingworth, John M. Henderson, Sridhar Mahadevan, and Fred C. Dyer .....	1026
<i>Comprehension of Active and Passive Sentences in Portuguese and English: The Prototypicality Effect</i>	
Rosângela Gabriel and Kim Plunklett .....	1027
<i>Stages of Phonological Processing in Spoken Production</i>	
Matthew Goldrick, Brenda Rapp, and Paul Smolensky .....	1028
<i>Why Are Some Problems Easy? New Insights into the Tower of Hanoi</i>	
Glenn Gunzelmann and Stephen Blessing .....	1029
<i>Of Words, Birds, Worms, and Weeds: Infant Word Learning and Lexical Neighborhoods</i>	
George J. Hollich, Peter W. Jusczyk, and Paul A. Luce .....	1030
<i>Modelling Language Acquisition at Multiple Temporal Scales</i>	
Steve R. Howell and Suzanna Becker .....	1031
<i>Perceptual and Experience-Dependent Influences on Location Memory Processes</i>	
Alycia M. Hund and John P. Spencer .....	1032
<i>A Study of Age-of-Acquisition Ratings in Adults</i>	
Gowri K. Iyer, Cristina M. Saccuman, Elizabeth A. Bates, and Beverly B. Wulfeck .....	1033
<i>The Development of Word Recognition: The Use of the Possible Word Constraint by 12-Month-Olds</i>	
Elizabeth K. Johnson, Peter W. Jusczyk, Anne Cutler, and Dennis Norris .....	1034
<i>Familiarity for Nouns and Verbs: Not the Same as, and Better than, Frequency</i>	
Natalie Kacirik, Connie Shears, and Christine Chiarello .....	1035

<i>Random Indexing of Text Samples for Latent Semantic Analysis</i>	
Pentti Kanerva, Jan Kristoferson, and Anders Holst .....	1036
<i>Auditory and Visual Continuity Perception: A Unifying Theory</i>	
Leah M. Knightly .....	1037
<i>The Role of Working Memory in Homograph Recognition</i>	
Yuki Kobayashi .....	1038
<i>Attentional Perseveration after the Inverse Base-Rate Effect</i>	
John K. Kruschke, Mark K. Johansen, and Nathaniel J. Blair .....	1039
<i>Do Readers Make Predictive Inferences about Conversations?</i>	
R. Brooke Lea, Patrick A. Kayser, Elizabeth J. Mulligan, and Jerome L. Myers .....	1040
<i>A Model of Prefrontal-Hippocampal Interactions in Strategic Recall</i>	
Jean C. Lim and Suzanna Becker .....	1041
<i>An Adaptive Model of Simple Communication</i>	
Michael Matessa and John R. Anderson .....	1042
<i>Morphological Influences on Phonetic Categorization</i>	
Kerstin Mauth .....	1043
<i>Unique Entropy as a Model of Linguistic Classification</i>	
Toben H. Mintz .....	1044
<i>Analogical Priming in a Word Naming Task</i>	
Robert G. Morrison, Keith J. Holyoak, and Barbara A. Spellman .....	1045
<i>Finding Common Ground in Children's Referential Communication</i>	
Aparna Nadig and Julie Sedivy .....	1046
<i>If Robots Make Choices, are they Alive?: Children's Judgments of the Animacy of Intelligent Artifacts</i>	
Milena K. Nigam and David Klahr .....	1047
<i>Effects of Visualization on Familiar Motion Problems</i>	
Matia Okubo .....	1048
<i>The Problem of Relevance in Blended Mental Spaces</i>	
David Paxman .....	1049
<i>Interpreting Eye-Movement Protocols</i>	
Dario D. Salvucci and John R. Anderson .....	1050
<i>Learning to Learn by Modular Neural Networks</i>	
Akio Sashima and Kazuo Hiraki .....	1051
<i>Modeling Embodied Cognition in a Complex Real-Time Task</i>	
Michael J. Schoelles and Wayne D. Gray .....	1052
<i>Does Human Memory Reflect the Environment of Early Hominids?</i>	
Lael Schooler, Juan Carlos Serio Silva, and Ramon Rhine .....	1053
<i>Knowledge Construction Links: Cues and Trajectories as Prior Experience and Knowledge</i>	
Kathy L. Schuh .....	1054
<i>A Delay-Dependent Switch in the Information Children Use to Remember Locations</i>	
Anne R. Schutte and John P. Spencer .....	1055
<i>19-Month-Olds' Sensitivity to Negation/Tense Dependencies</i>	
Melanie Soderstrom, Peter Jusczyk, and Kenneth Wexler .....	1056
<i>Scaling and Testing for Non-Euclidean Spaces</i>	
Jesse Spencer-Smith .....	1057
<i>Variation in Children's Word Production: Can 'Competence' Models Deal with Young Children's Truncation Patterns?</i>	
Helena Taelman and Steven Gillis .....	1058
<i>Is Musical Ability Related to the Prosody Learning of Second Language?</i>	
Akihiro Tanaka and Yohtaro Takano .....	1059

<i>Main Idea Identification: A Functional Imaging Study of a Complex Cognitive Process</i> Lêda Maria Braga Tomitch, Marcel Adam Just, and Patricia A. Carpenter.....	1060
<i>Schema Acquisition and Solution Strategy in Statistics Problem Solving</i> David Trumpower .....	1061
<i>Management of Multiple Goals on the Basis of Situational Urgency</i> Takafumi Tsuchiya.....	1062
<i>Transformational Analyses of Visual Perception</i> Douglas Vickers and Adrian K. Preiss.....	1063
<i>Use of Agent and Object-Oriented Information in Language Acquisition</i> Laura Wagner.....	1064
<i>Domains, Knowledge, and Constraints on Classification</i> William D. Wattenmaker, Kathleen A. Filak, and Josephine A. Mendoza.....	1065
<i>Knowledge Effects, Conceptual Structure, and Incidental Learning</i> William D. Wattenmaker, Josephine A. Mendoza, and Vanessa K. Nieves.....	1066
<i>Using Cognitive Models in the Design and Evaluation of Team Structure</i> Monica Z. Weiland and James L. Eilbert.....	1067
<i>Scene Context and Change Blindness: Memory Mediates Change Detection</i> Carrick C. Williams, Andrew Hollingworth, and John M. Henderson .....	1068
<i>Sequential Probability as a Segmentation Cue for Cantonese</i> Michael C. W. Yip .....	1069
<i>The Effect of Languages on Children's Use of Action Information</i> Hanako Yoshida and Linda B. Smith.....	1070
Author Index .....	1071



# Twenty-Second Annual Conference of the Cognitive Science Society

August 13-15, 2000

Institute for Research in Cognitive Science

University of Pennsylvania

Philadelphia, PA

## Conference Co-Chairs

Lila Gleitman, University of Pennsylvania

Aravind Joshi, University of Pennsylvania

## Conference Program Committee

Norm Badler, University of Pennsylvania  
Mark Baker, Rutgers University  
Gene Buckley, University of Pennsylvania  
Paul G. Chapin, National Science Foundation  
Robin Clark, University of Pennsylvania  
Veronica Dahl, Simon Fraser University  
Kostas Daniilidis, University of Pennsylvania  
Susan Epstein, Hunter College and the City  
University of New York  
Martha Farah, University of Pennsylvania  
Gilles Fauconnier, University of California, San  
Diego  
Kenneth D. Forbus, Northwestern University  
Bob Frank, Johns Hopkins University  
Randy Gallistel, University of California, Los  
Angeles  
Rochel Gelman, University of California, Los  
Angeles  
Dedre Gentner, Northwestern University  
Geoff Hall, University of British Columbia  
Gary Hatfield, University of Pennsylvania  
Frank Keil, Yale University  
Phil Kellman, University of California, Los  
Angeles  
Michael Kelly, University of Pennsylvania  
David Knill, University of Rochester  
Tony Kroch, University of Pennsylvania  
Barbara Landau, University of Delaware

David Lebeaux, NEC Research Institute  
Mitch Marcus, University of Pennsylvania  
Dimitris Metaxas, University of Pennsylvania  
Johanna Moore, University of Pittsburgh  
Rolf Noyer, University of Pennsylvania  
Martha Palmer, University of Pennsylvania  
Ellen F. Prince, University of Pennsylvania  
Zenon Pylyshyn, Rutgers University  
Owen Rambow, ATT Labs - Research  
Daniel Reisberg, Reed College  
Virginia Richards, University of Pennsylvania  
Whitman Richards, Massachusetts Institute of  
Technology  
Don Ringe, University of Pennsylvania  
Maria-Isabel Romero, University of  
Pennsylvania  
Paul Smolensky, Johns Hopkins University  
Matthew Stone, Rutgers University  
Gary Strong, NSF and DARPA  
Michael K. Tanenhaus, University of Rochester  
Paul Thagard, University of Waterloo  
Sharon L. Thompson-Schill, University of  
Pennsylvania  
John Trueswell, University of Pennsylvania  
Sandra R. Waxman, Northwestern University  
Bonnie Webber, University of Edinburgh  
Janet Werker, University of British Columbia  
Fei Xu, Northeastern University

## Conference Staff

Ann Bies, Nicole Bolden, Lee Leiber, Jennifer MacDougall, Laurel Sweeney, Trisha Yannuzzi

## 2000 Marr Prize

Eliana Colunga and Linda B. Smith, Indiana University  
*Committing to an Ontology: A Connectionist Account*

This conference was supported by the Cognitive Science Society, the National Institutes of Health, Microsoft Corporation, and the Institute for Research in Cognitive Science at the University of Pennsylvania.

# **The Cognitive Science Society**

## **Governing Board**

Jeffrey Elman, University of California at San Diego, La Jolla  
Vimla Patel, McGill University  
Susan L. Epstein, Hunter College and the City University of New York  
Kim Plunkett, Oxford University  
Martha Farah, Univ. of Pennsylvania  
Lawrence W. Barsalou, Emory University  
Paul Thagard, University of Waterloo  
Alan Lesgold, University of Pittsburgh  
Kenneth D. Forbus, Northwestern University  
Dedre Gentner, Northwestern University  
Douglas L. Medin, Northwestern University  
Michael Mozer, University of Colorado, Boulder  
Keith Stenning, Edinburgh University

## **Chair of the Governing Board**

Jeffrey Elman, University of California at San Diego, La Jolla

## **Chair Elect**

Lawrence W. Barsalou, Emory University

## **Journal Editor**

James G. Greeno, Stanford University

## **Executive Officer**

Colleen Seifert, University of Michigan

The Cognitive Science Society, Inc., was founded in 1979 to promote interchange across traditional disciplinary lines among researchers investigating the human mind. The Society sponsors an annual meeting, and publishes the journal *Cognitive Science*. Membership in the Society requires a doctoral degree in a related discipline (or equivalent research experience); graduate and undergraduate students are eligible for a reduced rate membership; and all are welcome to join the society as affiliate members. For more information, please contact the society office or see their web page at <http://www.umich.edu/~cogsci>

Cognitive Science Society, Univ. of Michigan, 525 East University, Ann Arbor, MI, 48109-1109; [cogsci@umich.edu](mailto:cogsci@umich.edu); phone and fax: (734) 429-4286.

# **Tutorial Program**

August 11-12, 2000

## **The Two Visual Systems and their Interactions**

Bruce Bridgeman, University of California, Santa Cruz

## **An introduction to the COGENT Cognitive Modelling Environment**

Dr. Richard Cooper, Birkbeck College, England

## **Cognitive Modeling and Simulation in Real-Time/Multi-tasking Domains Using the COGNET Framework**

Wayne Zachary, CHI Systems

## **Cognitive Neuroscience for Cognitive Scientists**

Martha Farah, University of Pennsylvania

## **Cognitive Science and Education: Creating, Implementing, and Funding Collaborations to Build a Cognitive Science Foundation for Teaching and Learning**

Christine Massey, IRCS, University of Pennsylvania

## **Tutorial Program Committee Co-Chairs**

Toben Mintz, University of Southern California

Ione Fine, Salk Institute for Biological Studies

## **Symposia**

### **Perspectives on Conceptual Change**

David Kaufman, University of California, Berkeley

### **The Nature of Human Errors**

Jiajie Zhang, University of Texas, Houston

### **The Role of the Cerebellum in Cognition and Affect**

Natika Newton, Nassau Community College

### **Bayesian Approaches to Cognitive Modeling**

Michael C. Mozer, University of Colorado at Boulder and  
Joshua Tenenbaum, Stanford University

### **Symposia Chair**

Bonnie Webber, University of Edinburgh

## **Special Session on Undergraduate Education in Cognitive Science**

Cognitive Science presents unique opportunities and problems as a component of the undergraduate curriculum. The opportunity is to provide a view of natural and artificial minds that is relevant to a wide range of careers in research, technology and the professions. The problem is to present this material in a form that coheres as a subject, and to ensure sufficient background to teach it in sufficient depth. The symposium presents a series of short lectures by distinguished educators and researchers from Europe and North America who will explore recent developments and future directions for the subject.

Among those participating are:

Andrew Brook, Carleton University  
Nils Dahlback, Linköping University  
Randy Jones, Colby College  
Keith Stenning, University of Edinburgh  
John C. Trueswell, University of Pennsylvania

### **Special Session Chair**

Mark Steedman, University of Edinburgh

## Reviewers for the Twenty-Second Annual Conference of the Cognitive Science Society

Agnar Aamodt	Mei Chen	Wayne D. Gray
Adele Abrahamsen	Sonu Chopra-Khullar	Prahlad Gupta
Woo-Kyoung Ahn	James Chumbley	Karl Haberlandt
Isabel Albers	Timothy C. Clausner	Udo Hahn
Martha W. Alibali	Catherine A. Clement	Koichiro Hajiri
Jan Allbeck	Charles Clifton	James Hampton
Amit Almor	Marc Cohen	Mary L. Hare
Erik M. Altmann	Phil Cohen	Trevor A. Harley
Jorge Alvoeiro	Fred Conrad	Catherine L. Harris
Mike Anderson	Andrew Conway	Nancy Hedberg
Jan Andrews	Martin Corley	Barbara Hemforth
Jennifer Arnold	Sharon A. Cote	Petra Hendriks
Richard Aslin	Gary Cottrell	Eduard Hoenkamp
Robert K. Atkinson	Hartvig Dahl	Douglas Hofstadter
Stephanie E. August	Hoa Trang Dang	Jorge Horas
Neville Austin	Zoltan Dienes	Eva Hudlicka
Norm Badler	Don Dulany	Elizabeth Ince
Nicolas Balacheff	Kevin Dunbar	Dennis Irons
Breck Baldwin	Karen Ehrlich	Ray Jackendoff
Linden Ball	Rene Elio	Dietmar Janetzko
Mark Baltin	Jeff Elman	Heisawn Jeong
Amy L. Baylor	Bruno Emond	Luis Jimenez
Bettina Berendt	Eileen Entin	Susan C. Johnson
David Beversdorf	Charles Erignac	Randolph M. Jones
Collin Billingsley	Martha W. Evens	Peter Juslin
Rama Bindiganavale	Howard T. Everson	Bertie Kaal
Harry Blanchard	Marte Fallshore	Michael B. Kac
Peter Bosch	Christiane Fellbaum	Bob Kachelski
Brian F. Bowdle	Antonio Fernandez-Caballero	Jim Kahn
Amy Brand	Rodolfo A. Fiorini	Pentti Kanerva
D. S. Bree	John Foxe	Barbara Kaup
Bruce Bridgeman	Reva Freedman	Mark Keane
Paul Brna	Robert M. French	Albert Kim
Andrew Brook	Stefan Frisch	John J. Kim
Patricia Brooks	Michael Gasser	Thomas King
Curtis Brown	Ted Gibson	Ishikawa Kiyoshi
Joanna Bryson	Andrew Glassner	Sheldon Klein
Lori Buchanan	Robert Goldstone	Guenther Knoblich
Curt Burgess	Laura M. Gonnerman	Christopher Koch
Bruce Burns	Barbara L. Gonzalez	Boicho Kokinov
Mike Byrne	Cleotilde Gonzalez	John Kounios
Thomas J. Capo	Paula Goolkasian	Rita Kovordanyi
Rich Carlson	Adrian Gordon	Seth Kulick
Laura Carlson-Radvansky	Art Graesser	Howard S. Kurtzman
Richard Catrambone	Jordan Grafman	Patrick Kyllonen
Violetta Cavalli-Sforza	Barbara Graves	William Langston

Nancy C. Lavinge  
David Leake  
Adrienne Lee  
Yuh-shiow Lee  
F. K. Lehman  
Ping Li  
Jeff Lidz  
Bowen Loftin  
Bradley C. Love  
Francis Lowenthal  
George F. Luger  
Paul Maglio  
Lorenzo Magnani  
Barbara Malt  
Arthur B. Markman  
Amy Masnick  
Chris Massey  
Mark E. Mattson  
Rachel McCloy  
Jean McKendree  
Ken McRae  
Susan McRoy  
David A. Medler  
Paul Messaris  
Craig S. Miller  
Naomi Miyake  
Joyce L. Moore  
Viv Moore  
Bradley J. Morris  
Tom Morton  
Erik T. Mueller  
Paul Munro  
Gregory L. Murphy  
Julien Musolino  
N. Hari Narayanan  
Aldo Nemesio  
Josef Nerb  
David Noelle  
Padraig G. O'Seaghdha  
Scott Overmyer  
Michael Pan  
Anna Papafragou  
Vimla L. Patel  
Neal Pearlmutter  
David Peebles

Jeff Pelletier  
Alexander Petrov  
Penny Pexman  
Steven Phillips  
Massimo Piattelli-Palmarini  
Robert F. Port  
Rashmi Prasad  
Athanasios Protopapas  
Joseph Psotka  
Sadhana Puntambekar  
Clark N. Quinn  
Athanasios Raftopoulos  
William J. Rapaport  
Eric Raufaste  
Stephen J. Read  
Michael A. Redmond  
Stephen K. Reed  
Alexander Renkl  
Jeff Rickel  
Frank Ritter  
Robert M. Roe  
Anoop Sarkar  
Gabriele Scheler  
Matthew Schlesinger  
Ute Schmid  
Christian Schunn  
Daniel L. Schwartz  
Julie Sedivy  
Michael G. Shafto  
Stuart C. Shapiro  
Richard Shillcock  
Thomas R. Shultz  
Jeff Siskind  
Derek Sleeman  
Peter Slezak  
Vladimir Sloutsky  
Eliot Smith  
Linda B. Smith  
Jesse Snedeker  
Cristina Sorrentino  
Ann Speed  
Michael Spivey  
Constance Steinkuehler  
Rosemary Stevenson  
Suzanne Stevenson

Michael Strait  
Robert Stufflebeam  
Patrick Sturt  
Ron Sun  
Prasad Tadepalli  
Heike Tappe  
Roman Taraban  
Virginia Teller  
Joshua Tenenbaum  
Paul Thagard  
Jean-Pierre Thibaut  
Nigel J. T. Thomas  
Charles Tijus  
Greg Trafton  
Roy Turner  
Barbara Tversky  
Jim Uleman  
Jody Underwood  
Lyle Ungar  
Frank Van Overwalle  
Alonso Vera  
Gregg Vesonder  
Douglas Vickers  
Michael R. Waldmann  
Iain Wallace  
J. G. Wallace  
William P. Wallace  
Hongbin Wang  
Pei Wang  
Robert Widner  
Shuly Wintner  
Andrew Wishart  
Wallace H. Wulfeck  
Takashi Yamauchi  
Yingrui Yang  
Wai-Kiang Yeap  
Richard M. Young  
Wayne Zachary  
Jeff Zacks  
Cornelia Zelinsky-Wibbelt  
Jiajie Zhang  
Corinne Zimmerman  
Ludger van Elst  
Emile van der Zee

## **Invited Speakers**

### **Spoken Language Systems and Human Communication**

James F. Allen, University of Rochester

A conversational agent is a system that can engage in natural language conversation in order to further its goals. Such a system requires a concerted effort to bring work in natural language understanding, dialog modeling, knowledge representation and reasoning into a single coherent system. The TRIPS system is a specific agent that can support unconstrained dialogue in order to assist the user in problem solving tasks. Significant effort has been made to make the system robust, so that it can perform well even in the face of inevitable speech recognition errors, and can continue the dialogue in a natural way under any circumstances. This talk will discuss the overall project and its accomplishments so far, and then focus on a few specific mechanisms that enable robust interaction. It will also review some recent evidence on the nature of human language comprehension and discuss the implications of this work for future dialogue systems.

### **The Symbolic Foundations of Conditioned Behavior**

Randy Gallistel, University of California, Los Angeles

The concept of associative learning has been central to psychology for more than a century. In the 60s and 70s, the associative conception of mind was challenged by the rise of the information processing approach, which emphasizes the construction of a symbolic representation of experience by means of computational operations. This latter conception has come to dominate thinking about sensory processing and perception at both the psychophysical and the neurobiological levels of analysis, but it has remained largely alien to our thinking about learning and memory. Conditioning paradigms – Pavlovian and operant conditioning protocols – were created in order to establish the laws of association formation, and it is widely assumed that the results of conditioning studies are in fact explainable on associative grounds. However, long neglected quantitative features of the data from conditioning experiments pose strong challenges to fundamental assumptions, such as that temporal pairing is a sine qua non for association formation or that reinforcement strengthens and non-reinforcement weakens associations. These assumptions are central not only to associative theories of conditioned behavior but also to connectionist models of behavioral processes and to efforts by neuroscientists to determine the cellular and molecular bases of learning and memory. The findings that challenge associative assumptions are readily explained by information processing theories of conditioning in which subjects measure and record intervals, estimate conditional rates of occurrence, and make decisions based on noisy decision variables computed from these remembered estimates. The success of information processing models on the home ground of associative models suggests that the brain is truly a symbol processing organ, whose operation must be understood in information processing terms.



## **Special Session on Undergraduate Education in Cognitive Science**

Cognitive Science presents unique opportunities and problems as a component of the undergraduate curriculum. The opportunity is to provide a view of natural and artificial minds that is relevant to a wide range of careers in research, technology and the professions. The problem is to present this material in a form that coheres as a subject, and to ensure sufficient background to teach it in sufficient depth. The symposium presents a series of short lectures by distinguished educators and researchers from Europe and North America who will explore recent developments and future directions for the subject.

### **Chair**

Mark Steedman, University of Edinburgh

### **Abstracts**

#### **The Cognitive Science Programme at Carleton University**

Andrew Brook, Carleton University

Carleton University in Ottawa offers free-standing, fully-integrated Cognitive Science programmes at both the undergraduate and the doctoral level. The undergraduate programme was recently reviewed by two senior Canadian cognitive researchers. Some of their findings and recommendations might be of interest to other cognitive science educators. In this presentation, I will describe some of the strengths and weaknesses of both our programmes, summarize some of the key findings and recommendations about the undergraduate programme, and discuss the implications of the latter.

#### **The Three Cultures of Cognitive Science**

Nils Dahlback, Linköping University

When designing an educational program for cognitive science, it is important to base this on some coherent view of the field. If not, there is a risk that the students are presented with a haphazard selection of courses that is more a reflection of the interests of the current available faculty than anything else. Lacking an integrative framework, it will be difficult for the students to relate the different topics and perspectives presented to each other, and it will be difficult for prospective employers in both industry and academia to know which competence the students bring with them. In this talk, an alternative view of cognitive science as neither one unified cognitive science nor just a multidisciplinary field of a number of sciences (psychology, AI, linguistics, philosophy, neuroscience, etc.) is presented. It is argued that cognitive science is best described as a matrix of two dimensions, a content or domain dimension (e.g., language, problem solving etc., and subsets of these) and a methods dimension, comprising three basic approaches to research: empirical, formal, and model building. The latter are seen not only as methods per se, but rather as scientific 'cultures,' carriers of differing explicit and implicit views of what constitutes 'good research.' Since cultural knowledge to a large extent can be acquired only by 'living' in the culture, the Linköping Cognitive Science Master's program is built on the assumption that the students should early in their studies learn all three scientific traditions, both the theoretical and methodological aspects, on an equal footing, before specializing in a particular sub-field. The talk presents the 'three cultures' view of cognitive science, how this has influenced the general design of the program, and describes theoretical and applied courses which illustrate our approach to supporting the students acquiring their own perspective of a multicultural but still unified field of cognitive science.

## **Using Project Work in Teaching Cognitive Science**

Randy Jones, Colby College

In recent years I have developed and taught introductory undergraduate courses in cognitive science and artificial intelligence, as well as one graduate-level course on cognitive science. In developing my courses, I have strongly subscribed to a proposition that I assume most of us (as teachers and cognitive scientists) believe: an effective form of education requires students to participate actively in constructive projects that exercise the material they are learning in class. This presentation describes a set of class projects I have developed, in the hopes that others will find them useful in teaching courses on cognitive science. Some of the projects are directly applicable to a cognitive science course. Others I have developed for a course on artificial intelligence, but would also be appropriate for cognitive science, depending on the emphasis of the course. Among other topics, the projects include study of knowledge representation, learning, production systems, cognitive modeling, and interactive systems. Project descriptions and executable code are available on-line at <http://www.cs.colby.edu/~rjones/courses/cs397/projects/> and <http://www.cs.colby.edu/~rjones/courses/cs353/projects/>.

## **Teaching Multiple Disciplinary Perspectives: A First Year Course in Cognitive Science**

Keith Stenning, University of Edinburgh

First year Edinburgh undergraduate students arrive knowing little of AI, computer science, linguistics, logic, philosophy, psychology – the component disciplines of cognitive science. They don't even know whether these component disciplines are of interest to them. For the last four years we have been teaching a half-year course on Human Communication which is designed to give students from all departments in the university a grasp of what cognitive science is, and how the component disciplines contribute to it. This talk will describe some of our experiences. Are we best teaching single disciplines first, and integrating only after? Or is it better to start by treating disciplines as merely perspectives on a common subject matter?

## **Cognitive Science Education at Penn and the Undergraduate Summer Workshop**

John C. Trueswell

In this talk, I will try to convey the flavor of cognitive science education from the perspective of the group at the University of Pennsylvania. Over the years, Penn has maintained a loose federation approach to cognitive science, in the sense that the departments that make up the participating subdisciplines of cognitive science have used the Institute for Research in Cognitive Science (IRCS) as a gathering place and intellectual-exchange center, but also as a catalyst for interdisciplinary course offerings at the undergraduate and graduate level. I'll discuss the pros and cons of such an educational system, and emphasize how it leaves open the definition of cognitive science, recognizing the current dynamics of the field. As an illustration of this, I will focus on a recent educational initiative stemming from IRCS: The Undergraduate Summer Workshop in Cognitive Science and Cognitive Neuroscience. Each year, IRCS brings together a select group of undergraduate students from around the world who are interested in pursuing graduate work within some area of cognitive science or cognitive neuroscience. The workshop provides students with an intense two-week introduction to Penn's perspective on these emerging disciplines. Penn faculty offer day-long seminars and labs in their area of specialization, permitting in depth discussion of a particular research topic. Each week ends with a panel discussion by the faculty, relating the topics of that week, and providing spontaneous discussion of where the field of cognitive science might be headed in the coming years. By acknowledging and educating students that cognitive science is still an ill-defined rapidly changing field, we stimulate students to learn more about cognitive science, and entice them to contribute to its development and definition.

**This page left blank intentionally.**

# Scientific Explanation, Systematicity, and Conceptual Change

## **Organizer and Chair: David R. Kaufman**

Cognition and Development, Graduate School of Education  
University of California, Berkeley; Berkeley, CA, 94720  
email: davek@socrates.berkeley.edu

## **Speakers: Stella Vosniadou**

Department of History and Philosophy of Science  
National and Capodistrian University of Athens; Athens, Greece  
email: svosniad@athena.compulink.gr

## **Andy diSessa**

Cognition and Development, Graduate School of Education  
University of California, Berkeley; Berkeley, CA, 94720  
email: disessa@soe.berkeley.edu

## **Paul Thagard**

Philosophy Department  
University of Waterloo: Waterloo, Ontario, N2L 3G1  
email: pthagard@watarts.uwaterloo.ca

## **Introduction**

Humans possess remarkably rich and adaptive conceptual knowledge systems that enable them to form relatively stable representations about the world, perceive coherence amidst noise and chaos, and communicate elaborate explanations to others who see the world in strikingly similar ways. On the other hand, knowledge can sometimes be surprisingly brittle and context-bound, coherence may be more illusory than real, and individuals (e.g., teachers and students) may repeatedly fail to achieve common ground during routine discourse. How can we account for such apparent contradictions? Conceptual change names a family of theories, methodological approaches, and research traditions concerned with the origin, ontogenesis, and evolution of knowledge systems as a result of formal and informal learning. Conceptual change is the subject of considerable research across all of the cognitive sciences. In particular, it is central to investigations in the philosophy of science, cognitive development, and science education.

The speakers in this symposium will address issues in conceptual changes as they pertain to children, students learning science, lay adults, and practicing scientists. They will consider philosophical, developmental, computational, and instructional issues related to the characterization of systematicity and coherence in scientific explanation. The participants will offer distinct and sometimes divergent points of view on conceptual change with particular

attention to the reasons and mechanisms that produce systematicity and coherence (and alternatively incoherence) within and across individuals in generating scientific explanations. The speakers will address a range of related questions, including the following:

1. How can we characterize the state of knowledge structures prior to formal learning? What happens to students' knowledge when it makes contact with formal learning?
2. What are the knowledge elements that undergo change in conceptual change (e.g., beliefs, theories, schemata, propositions, and coordination classes)? What constitutes evidence for such changes?
3. What are "common" or "typical" trajectories in conceptual development (e.g., from atheoretical to theoretical, incoherent to increasingly coherent)? How can we account for periods of stability and instability in the generation of scientific explanations?
4. What are the mechanisms of change (e.g., differentiation, belief revision, enrichment, conceptual combination, re-organization and reprioritization of knowledge elements)?

5. What factors or criteria contribute to the acceptability, plausibility, and overall appraisal of scientific explanations in children, lay people, and scientists?
6. How can we expand the scope of conceptual change research to incorporate emotional and motivational variables?

Research in the philosophy of science, cognitive development, and science learning has several interesting points of convergence, despite the fact that they constitute different research programs. Philosophy of science is a discipline devoted to analyzing the character of scientific investigations (Bechtel, 1988). It endeavors to answer questions such as what constitutes a valid scientific explanation and how do scientific theories change over time. Conceptual development research is devoted to the study of age-related transitions in domain-specific (e.g., physics, biology) understandings. Conceptual change investigations in science education focus on a) characterizing transformations in learners that (with varying success) result in transformations in understanding of scientific phenomena and b) promoting instructional situations that increase the likelihood of robust and generative understanding.

Each of these disciplines is focally concerned with changes in knowledge systems that go well beyond mere knowledge accretion or belief revision. There is general agreement that conceptual change necessitates a substantial reorganization of knowledge. The history and philosophy of science (HPS) has had an enormous influence on both cognitive development and science education research (Brewer, Chinn, & Samarapungavan, 1998). HPS has provided an explanatory vocabulary for characterizing changes in scientific understanding and criteria for evaluating the quality of explanations. It has also served to highlight the fundamental commonalities underlying the conceptual change process and has led to some strong claims about the deep structural similarities between children (or naïve students) and practicing scientists. Clearly, not every theorist views the “scientist as child” metaphor as equally illuminating. In fact, each of the participants in this symposium has been critical of this perspective. Nevertheless, this point of view serves to introduce some important distinctions about the “theoretical character” of conceptual learning.

Theory theory proponents claim that there are deep similarities between scientists and children in the formation of theories (e.g., Gopnik & Wellman, 1994). Children’s naïve theories embody causal notions, enable distinct types of interpretations, explanations, and predictions, and are similarly subject to processes of modification and revision as the evidence dictates. The process of conceptual change in children is very similar in character to the process of theory revision in science. Vosniadou (1994) views conceptual

change in children and science students as differing substantially in character from scientific theory change in that children lack systematicity, abstractness, and metaconceptual awareness (i.e., understanding the hypothetical nature of their beliefs). She proposes the notion of framework theories, which consists of basic presuppositions about the way the world works and serves to constrain the acquisition of science concepts. These framework theories guide children’s interpretation of scientific phenomena and enable them to generate scientific explanations and predictions in a reasonably consistent fashion (Ioannides & Vosniadou, submitted). These “theories” are continuously enriched, differentiated, and revised as children encounter new information. However, when framework theories come into contact with formal science instruction, fragmentation, incoherencies, and misconceptions are often the result.

diSessa (1993) begins with the premise that naïve understandings of the physical world constitute a rich, complex, and diverse knowledge system. However, the system as a whole is only weakly organized and students’ intuitive scientific understandings are often a fragmented, loosely connected, collection of ideas, having none of the commitment or systematicity attributable to theories. The elements of knowledge called “p-prims” reflect minimal abstractions from common experience. Through learning and instruction, p-prims get tuned to newer contexts, refined, and reprioritized as the knowledge system is reorganized. They become supplanted in many contexts by more complex explicit knowledge structures that include physical laws. However, p-prims continue to exert substantial influence even in the reasoning of experts. Growth in scientific understanding involves a major structural change toward systematicity. Recently, diSessa and Sherin (1998) introduced the notion of coordination class, which involve systematically connected ways of gaining information from the world. Coordination classes include strategies of selective attention and systematic integration of observations.

In characterizing the nature of change in the history of science, Thagard (1992) identifies degrees of conceptual reorganization, ranging from belief revision to wholesale changes in the organizing principles underlying a conceptual system. For example, Darwin’s theory of natural selection redefined the classification of organisms according to historical lineage rather than feature similarity. The theory of explanatory coherence (instantiated in a connectionist model, ECHO) is integral to understanding the differential evaluation of competing hypotheses for best explanation and more generally, the process of conceptual change/theory adoption in science. The theory provides a set of principles (e.g., symmetry, simplicity, and data priority) that establish relations of coherence and incoherence between propositions. Thagard has used the theory of explanatory coherence, as instantiated in ECHO, to model numerous theoretical

disputes in the history of science. Thagard (1992) also considered whether conceptual change is similar in scientists and children. His analysis of the kinds of epistemic changes and process of “theory revision” reported in the developmental literature suggests that they are not typically characteristic of the kinds of dramatic changes evidenced in scientific conceptual revolutions. Thagard has also considered how other forms of coherence such as analogical, deliberative, and most recently, emotional coherence affect argumentation and theory change (in press).

## **Conceptual Change in Science Learning: From Coherence to Fragmentation**

Stella Vosniadou

Accounts of the knowledge acquisition process have customarily assumed that knowledge acquisition proceeds in a continuous manner enriching initially fragmented conceptual structures and making them increasingly more systematic, and more coherent. In this paper I will try to develop a different point of view based on a series of empirical studies investigating the development of science concepts. More specifically, the following arguments will be made with respect to a) the nature of children’s initial conceptual structures and b) the process of conceptual change.

*Initial conceptual structures:* There is considerable agreement in the cognitive science and science education literature that by the time children go to school they have acquired considerable knowledge about the physical world (an intuitive physics) that exerts considerable influence on subsequent learning and particularly on learning science. Researchers disagree, however, on the exact nature of such an intuitive physics. One view, expressed by diSessa (1988) is that initial knowledge structures about the physical world consist of an unstructured collection of small knowledge elements, which he calls phenomenological primitives (p-prims). These pieces of knowledge are generated as abstractions of common phenomena and are activated in certain characteristic cases. According to this view the process of conceptual change is one of collecting and systematizing the fragments of knowledge into consistent wholes. This happens as p-prims change their function in order to be integrated into the scientific framework.

Unlike the above view of knowledge acquisition, a number of empirical studies investigating the process of knowledge acquisition in science conducted in our lab, show that preschool children answer questions about force, matter, heat, the earth, etc., in a relatively consistent way, revealing the operation of a common explanatory framework

(Vosniadou & Brewer, 1992; 1994; Ioannides & Vosniadou, 1991; submitted). These results are consistent with research on conceptual development in infancy showing that the process of knowledge acquisition starts immediately after birth and proceeds in an orderly fashion towards the construction of an initial framework theory of physics that allows children to function adequately in the physical environment. The term theory is used here to denote a causal, relational explanatory structure and not an explicit, well formed and socially shared scientific theory. In other words, the empirical results support the hypothesis that children’s initial conceptual structures are not as fragmented as initially thought, but rather, children start the knowledge acquisition process by forming rather narrow but nevertheless internally consistent explanatory frameworks.

*The process of conceptual change:* The results of the empirical studies mentioned above show that the process of conceptual change is a slow and gradual affair that happens over a long period of time. During this process, we do not observe a change from fragmentation to increased coherence. Rather, the initial explanatory structures become more fragmented as aspects of the scientific theory are assimilated into the framework theory either creating synthetic models (which are internally consistent but scientifically wrong) or internally inconsistent structures.

In order to better understand the debate regarding coherence vs. fragmentation we should take into consideration the fact that most of the studies that support the argument that knowledge of physics’ concepts consists of “disconnected knowledge fragments” (see also Reif & Allen, 1989), are studies of older students (either college students or late high school students). On the contrary the arguments that support the coherence to fragmentation view (e.g., Vosniadou, 1994) are based on experimental evidence coming from younger children. The argument advanced here is that both of the experimental findings are correct, but that the fragmentation observed in older children and naïve adults represents a change from initial coherence to fragmentation. This is the case because a) initial conceptual structures are much more cohesive than originally thought, and b) because science instruction proceeds by fragmenting initial explanatory frameworks without succeeding in building an alternative cohesive scientific explanatory framework. Although increased fragmentation appears to be the common result of much science instruction, the process of knowledge acquisition does not stop there. The students that proceed to become experts acquire increasingly less fragmented and more cohesive science concepts.

## A Complex Adaptive Systems View of Conceptual Change

Andrea A. diSessa

Although it seems largely unacknowledged in the conceptual change community, there is a huge diversity of views about basic issues in conceptual change. Even the seemingly innocuous question “what changes in conceptual change (and what does not)?” leads to a plethora of views invoking not obviously commensurable explanatory constructs such as concepts, beliefs, models, ontologies, ontological commitments, nodes and links, schemata, and so on. One of the major fault lines in the community concerns whether conceptual change is *localizable* (e.g., in a few discrete entities, such as concepts), or whether, in contrast, it is more appropriate to think of conceptual change as emergent within a *complex system*, implicating many types of mental entities and many possible configurations. Localizability will be the focal issue of this talk.

I propose to explain and advocate the “complex adaptive systems view” (CASV) of conceptual change beginning by motivating the CASV approach methodologically. If we are to settle issues such as the localizability of conceptual change, it is imperative that we announce and debate standards for explanation within this research area. I put forward a beginning set of standards:

- Theoretical accountability — Accounts of conceptual change should employ technically well-developed explanatory constructs (e.g., concept, ontology, etc.). Dictionary definitions don’t come close to the level of specificity one needs in scientific accounts of cognitive development. At a minimum, an appropriately developed explanatory construct should allow distinguishing, in principle, between instances and non-instances of such a construct. One should expect, sooner or later, that accounts of such constructs include specification of the processes of normal deployment the construct and processes of change.
- Empirical accountability — Although a wide range of empirical methods are appropriate for studying conceptual change, we may still hold some general principles. In particular: (a) we should expect empirical work to include, in some measure, process data that can confirm or disconfirm assumptions about theoretical entities and processes hypothesized to be involved in conceptual use and conceptual change; (b) we should expect sufficient breadth of experimentation to allow limits of contextual dependence of both the construct involved and particular instances of the construct. Both (a) and (b) are plausible general accountabilities. The

vast majority of conceptual change studies collect no process data, and (b), *contextuality*, is particularly important in settling localizability questions.

I will exemplify these principles in two ways: First, I will introduce some particulars of my own CASV approach, including (a) two claimed-to-be well-rationalized explanatory constructs, *p-prims* (diSessa, 1993) and *coordination classes* (diSessa & Sherin, 1998), and (b) examples of process and other data that support theoretical claims and entailments. P-prims constitute a large class of simple “intuitive” schemata, and most directly limit claims of localizability in conceptual change. A coordination class is a model of a large-scale knowledge system constituting (a step toward) a technically precise and cogent definition of a particular class of concepts. Coordination classes define a number of obvious partial constructions of a full “concept,” which, once again, provides opportunities to examine localizability of conceptual development empirically.

Finally, I will enter into an abbreviated “competitive argumentation” comparing the success of this version of CASV with excellent recent work by Vosniadou and Ioannides (Ioannides & Vosniadou, submitted) studying the same conceptual terrain, mechanics—work that is, however, open to criticisms on the basis of the above principles. In particular, we have recently begun to gather and analyze data to bridge age-of-subjects and other methodological differences that has, so far, kept comparisons from being as compelling as they might be. We are gathering data over the same age ranges as Vosniadou and Ioannides, using similar methods, but with a slightly more open protocol to allow better contextuality analysis and some relevant process data. By the time of the conference, we should have preliminary results that bear directly on the cogency of Vosniadou’s theoretical frame, and, in particular, on the localizability of conceptual change concerning the concept “force.”

## Emotional Coherence in Scientific Explanation and Conceptual Change

Paul Thagard

Scientists are supposed to be dispassionately rational, but they are as emotional as other people. Theories are not only accepted or rejected: sometimes they are loved or hated. Good theories are often praised for their beauty and elegance, while bad theories are sometimes derided as ugly or crazy. This talk will interpret the cognitive-emotional judgments of scientists in terms of a recently developed theory of emotional coherence (Thagard, in press). My earlier work used a computational model of explanatory coherence to explain the acceptance and rejection of hypotheses on the

basis of the degree to which they satisfy a set of constraints defined in terms of explanation and evidence (Thagard, 1992). In line with much recent research in psychology and neuroscience on the ubiquity of emotions in cognition, my new model incorporates emotion into coherence computations and shows how emotional judgments can emerge from explanatory and other kinds of inference. The diversity and intensity of reactions to controversial theories such as evolution by natural selection can be explained by the theory of emotional coherence. Theory change is in part a matter of emotional change, as scientists shift their emotional attitudes toward hypotheses from positive to negative and vice versa. For example, in recent years most gastroenterologists have shifted their attitudes concerning the bacterial theory of ulcers from feeling it was ridiculous to viewing it as powerful and exciting (Thagard, 1999).

Conceptual change is also an emotional as well as a cognitive process. Concepts are mental representations corresponding to words, whereas propositions are mental representations corresponding to sentences. Both kinds of mental representations usually have emotional valences attached to them. For example, the valence of "baby" is typically positive, and the valence of "garbage" is typically negative. Conceptual change is emotional change when it involves a shift in valence from positive to negative or vice versa. In the Darwinian revolution, for example, many people shifted the valence attached to "evolution" from negative to positive. I will describe how the theory of emotional coherence can account for this aspect of conceptual change.

## References

- Bechtel, W. (1988). *Philosophy of science: An overview for cognitive science*. Mahwah, NJ: Erlbaum
- Brewer, W. F., Chinn, C. A. & Samarapungavan, A. (1998). Explanations in scientists and children. *Minds and Machines*, 8, 119-136.
- diSessa, A. (1988). Knowledge in pieces. In G. Forman & P.B. Pufall (eds.), *Constructivism in the computer age*. Hillsdale, NJ: Erlbaum.
- diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, 10(2 & 3), 105-225.
- diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? *International Journal of Science Education*, 20(10), 1155-1191.
- Gopnik, A. & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (eds.), *Mapping the mind: Domain specificity in cognition and culture*. New York: Cambridge University Press.
- Ioannides, C. and Vosniadou, S. (1991) *The development of the concept of force in Greek children*. Paper presented at the biennial meeting of the European Society for Research on Learning and Instruction, Turku, Finland, August.
- Ioannides, C., Vosniadou, S. (Submitted). *The changing meanings of force: A developmental study*.
- Reif, F. & Allen, S. (1989). Interpreting and teaching scientific concepts: A study of acceleration. *Cognition and Instruction*. 9, 1-44.
- Thagard, P. (in press). *Coherence in thought and action*. Cambridge, MA: MIT Press, fall 2000.
- Thagard, P., (1992). *Conceptual revolutions*. Princeton: Princeton University Press.
- Thagard, P., (1999). *How scientists explain disease*. Princeton: Princeton University Press.
- Vosniadou, S. (1994) Capturing and modelling the process of conceptual change *Learning and Instruction* 4, 45-69.
- Vosniadou, S. & Brewer, W.F. (1992). Mental models of the earth: A study conceptual change in childhood. *Cognitive Psychology* 24, 535-585.
- Vosniadou, S. & Brewer, W.F. (1994). Mental models of the day/night cycle *Cognitive Science* 18, 123-183.



# THE FUNCTION OF THE CEREBELLUM IN COGNITION, AFFECT AND CONSCIOUSNESS: EMPIRICAL SUPPORT FOR THE EMBODIED MIND

## Introduction

Natika Newton (nnewton@suffolk.lib.ny.us)

Department of Philosophy  
Nassau Community College  
Garden City, NY 11530

A growing movement in cognitive science views consciousness and cognition as self-organizing systems involving emotion and sensory-motor agency (e.g. Damasio 1994, 1999; Clark 1996; Glenberg, 1997; Hurley 1998). The view that cognition is best understood as *embodied* is replacing models involving amodal symbol systems like the arbitrary, intrinsically meaningless symbols of computer programs, which notoriously fail to explain common-sense reasoning and consciousness. The embodied-cognition approach sees such behavior as extensions of the animal's value-laden interaction with its environment.

How can abstract reasoning (e.g. logic and mathematics) make use of bodily action abilities? Briefly: sensorimotor imagery, conscious or semiconscious activated memory traces of the experiences of performing basic actions, functions not only in action contemplation and planning but also in the mental manipulation of objects in abstract reasoning. Abstract thought builds on basic action schemas: bodies interacting with objects in space (e.g. Huttenlocher 1968). To those claiming to lack such imagery, it can be argued that such images are not necessarily fully conscious, and brain imaging studies are now available that can decide such matters.

Actions require motivation. Even covert attention shifts depend on emotional interests of the organism; subcortical structures such as the amygdala, hippocampus and the hypothalamus influence voluntary attention mechanisms in the anterior cingulate. Actions imagined but not performed are both activated and inhibited in the frontal lobes and motor cortex; inhibition, controlled in large part by the hypothalamus, allows action images to be consciously experienced (Jeannerod 1998) along with the emotional values associated with the actions.

The combination of the above approach with recent work on emotion is powerful, allowing the formation of a global theory of brain function in which dynamic interactions among brain areas and brain events can be mapped at many levels of organization. An important prediction of the approach is that brain mechanisms once thought devoted to motor activity are also active in emotional and cognitive activities. Our example is the cerebellum. As we shall see, it appears that the cerebellum is not only a coordinator of motor actions, but also of reasoning and, most recently discovered, of emotional with cognitive states. If reasoning and other cognitive activities make use of motor schemas, this is exactly what one would expect. The cerebellum appears to be not just an organ for the coordination of actual motor activities, but also for coordinating the output of both cortical and subcortical structures involved in affect-laden cognitive activity at all levels.

## The Role of the Cerebellum in Cognition and Affect

Jeremy Schmahmann (schmahmann@helix.mgh.harvard.edu)

Associate Prof. of Neurology  
Harvard Medical School, Mass. General Hospital  
Boston, MA 02114

Many studies suggest that the cerebellum is essential to the neural circuitry subserving cognition and emotion. It connects with the reticular system (arousal), hypothalamus (autonomic function and emotional expression), limbic system (experience and expression of emotion), and paralimbic and neocortical association areas critical for higher order function (cognitive dimensions of affect). Behavioral changes in adults and children with focal cerebellar lesions provide clinical support for the relationship between the cerebellum and cognition. A cerebellar cognitive affective syndrome in adults and children is defined by impairments in executive, visual spatial, and linguistic function and dysregulation of affect. A cerebellar role in the modulation of aggression and mood appears in children with the posterior fossa syndrome following surgery involving the vermis, and during clinical and experimental neurosurgical manipulation. Functional imaging studies reveal cerebellar involvement in nociception, autonomic behaviors, affective experiences, and multiple cognitive paradigms. These suggest topographic organization in the human cerebellum with the somatosensory homunculus in the anterior lobe, cognitive operations in the neocerebellum in lobules VI and VII, and emotion particularly influenced by the vermis. We have extended the hypothesis that the phylogenetically older fastigial nucleus, vermis and flocculonodular lobe constitute the "limbic cerebellum" to include these structures in the Papez circuit.

The cerebrocerebellar system appears to consist of discretely organized parallel anatomic subsystems that serve as substrates for differentially organized functional subsystems. We have proposed that there is a universal cerebellar transform (UCT), possibly error detection, prevention, and correction utilizing an internal model that facilitates the production of harmonious motor, cognitive, and affective/autonomic behaviors: the cerebellum detects, prevents, and corrects mismatches between intended and perceived outcome of interactions with the environment. Disruption of circuitry linking the cerebellum with the cerebral hemispheres prevents cerebellar modulation of functions subserved by the affected subsystems, and produces dysmetria, the universal cerebellar impairment (UCI). Dysmetria of movement, or ataxia, is matched by "dysmetria of thought", the proposed fundamental mechanism underlying disorders of intellect and emotion resulting from cerebellar dysfunction, including the cerebellar cognitive affective syndrome, abnormalities of affect, and psychotic thinking.

### **The Integrative Role of the Cerebellar Vermis in Cognition and Emotion**

**Carl Anderson (carl\_anderson@hms.harvard.edu)**

Developmental Biopsychiatry Research  
Program, and Brain Imaging Center,  
McLean Hospital, Belmont, Mass. 02178

The cerebellar vermis and the fastigial nucleus are treated here as the "the cerebellar vermis-fastigial nucleus" (VFN) complex. The fastigial nucleus, phylogenetically the oldest of the four cerebellar nuclei, influences eye movements, posture, equilibrium and autonomic activity (Beitz, 1982). The role of VFN complex in cognitive/emotional synergy is supported by imaging studies and by observations of its focal and myriad interconnections among key brainstem nuclei and bi-hemispheric networks of motor and limbic system structures. The neocerebellar hemispheres, lateral extensions of the VFN, are activated during mental imagery, tactile learning and language and sensory-processing. Although motoric aspects of articulation, balance and bimanual coordination may involve more medial areas of the cerebellum or VFN, this region is also nonspecifically activated during many of the same functional imaging studies of neocerebellar activation. This apparent lack of specificity hints at the role of the VFN in consciousness.

As it bridges the hemispheres, pathology of the VFN appears to connect many apparently unconnected psychiatric disorders. Although the cerebellum occupies only 10% of the human brain, it contains more than half of its neurons (except in the psychiatric disorders detailed below where the VFN is significantly smaller). Among mammals, *Homo Erectus* relies most heavily on the VFN for integrating visual, vestibular and proprioceptive cues. Bipedal standing robustly activates anterior and posterior vermal regions (Ouchi et al., 1999). Destruction of the vermal cortex (Sullivan et al., 2000), a result of chronic alcoholism, produces out-of-balance staggering gait ataxia through inflexible coordination of visual, vestibular and proprioceptive feedback. Childhood trauma appears to result in pervasive cerebellar damage. This may be due to the protracted postnatal ontogeny of the cerebellum rendering it sensitive to early corticosteroid exposure (Lauder, 1983). In this case VFN pathology is associated with the development of limbic seizures (Heath, 1976; Strain et al., 1979; Cooper et al., 1974, 1985; Riklan et al., 1976). An association between the VFN and limbic seizures was first observed in electrical recordings from the hippocampus and fastigial nucleus of violent adult Harlow monkeys (Heath, 1972). Aggressive behavior in these animals resulted from the stress of total maternal deprivation (Harlow, 1971). Bremer (1997) found that lesions of the vermis, but not the cerebellar hemispheres, tamed their aggression, suggesting that deprivation disordered the developing vermal cortex. Heath accordingly used electrical stimulation of the vermis (which inhibits cortex and disinhibits the fastigial nucleus) to relieve psychotic symptoms in humans (Heath, 1980). Interestingly, research by Mason and Harlow (1975) has shown that rocking during early life, which stimulates the VFN, mitigates the adverse effects of maternal deprivation. Child abuse is also associated with dissociation, increased prevalence of abnormal EEG=s (Ito, 1998), and symptoms suggestive of limbic seizures (Teicher et al., 1993). It seems to cause a limbic "kindling" that produces epilepsy in experimental animals. Repeated electrical stimulation of the limbic system in experimental animals can lead to seizures. Repeated abuse in humans may also result in limbic electrical abnormalities associated with epileptic-like behavioral experiences. Electrical stimulation of the VFN in humans suppresses the spread of epileptic seizures (Cooper et al., 1974; Cooper and Upton, 1985). We used fMRI to assess the relationship between behavioral measures of limbic kindling and blood volume in the VFN of young adults with a history of childhood abuse and found a strong correlation between VFN blood flow and kindling (Anderson et al., 1999; Teicher et al., 1993); early abuse seems associated with a functional deficit in limbic-VFN networks. Converging data suggest involvement of VFN abnormalities in various disorders including depression (Fischler et al., 1996; Lauterbach, 1996; Beauregard et al., 1998), schizophrenia (Loeber et al., 1999; Jacobsen et al., 1997),

autism (Courchesne et al., 1991) and ADHD (Berquin et al., 1998). The VFN innervates the locus coeruleus (LC), ventral tegmental area (VTA), substantia nigra (SN), and midline raphe, cell body regions of the dopaminergic, noradrenergic and serotonergic pathways (Reis & Golanov, 1997; Snider & Maiti, 1976; Snider et al., 1976). We also found a strong dose-dependent effect of methylphenidate on blood flow in the VFN of ADHD children (Anderson et al., 2000).

Pettigrew (1998) shows that binocular rivalry occurs between, not within, cerebral hemispheres, and that the rate of perceptual rivalry is slow in bipolar disorder. Pettigrew's interhemispheric "sticky switch" in manic depression could be due to VFN pathology observed in bipolars (Lauterbach, 1996). The VFN complex projects to pons and reticular formation sites where network cascades are easily activated. Fastigial electrical stimulation desynchronizes EEG, characteristic of the behavioral states of REM sleep and attention orienting. Snider (1976) demonstrated direct and indirect projections from the VFN to the LC (Ruggiero et al., 1997) and paraventricular nucleus (Astier et al., 1990) and parabrachial nuclei (Supple & Kapp, 1994). All regions that could contribute to desynchronized EEG and facilitate cortical binding. Usher (1999) proposes that electrotonic coupling within the LC plays a role in attentional modulation and regulation of goal-directed versus exploratory behaviors. This electrotonic syncytium structure may represent the fingerprint of dynamical internal models in the cerebellum (Imamizu et al., 2000). Makarenk (1998) demonstrates that synchrony of inferior olive neurons arises from chaotic subthreshold oscillations. These neurons, while having maximum functional permissiveness, can also transform rapidly into robustly determined functional patterns of multicellular coherence. Along with a hypothesis of how pontine organization may be controlled by the cerebellum for binding assemblies of cortical-striatal-thalamic loops into coherent motor strategies (Schwarz & Theier, 1999), these findings suggest that consciousness (in terms of motor patterns) may develop in the spatial-temporal chaos of phase synchronized pontine olive ensembles.

Drug addiction is also associated with early child abuse, ADHD and bipolar disorder. A potent anti-addiction drug, ibogaine, causes hallucinations, cerebellar tremor, transient ataxia, and vermal lesions in rats. Ibogaine strongly activates climbing fiber activity in VFN Purkinje cells (O'Hearn & Molliver, 1997). Hallucinations associated with ibogaine treatment indicate impairment of the cerebellar efferent copy mechanism. Subjects may then sleep for 24 hrs and awaken free from addiction (Kovera et al, 1999), suggesting a connection between ibogaine action and changes in the fractal organization of REM sleep (Anderson, 1998). The cerebellar vermis is most active during REM, especially in human infants (Chugani, 1998). Schlesinger et al. (1998) found that REM deprivation implicated VFN function in postural control and attention. Interpreted in the context of vertically convergent fractal time processes, early stress appears to disrupt organization of fractal REM patterns, leading to alteration of patterns of nuchal atonia occurring during REM sleep in fetuses and neonates (Anderson, 1995; Anderson et al., 1998); suggesting a connection between REM sleep, the VFM and cognitive/emotional synergy. Our hypothesis is that early stress results in pathological fastigial regulation of chaotic spatial-temporal patterns of pontine olive ensembles. As the VFM appears to influence the architecture of REM sleep, abnormal spatial-temporal patterns lock into a negative feedback loop, with further disorganization of pontine olive ensembles. Ibogaine, by overdriving VFM climbing fiber activity, breaks these cycles, resulting in REM rebound and the flooding of abuse memories as efferent copy mechanisms are reset during the treatment. Thus the VFN may represent a key node in the interface of limbic-brainstem network oscillations and self-organized attentional/orienting mechanisms with dynamic internal motor schemas during our ongoing emotion-laden experience of consciousness.

**Consciousness and the Cerebellum**  
**Ralph D. Ellis (ralphellis@mindspring.com)**  
Department of Philosophy  
Clark-Atlanta University  
Atlanta, GA 30314

If the cerebellum plays a coordinating and synchronizing role in the brain, this role must be an important determinant of the structure of conscious processing, because conscious processing is a whole brain activity of a self-organizing system. That consciousness does not result from passive stimulation is clear since occipital activity alone does not yield perceptual consciousness (Aurell 1989; Posner & Rothbart 1995; perceptual studies by Mack & Rock 1998 also entail this). Moreover, cerebellar activity is similar for abstract thought, physical movement, or imagination of physical movement (Ito 1993); this suggests that the cerebellum

subserves intellectual operations, and that intellectual understanding is an extension of manipulation of action affordances (Newton 1996).

Wakefulness results from synchronization of wave patterns in diverse regions, especially between the thalamus and the cortex (Asimov 1965). Since the cerebellum controls widely distributed synchronizations, it is crucial for the difference between sleep and wakefulness; hence again, for consciousness. Equally important, coordination between hippocampus and other subcortical regions, and the effect of this coordination on the extended reticular thalamic activating system (ERTAS), which attunes the thalamus to select for incoming stimuli with emotionally important action affordances, is also needed for *perceptual* consciousness. Occipital processing alone is not conscious; other areas, including anterior areas such as anterior cingulate and frontal and parietal lobes; must be activated in response to input from emotional brain areas; since this activity involves widespread synchronization, the cerebellum also plays a part. If consciousness begins with action affordances, or Damasio's (1999) "as if body loop," then the cerebellum is pivotal for consciousness.

The hippocampus shows an event related potential within 20 ms. of a perceptual stimulus (Coles 1990), indicating subcortical activation with wave synchronization phenomena, this implies cerebellar involvement as well: the first occipital ERP does not begin until around 100 ms. Rather than anterior and subcortical activation's being a *response* to an *occipital stimulus*, this activation must *already* have taken place prior to consciousness. ERTAS, guided by emotional subcortical purposes, *determines* registration of perceptual input in consciousness.

When a visual stimulus is unexpected, there is a 1/4 second delay from occipital processing to the other brain processes needed for perceptual consciousness (Srebro 1985) — too long to be explained by the speed of spreading activation. The delay allows emotional areas to activate thalamus, frontal and parietal areas in response to hippocampal and cerebellar tuning toward relevant action affordances; this "looking for" activity has already begun *prior to* occipital effects on perceptual consciousness (since the occipital P200 has not yet occurred). Even in *involuntary* attention and in cases of frontal lobectomy, the limbic system selectively gates incoming stimuli according to general motivational purposes *via* subcortical control of neurotransmitters, tuning the thalamus (Damasio 1999; Faw forthcoming) and allowing consciousness rather than blindsight. Persons with no anterior cingulate are in vegetative states (Damasio 1999), so even involuntary attention is impossible without it.

This reverses traditional thought about the causal ordering of brain events: perception was thought to drive emotion, which in turn drove action. Instead, the organism must first be geared up to seek important data, the most fundamental of which involve action and thus cerebellar functions. Subcortical tuning activates frontal and limbic regions to form preconscious image schemas associated with important perceptual categories, prior to processing of the stimulus. If the stimulus resonates with this self-generated activity, a more vivid image is formed, and one that is felt as the image of a present perceptual object rather than as an image (Aurell, 1989).

In tracking a soccer ball, expectation is at each moment motivated by categories of utility and retention of the ball's previous location. When the ball suddenly turns up where we are not looking, we do not see it, but have a vivid image of where it *should have been*. It then takes a fourth of a second to find the ball. It catches attention by presenting affordances to the motivated organism. Self-organization must be holistically coordinated, requiring cerebellar synchronizations. Since perception is motivated by utility (Newton 1996) and frontal and parietal areas are tuned by emotional areas (hypothalamus, hippocampus, cerebellum, amygdala), we see the sinister smile without noticing its sinister details; we note a room's disorder but not the crooked picture frame that makes it disorderly (Merleau-Ponty, 1942, p. 173).

Emotion is not sufficient for consciousness. Plants and low animal species have organismic purposes, but little consciousness. Consciousness occurs only when emotion combines with representation, occurring not passively but as an activity of the organism. Emotional agnosics can't represent what emotions are "about." We are conscious of emotions through representation. Even unconscious emotions still drive the representational processes in which we do engage, and even pure curiosity is an emotion that motivates us to explore our environment and represent what is there (Panksepp 1998).

### References

- Anderson, C. (1995). *Spontaneous perinatal behaviors associated with REM sleep: possible ontogenetic adaptation and source of plasticity underlying emergence of behavioral neophenotypes*. Ph.D., Florida Atl. Univ.
- Anderson, C.M., Maas, L.C., Renshaw, P.F., & Teicher, M.H. (2000). Methylphenidate Dose-Dependently Alters Blood Flow in the Vermis But Not Basal Ganglia of ADHD Boys. *Biological Psychiatry*, 47, 106S.

- Anderson, C.M., Mandell, A.J., Selz, K.A., Terry, L.M., Wong, C.H., Robinson, S.R., Robertson, S.S., & Smotherman, W.P. (1998). The development of nuchal atonia associated with active (REM) sleep in fetal sheep: presence of recurrent fractal organization. *Brain Research*, 787(2), 351-7.
- Anderson, C., Polcari, A., McGreenery, C., Maas, L., Renshaw, P., & Teicher, M. (1999). Cerebellar vermis blood flow: associations with psychiatric symptoms in child abuse and ADHD. *Soc. for Neuroscience Abstracts*, 25(part 2), 1637.
- Anderson, C.M. (1998). Ibogaine Therapy in Chemical Dependency and Posttraumatic Stress Disorder: A Hypothesis Involving the Fractal Nature of Fetal REM Sleep and Interhemispheric Reintegration. *Multidisciplinary Association For Psychedelic Studies, Vol. VIII* (pp. 5-14).
- Asimov, Isaac (1965). *The Human Brain*. New York: Mentor.
- Astier, B., Van Bockstaele, E.J., Aston-Jones, G., & Pieribone, V.A. (1990). Anatomical evidence for multiple pathways leading from the rostral ventrolateral medulla (nucleus paragigantocellularis) to the locus coeruleus in rat. *Neuroscience Letters*, 118(2), 141-6.
- Aurell, Carl G. (1989). Man's triune conscious mind. *Perceptual and Motor Skills*, 68: 747-754.
- Beauregard, M., Leroux, J.M., Bergman, S., Arzoumanian, Y., Beaudoin, G., Bourgouin, P., & Stip, E. (1998). The functional neuroanatomy of major depression: an fMRI study using an emotional activation paradigm. *Neuroreport*, 9(14), 3253-8.
- Beitz (1982). *The cerebellum--new vistas*. Berlin: Springer-Verlag.
- Berman, A.J. (1997). Amelioration of aggression: response to selective cerebellar lesions in the rhesus monkey. *International Review of Neurobiology*, 41, 111-9.
- Berquin, P., Giedd, J., Jacobsen, L., Hamburger, S., Krain, A., Rapoport, J. & Castellanos, F.X. (1998). Cerebellum in attention-deficit hyperactivity disorder - a morphometric mri study. *Neurology*, 50(4), 1087-1093.
- Chugani, H.T. (1998). A critical period of brain development: studies of cerebral glucose utilization with PET. *Preventive Medicine*, 27(2), 184-8.
- Coles, M., G. Gratton, and M. Fabiani (1990). Event-related brain potentials. In *Principles of Psychophysiology*. Cambridge: Cambridge University Press, pp. 413-453.
- Cooper, I., Riklan, M., Waltz, J., Amin, I., & Pani, K. (1974). A study of chronic cerebellar stimulation in disorders of sensory communication in central nervous system. *Boletín de Estudios Médicos y Biológicos*, 28(8-10), 347-90.
- Cooper, I., & Upton, A. (1985). Therapeutic implications of modulation of metabolism and functional activity of cerebral cortex by chronic stimulation of cerebellum and thalamus. *Biological Psychiatry*, 20(7), 811-3.
- Courchesne, E. (1991). Neuroanatomic imaging in autism. *Pediatrics*, 87(5 Pt 2), 781-90.
- Damasio, A. (1999). *The Feeling of What Happens*. New York: Harcourt Brace.
- Faw, B. (forthcoming). Consciousness, motivation, and emotion: Biopsychological reflections. in R. Ellis & N. Newton (eds.), *The Caldron of Consciousness: Affect, Motivation, and Self-Organization*. Amsterdam: Benjamins.
- Fischler, B., D'Haenen, H., Cluydts, R., Michiels, V., Demets, K., Bossuyt, A., Kaufman, L., & DeMeirleir, K. (1996). Comparison of 99m Tc HMPAO SPECT scan between chronic fatigue syndrome, major depression & healthy controls: an exploratory study of clinical correlates of cerebral blood flow. *Neuropsychobiology*, 34(4), 175-83.
- Harlow, H.F., & Mc Kinney, W.T., Jr. (1971). Nonhuman primates and psychoses. *Journal of Autism & Childhood Schizophrenia*, 1(4), 368-75.
- Heath, R.G. (1972). Electroencephalographic studies in isolation-raised monkeys with behavioral impairment. *Diseases of the Nervous System*, 33(3), 157-63.
- Heath, R.G., Llewellyn, R.C., & Rouchell, A.M. (1980). The cerebellar pacemaker for intractable behavioral disorders and epilepsy: follow-up report. *Biological Psychiatry*, 15(2), 243-56.
- Imamizu, H., Miyauchi, S., Tamada, T., Sasaki, Y., Takino, R., Putz, B., Yoshioka, T., & Kawato, M. (2000). Human cerebellar activity reflecting an acquired internal model of a new tool [see comments]. *Nature*, 403(6766), 192-5.
- Ito, Maseo (1993). Movement and thought: Identical control mechanisms by the cerebellum. *Trends in Neurosciences* 16, 448-450.
- Ito, Y., Teicher, M.H., Glod, C.A., & Ackerman, E. (1998). Preliminary evidence for aberrant cortical development in abused children: a quantitative EEG study. *Journal of Neuropsychiatry and Clinical Neurosciences*, 10, 298-307.
- Jacobsen, L., Giedd, J., Berquin, P., Krain, A., Hamburger, S., Kumra, S., & Rapoport, J. (1997). Quantitative morphology

of the cerebellum and fourth ventricle in childhood-onset schizophrenia. *AmJ Psych*, 154(12), 1663-9.

Kovera, C.A., Kovera, M.B., Pablo, F.R., Ervin, I.C., Williams, I.C., & Mash, D.C. (1999). Anti-Addiction Benefits of Ibogaine: Mood Elevation and Drug Craving Reduction. *Society for Neuroscience Abstracts*, 25(part 2), 1564.

Ito, M.(1993).Movement and thought: Identical mechanisms by cerebellum.*Trends Neurosciences* 16, 448-450.

Lauder,J.(1983).Hormonal & humoral influences on brain development.*Psychoneuroendocrinology*,8(2),121-55.

Lauterbach, E.C. (1996). Bipolar disorders, dystonia, and compulsion after dysfunction of the cerebellum, dentatorubrothalamic tract, and substantia nigra. *Biological Psychiatry*, 40(8), 726-30.

Loeber, R.T., Sherwood, A.R., Renshaw, P.F., Cohen, B.M., & Yurgelun-Todd, D.A. (1999). Differences in cerebellar blood volume in schizophrenia and bipolar disorder. *Schizophrenia Research*, 37(1), 81-89.

Mack, Arien, and Irvin Rock (1998). *Inattentive Blindness*. Cambridge: MIT/Bradford.

Makarenko, V., & Llinas, R. (1998). Experimentally determined chaotic phase synchronization in a neuronal system. *Proceedings of the National Academy of Sciences of the United States of America*, 95(26), 15747-52.

Mason, W.A., & Berkson, G. (1975). Effects of maternal mobility on the development of rocking and other behaviors in rhesus monkeys: a study with artificial mothers. *Dev Psychobiol*, 8(3), 197-211.

Melchitzky,D.S.,& Lewis,D.A.(2000). Tyrosine hydrolase- and dopamine transporter-immunoreactive axons in the primate cerebellum: evidence for a lobular- and laminar-specific dopamine innervation. *Neuropsychopharmacology*, 22(5), 466-72.

Merleau-Ponty, Maurice (1942/ 1963). *The Structure of Behavior*. Boston: Beacon.

Newton, Natika (1996). *Foundations of Understanding*. Amsterdam, John Benjamins.

O'Hearn, E., & Molliver, M.E. (1997). The olivocerebellar projection mediates ibogaine-induced degeneration of Purkinje cells: a model of indirect, trans-synaptic excitotoxicity. *Journal of Neuroscience*, 17(22), 8828-41.

Ouchi, Y., Okada, H., Yoshikawa, E., Nobezawa, S., & Futatsubashi, M. (1999). Brain activation during maintenance of standing postures in humans. *Brain*, 122(Pt 2), 329-38.

Panksepp, Jaak (1998). *Affective Neuroscience*. New York: Oxford

Pettigrew, J.D., & Miller, S.M. (1998). A 'sticky' interhemispheric switch in bipolar disorder? *Proc R Soc Lond B Biol Sci*, 265(1411), 2141-8.

Posner, Michael I. and Mary K. Rothbart (1992). Attentional mechanisms and conscious experience. In A.D. Milner and M.D. Rugg (eds). *The neuropsychology of consciousness*. London: Academic Press, 187-210.

Reis, D.J., & Golanov, E.V. (1997). Autonomic and vasomotor regulation. *International Review of Neurobiology*, 41, 121-49.

Riklan, M., Kabat, C., & Cooper, I.S. (1976). Psychological effects of short term cerebellar stimulation in epilepsy. *J Nerv Ment Dis*, 162(4), 282-90.

Ruggiero,D.A.,Anwar,M.,Golanov,E.V.,&Reis,D.J.(1997).The pedunculo pontine tegmental nucleus issues collaterals to fastigial nucleus and rostral ventrolateral reticular nucleus in the rat. *Brain Res.*,760(1-2), 272-6.

Schlesinger, A., Redfern, M.S., Dahl, R.E., & Jennings, J.R. (1998). Postural control, attention and sleep deprivation. *Neuroreport*, 9(1), 49-52.

Schwarz, C., & Thier, P. (1999). Binding of signals relevant for action: towards a hypothesis of the functional role of the pontine nuclei. *Trends in Neurosciences*, 22(10), 443-51.

Snider,R.S.,& Maiti,A.(1976).Cerebellar contributions to Papez circuit. *J. Neuroscience Research*, 2(2), 133-46.

Snider,R.,Maiti,A.,& Snider,S.(1976).Cerebellar pathways to ventral midbrain & nigra.*Exp.Neur.*53(3),714-28.

Srebro, R.,(1985).Localization of visually evoked cortical activity in humans. *J. Physiology* 360, 233-246.

Strain, B.M., Babb, T.L., Soper, H.V., Perryman, K.M., Lieb, J.P., & Crandall, P.H. (1979). Effects of chronic cerebellar stimulation on chronic limbic seizures in monkeys. *Epilepsia*, 20(6), 651-64.

Sullivan, E.V.,Deshmukh, A.,Desmond, J.E.,Mathalon, D.H.,Rosenbloom, M.J.,Lim, K.O., & Pfefferbaum,A. (2000).Cerebellar gray matter deficits in schizophrenia and alcoholism.*Biological Psychiatry*,47(1), 37S.

Supple, W.F., Jr., & Kapp, B.S. (1994). Anatomical and physiological relationships between the anterior cerebellar vermis and the pontine parabrachial nucleus in the rabbit. *Brain Research Bulletin*, 33(5), 561-74.

Teicher, M.H., Glod, C.A., Surrey, J., & Swett, C. (1993). Early Childhood Abuse and Limbic System Ratings in Adult Psychiatric Outpatients. *Journal of Neuropsychiatry and Clinical Neurosciences*, 5, 301-6.

Usher, M., Cohen, J.D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G.(1999). The role of locus coeruleus in the regulation of cognitive performance. *Science*, 283(5401), 549-54.

# Bayesian Approaches to Cognitive Modeling

**Joshua Tenenbaum**

*Department of Psychology  
Stanford University  
Stanford, CA 94305*

**Michael C. Mozer**

*Department of Computer Science  
University of Colorado  
Boulder, CO 80309–0430*

Many, if not all, aspects of human cognition depend fundamentally on inductive inference: evaluating degrees of belief in hypotheses given weak constraints imposed by observed data. In logic-based models of cognition, the currency of belief is a binary truth value. In connectionist models of cognition, the currency of belief is an activation level. In Bayesian models of cognition, the currency of belief is a probability. The term “Bayesian” comes from Thomas Bayes, an 18th century minister who introduced a key theorem which serves as the mathematical basis of probabilistic inference. Under the single assumption that degrees of belief be represented as probability distributions, Bayes’ theorem describes how the degree of belief in a hypothesis,  $h$ , should be updated as a result of some new evidence,  $e$ :

$$P(h|e) = P(e|h)P(h) / \sum_{h' \in H} P(e|h')P(h')$$

where  $P(h|e)$  denotes the conditional (*posterior*) probability that  $h$  is true given that  $e$  is true,  $P(h)$  denotes the unconditional (*prior*) probability that  $h$  is true, and  $P(e|h)$  denotes the *likelihood* of observing  $e$  given that  $h$  is true.  $H$  denotes a set of mutually exclusive and exhaustive alternative hypotheses that could be invoked to explain  $e$ . A Bayesian’s belief in  $h$  given  $e$  is thus a measure of how well  $h$  explains  $e$  relative to how well alternative hypotheses  $h' \in H$  explain  $e$ .

As a normative theory of inductive inference, the Bayesian paradigm provides a principled, general-purpose framework for constructing rational models of cognition across a wide range of domains (Anderson, 1990; Knill & Richards, 1996; Oaksford & Chater, 1998). This symposium will provide a forum for representatives of Bayesian approaches from various areas of cognitive science—perception, learning, reasoning, memory, and language acquisition—to discuss both the successful aspects and the open challenges of the Bayesian paradigm. Questions to be addressed include:

- How does a Bayesian analysis provide a rational explanation for phenomena that have previously been addressed by mechanistic models? When and why does Bayes predict new phenomena that mechanistic models fail to predict? When do Bayesian analyses result in emergent predictions that are not intuitively obvious from the model’s design?
- How does Bayes support the integration of disparate sources of information into a single coherent inference?
- How does Bayes allow the unification of two or more apparently distinct modes of processing into a single compu-

tational framework?

- Where does a Bayesian agent’s hypothesis space come from? What kind of extra-Bayesian assumptions are needed in deriving the probabilistic generative model (prior probabilities and likelihoods) that is the foundation of a Bayesian analysis?
- The Bayesian paradigm conceives of perception and cognition as being adapted to the structure and statistics of the environment, but the mechanisms of this adaptation may vary across domains. What are the roles of evolution, learning, and habituation in adapting a Bayesian agent to the structure of a particular domain?
- There are typically many different ways to give a Bayesian analysis of a particular task. Is there always one “correct” Bayesian model? What are the criteria for deciding that one is correct?
- How can Bayesian models be tested empirically? Is the Bayesian approach falsifiable? Should it be?
- How can we reconcile the success of Bayesian models of cognition with the well-known findings from the heuristics and biases literature that “people are not Bayesian”? Could these discrepancies reflect different ways of formulating Bayesian analyses of the same tasks?
- Bayesian models, when fully implemented, are often computationally intractable. What are the implications of this intractability for a model’s psychological or neural plausibility? What are the possibilities for principled approximations that might preserve the rigor of the approach in a more tractable setting? How might familiar, cognitively plausible heuristics be viewed as approximations to the full Bayesian competence?
- “Probability is not really about numbers; it is about the structure of reasoning” (G. Shafer, as quoted in Pearl, 1988). How might the structural aspects of Bayesian inference, as captured in Bayes nets and other graphical models, be important for understanding human cognition?

Speakers at the symposium will include: Michael Brent (*Bayesian modeling of segmentation and word discovery*), Evan Heit (*A Bayesian account of category-based induction*), Michael Mozer (*Temporal dynamics of information transmission in a Bayesian cognitive architecture*), and Joshua Tenenbaum (*Rules and similarity in concept learning*).

## The Nature of Human Errors: An Emerging Interdisciplinary Perspective

**Jiajie Zhang, Ph.D. (Chair)**

Associate Professor, Dept of Health Informatics  
University of Texas at Houston

**Vimla L. Patel, Ph.D., DSc**

Professor, Departments of Medicine and Psychology  
Director, Center for Medical Education, McGill University

**Edward H. Shortliffe, M.D., Ph.D.**

Professor and Chair, Department of Medical Informatics  
Columbia College of Physicians and Surgeons

**Michael Freed<sup>1</sup>, Ph.D. & Roger Remington<sup>2</sup>, Ph.D.**

Research Associate<sup>1</sup>, Director<sup>2</sup>, Cognition Lab  
NASA Ames Research Center

### Introduction

In light of a growing awareness of the role of human errors in widely publicized incidents such as airline accidents and complications of medical procedures, now is the right time for cognitive science to make a contribution to the study and prevention of human errors. As shown in Figure 1, human errors account for more than half of accidents in most industries. In air traffic control, the rate is over 90%. Human errors occur primarily due to inadequate information processing. As an interdisciplinary field for the study of information processing in humans and machines, cognitive science can make a significant contribution to human error studies.

In this symposium, the four presentations will address human errors from four different perspectives.

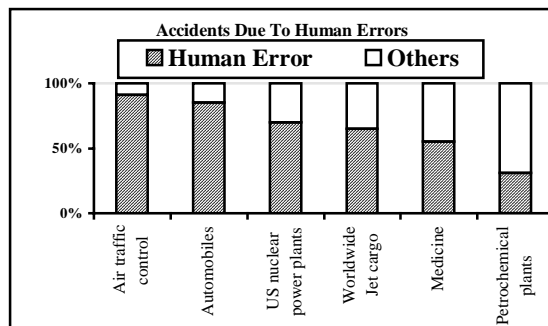


Figure 1. Accidents due to human errors

### Human Errors: Cognitive Theory & Interface Design

*Jiajie Zhang*

There are two major types of human errors (Reason, 1990): planning and execution errors. Slips are errors of execution in which the correct action does not proceed as intended. Mistakes are errors of planning in which the original intended action is not correct. This presentation will focus on four types of slips (Norman, 1981). Caption slips result from automatic activation of a well-learned routine that overrides the current intended activity (e.g., driving home directly instead of picking up a prescription on the home way). Description slips are due to incomplete or ambiguous specification of intention that is similar to a familiar intention (e.g., inserting a Zip disk to a floppy drive). Associative activation slips are due to activation of similar but incorrect schemas (e.g., picking up the desktop phone when the cell phone rings). Loss-of-activation slips are due to loss of the activation of current intention (e.g., forgetting an idea for this symposium proposal after answering an interruptive phone call). The first part of this presentation will describe a

cognitive theory of slips (based on Norman's schema theory) that attempts to explain why slips occur and predict when they occur.

The second part will be about the design of systems that minimize human errors. The cognitive theory of slips points out the causes and predicts what types of slips will happen under what circumstances. With such a theoretical guideline, we can design systems that have properties that can make certain types of slips impossible to occur or minimize the factors that can cause errors (e.g., a good user interface that minimizes mental workload).

### Conceptual and Procedural Errors in Medical Decision Making

*Vimla L. Patel*

Cognitive studies of errors in medical decision making have traditionally focused on biases and faulty heuristics that lead health professionals to fail to attend to, or properly consider, relevant data. The error is sometimes attributed to physicians' lack of competency in probabilistic reasoning. In our view, decision making is an inherently complex cognitive and social process and errors can have multiple etiologies. It is convenient to partition sources of error into three categories: 1) individual/cognitive, 2) social/communicative and 3) systemic/institutional. Errors can arise due to actions (or neglect) of a single individual. Decision making critically depends on the availability of current information, a level of understanding, and the use of appropriate decision strategies.

The most serious cognitive errors are those that arise for reasons other than simple neglect or oversight (e.g., unintended slips). Possible causes include procedural errors and faulty conceptual knowledge. In addition, several studies have documented errors due to dissociations between subjects' conceptual understanding and their application of knowledge in solving patient problems. For example, a subject may understand that certain levels of serum cholesterol coupled with other symptoms necessitate pharmaceutical intervention, but may fail to incorporate this knowledge into an action plan. Similarly, an individual may know how to carry out an effective procedure, but lack the prerequisite conceptual knowledge required to determine its suitability or to cope with problems that arise when it is being performed. This can lead to errors of over-generalization or contribute to use of an overly narrow perspective (violation of constraints).



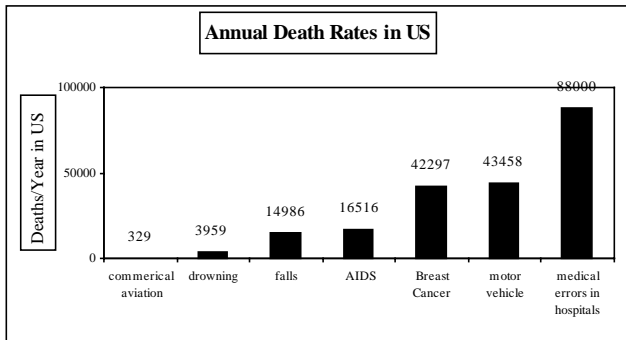
High velocity decision-making environments, such as intensive care settings, are vulnerable to multiple sources of social and communicative errors. These errors can emerge from disruptions in the flow of information such as the failure of coordination and communication between an overnight and daytime nurse who must achieve mutual understanding about the state of a patient for whom they both care. Systemic and institutional errors are caused by problems that are not due to any individual or team of individuals, but rather are caused by some fault in a system. This category may include problems with technological systems, the physical design of the workspace, or the use of institutionally sanctioned, but faulty protocols.

This presentation will consider a range of medical decision making errors, drawing on both laboratory and naturalistic studies, and will attempt to relate these errors of reasoning to issues of education and training.

**Information Technology’s Role in the Prevention of Human Errors in Clinical Medicine**

*Edward H. Shortliffe*

A recent report from the Institute of Medicine (IOM, December 1999) indicates that 44,000 to 98,000 patients die from medical errors every year in US hospitals. The study suggests that more people die from medical errors in hospitalization than from motor vehicle accidents, breast cancer, or AIDS (see Figure 2). When deaths from ambulatory settings are considered, the estimate can go much higher.



**Figure 2.** Annual death rates in US

Several industries (e.g., aviation and nuclear power plants) have been very successful in preventing human errors, perhaps because their accidents, when they do occur, make bigger headlines than medical incidents due to their catastrophic nature. For example, 329 people die a year from commercial aviation in US. In contrast, the death rate due to medical errors in the US is equivalent to one jumbo-jet crash every day. The IOM report has increased the public’s awareness of the frequency and significance of medical errors, and the Clinton Administration has authorized the creation of a Center for Patient Safety with initial funding of \$35 million a year, setting a goal of reducing medical errors by 50% in five years.

This presentation argues that there is much that the health-care industry can do to prevent the kinds of errors described in the IOM report. Many of the problems are re-

lated to inadequacies in process rather than to incompetence in health workers, and information technology can play a particularly important role in dealing with such errors. Examples include the computer-based verification of dosing information at the time that a drug regimen is ordered, or improved access to (and legibility of) pertinent clinical information that may prevent decision-making errors before they occur. Challenges in implementing and integrating such facilities into clinical environments will be discussed, along with examples of systems currently in use to address these kinds of human errors in clinical settings. The role of computer-based clinical decision-support systems will be emphasized.

**Human Error Modeling in Aviation**

*Mike Freed and Roger Remington*

In commercial aviation, as in many other demanding tasks, human error is among the most significant sources of cost and risk. One important consequence is a necessary conservatism about introducing new aviation technologies. In particular, new procedures and devices that affect the type, pace, or amount of work of an operator may inadvertently facilitate error. Designers typically assess the human performance impact of new technology by building system prototypes, training users, and then running “human in the loop” studies in which operators are observed carrying out tasks in a variety of scenarios. This tends to be very costly, limiting the amount of testing that can be done and, indirectly, the flexibility of the system to accommodate innovation and adjustment.

One solution is to develop better methods for evaluating at an early stage in design (before a physical prototype), when altering the design is inexpensive. For some domains, methods such as guideline based critiquing and cognitive walkthrough can be used to detect human factors problems in a design at this early stage. For more complex domains, computer simulation is needed to handle the vast amount of situational detail that must be considered as possibly contributing to operator error.

We will describe APEX, a tool for simulating human operators in complex, dynamic environments (so far including TRACON air traffic control and Boeing 757 flight deck). The human performance model used in APEX adapts an AI technology called reactive planning to enable capable behavior in such demanding environments. This technology turns out to be especially suited for simulating human proneness to certain forms of systematic procedural error, especially what Reason calls “frequency gambling errors.” Ultimately APEX is intended to help designers sift through thousands of possible scenarios to identify possibilities for error that might otherwise have been overlooked. To this end, we have used examples of operator error from “human in the loop” simulation studies and from reported incidents in the Aviation Safety Reporting System database to drive development of the model. We are also beginning to study procedural errors empirically in order to test the effect of certain possible causal factors.

# Induction of causal chains

Woo-kyoung Ahn

(woo-kyoung.ahn@vanderbilt.edu)

Department of Psychology, Vanderbilt University; 111 21st Avenue, Nashville, TN 37240 USA

Martin J. Dennis

(martin.dennis@yale.edu)

Department of Psychology, Yale University; 2 Hillhouse Avenue, New Haven, CT 06520 USA

## Abstract

The current study examined one way in which people learn complex causal relations from covariation. When participants were presented with covariation information between X and Y and covariation information between Y and Z only, they were willing to infer a causal relationship between X and Z, although it is not warranted by the evidence. Furthermore, the perceived strength of the terminal relationship was tied to the perceived strength of the intermediate relationships, as manipulated through the order of evidence. These results imply that people do not follow normative, contingency-based theories, but instead carry out a hypothesis-testing process and combine piecemeal relationships into an overarching causal induction.

## Introduction

Learning causal relations between two events is a fundamental cognitive activity. Although there are many ways to acquire causal knowledge (Ahn & Kalish, in press), the current study examines one way of learning causal relations, namely, through covariation.

In a simple case of covariation involving two events, a possible cause is either present (X) or absent ( $\sim X$ ), and the target effect to be explained is either present (Y) or absent ( $\sim Y$ ) as shown in Figure 1. One way to define covariation between two factors is to calculate an index  $\Delta P = P(Y/X) - P(Y/\sim X)$ , the difference between the probability that the effect occurs, given that the cause is present, and the probability that the effect occurs, given that the cause is absent (e.g., Cheng & Novick, 1992; Jenkins & Ward, 1965). In the example shown in Figure 1,  $\Delta P$  is 0.2. Numerous studies have demonstrated positive correlations between objective  $\Delta P$  and the perceived causal strengths between events (e.g., Wasserman, Chatlosh, & Neunaber, 1983).

	Y	$\sim Y$
X	6	4
$\sim X$	4	6

Figure 1. Example contingency between X and Y. Numbers show example frequencies of each evidence type.

In addition to the simple causal relations examined in these previous studies, people have knowledge about com-

plex causal mechanisms (e.g., Ahn, Kalish, Medin, & Gelman, 1995). For instance, in explaining why Kim had a traffic accident, one might refer to a mechanism of drunk driving rather than just the fact that there is a positive covariation between a traffic accident and drunk driving. Most people understand the mechanism underlying the effect of drunk driving on a traffic accident to be that when drunk, a person's motor responses are uncoordinated, in which case a person might not stay in the road, and so on. The question is, how do we acquire understanding of these causal mechanisms?

In answering this question, it is important to understand first how our knowledge about causal mechanisms might be represented. One useful tool is conditional dependencies or Bayesian networks<sup>1</sup>. The idea is that mechanisms can be represented in terms of a complex web of covariation, or more specifically, as a directed graph in which nodes representing variables are connected with arrows indicating causal directions (Glymour, 1998; Glymour & Cheng, 1998; Pearl, 1996; Spirtes, Glymour, & Scheines, 1993; Waldmann & Martignon, 1998). For instance, a mechanism underlying the covariation between drunk driving and a traffic accident might be represented as follows:

drink alcohol  $\rightarrow$  uncoordinated motor responses  $\rightarrow$  traffic accident

Glymour (1998) proposes that B is a mechanism for a correlation between A and C, if, conditional on B, the correlation of A and C goes to zero. In the above example, one observes that as drunk driving increases, the number of

<sup>1</sup> Although we agree that Bayesian networks are a useful tool for representing people's causal mechanism knowledge, we do not endorse the view that conditional dependencies are all there is to that knowledge. Ahn and Kalish (in press) state that conditional dependencies are consistent with causal mechanisms because people's ideas about mechanisms support patterns of association. For instance, if someone believes that getting sneezed on causes illness via the mechanism of the transmission of germs, they should expect that the covariation between sneezing and illness is conditional on the transmission of germs. However, Ahn and Kalish disagree with Glymour (1998), who argues that patterns of covariation *are* mechanisms, and not just evidence for them. That is, a pattern of covariation might be one useful piece of evidence for identifying a relation as causal (i.e., addressing an epistemic question), but they are not what people mean by causation (i.e., addressing a metaphysical question). The current study addresses the epistemic question rather than the metaphysical question. Thus, we do not focus on this debate and instead assume that mechanisms can be represented in terms of conditional dependencies.

	X	~X
Y	6	4
~Y	4	6

	Y	~Y
Z	6	4
~Z	4	6

Figure 2. Example contingency between events X and Y and contingency between events Y and Z.

traffic accidents increases. Conditional on the number of uncoordinated motor responses, however, the covariation between drunk driving and traffic accidents would be greatly reduced. Thus, uncoordinated motor responses serve as a mechanism for this covariation.

Little is known about how people actually learn these complex patterns of covariations. Waldmann and Martignon (1998), who make use of a Bayesian network to represent mechanism knowledge, admit that it is improbable that humans learn such networks bottom-up, as instantiated in some computational models (e.g., Spirtes et al., 1993). For instance, Hashem and Cooper (1996) generated nine sets of relatively simple causal networks (e.g.,  $A \rightarrow B \rightarrow C$ , or  $A \leftarrow B \rightarrow C$ ) instantiated as diseases. Second and third year medical students were instructed to ask for any conditional probabilities among the three variables in each network, and to estimate the causal strength between B and C after receiving answers to their questions. Even from these simple causal networks, their estimates significantly deviated from the normative answers. The results suggest that it is unlikely that people can keep track of all conditional probabilities necessary for acquiring causal networks.

A simpler way of acquiring mechanism knowledge is by combining piecemeal causal relations. In this study, we attempt to show that upon learning that X sometimes causes Y and Y sometimes causes Z, people conclude (albeit erroneously) that X sometimes cause Z. Such inference is non-normative in that even if there is a contingency between X and Y and a contingency between Y and Z, it does not guarantee that there will be a positive contingency between X and Z. A normative conclusion would be that no inference about the relationship between X and Z can be made.

Consider Figure 2 which shows 40 individual cases, 20 of which depict the covariation between X and Y and 20 of which depict the covariation between Y and Z ( $\Delta P = 0.2$  in both cases). Each of these 40 cases represents a different observation. On the left, we know, for instance, that there are six cases in which X and Y co-occurred, but we do not know what might have happened about Z in these six cases. Depending on this unknown information, the contingency

between X and Z can vary widely. To demonstrate this point, the top half of Figure 3 shows three possible distributions of these patterns of co-occurrence within the different levels of X and Y. For instance, in the six cases in Figure 3a. where both X and Y occur (the upper left-hand cell of the first contingency table), in two of the cases Z occurs and in the other four Z does not occur. The resulting co-occurrence patterns between X and Z are shown in the bottom half of the figure. Note that, not only is the pattern of co-occurrence between X and Y identical in each example, but the pattern between Y and Z is also identical (i.e.,  $\Delta P = 0.2$ ). However, the contingency between X and Z varies widely. Thus, a normative answer given covariation information about X and Y and covariation information about Y and Z only is that contingency between X and Z cannot be determined.

We propose that people would not make such normative judgments, and instead they would frequently assume that they can estimate the relationship between X and Z only from the covariation between X and Y (X-Y covariation) and covariation between Y and Z (Y-Z covariation). The reason for this is two-fold. First, as discussed earlier, keeping track of multiple conditional dependencies seems to be beyond the capacity of human cognition, but people have complex causal mechanism knowledge that can be represented in terms of conditional dependencies. Thus, people must have acquired this knowledge through other means. Second, in real-life situations, constituent covariations are oftentimes revealed in different cases. For instance, one might observe that eating a lot of food high in fat increases one's cholesterol level, and one might also observe that *other* people with high cholesterol die of a heart attack (not knowing whether these people had high-fat diets when alive). Therefore, it is adaptive, although non-normative, to make unwarranted inferences about unobserved covariations based on piecemeal covariations.

Specifically, we propose that people carry out a sort of syllogistic reasoning in this situation (Goldvarg & Johnson-Laird, 1999). Given that X causes Y and Y causes Z, people would subsequently conclude that X causes Z. We also propose that the stronger the perceived intermediate

	a. $\Delta P = -0.2$		b. $\Delta P = 0$		c. $\Delta P = 1$	
	X	~X	X	~X	X	~X
Y	6 (2 Z, 4 ~Z)	4 (4 Z)	6 (5 Z, 1 ~Z)	4 (1 Z, 3 ~Z)	6 (6 Z)	4 (4 ~Z)
~Y	4 (2 Z, 2 ~Z)	6 (2 Z, 4 ~Z)	4 (4 ~Z)	6 (4 Z, 2 ~Z)	4 (4 Z)	6 (6 ~Z)
Z	4	6	5	5	10	0
~Z	6	4	5	5	0	10

Figure 3. Example frequencies of co-occurrence between X and Z, holding constant the co-occurrence between X and Y, and between Y and Z. Note:  $\Delta P$ 's show contingency between X and Z.

causal relations are, the stronger the perceived causal relationship between the two terminal events would be. Thus, if causal relations between X and Y, and Y and Z are weak, one would infer a weak causal relation between X and Z. If so, any manipulation that increases the perceived intermediate causal strengths should also increase the perceived causal strength between the terminal events. One such manipulation is presented in Dennis and Ahn (in press) who manipulated the order of evidence supporting a positive causal relationship versus evidence supporting a negative causal relationship. Because the current study utilized the same manipulation, we will first describe this study in detail, and then return to the issue of deducing overarching causal relations from piecemeal covariations.

### Order manipulation

Consider four cells in Figure 1 again. Cells XY and  $\sim X\sim Y$  serve to confirm that a positive causal relationship exists between X and Y. Henceforth, we will call these two cells positive evidence. Cells X $\sim Y$  and  $\sim XY$  serve to confirm that there is a negative causal relationship between X and Y (negative evidence, henceforth). Participants in Dennis and Ahn (in press) observed a sequence of trials, each of which described presence or absence of two events, and judged the causal strength between the two events at the end of the sequence. Participants in one condition observed the bulk of positive evidence followed by the bulk of negative evidence (positive-first condition). In the other condition, participants observed the bulk of negative evidence followed by the bulk of positive evidence (negative-first condition). Although the order was different, all participants observed an identical covariation between X and Y, namely zero, in their experiment. The three possible results from this experiment were; (1) no effect of order, (2) a recency effect in which the negative-first condition leads to more positive causal estimates than the positive-first condition, and (3) a primacy effect in which the positive-first condition leads to more positive causal estimates than the negative-first condition. Existing models of causal induction predict either no effect (Cheng, 1997) or a recency effect (Rescorla & Wagner, 1972; see Dennis & Ahn, in press for more details of this prediction.) However, the results showed a strong primacy effect. This result was obtained even with the prospect of receiving reward for accurate judgments, indicating that the results are unlikely to be due to a fatigue effect.

Dennis and Ahn (in press) proposed that the primacy effect is obtained because causal learning occurs through a process of belief formation and updating. In this view, the information that a person receives at the beginning is used to construct an initial hypothesis about possible causal relationships. This initial belief then helps to provide an anchor point for future adjustments (Hogarth & Einhorn, 1992). However, as shown by Tversky and Kahneman (1974), people do not sufficiently adjust their initial anchor, resulting in the primacy effect.

### Order effect in the two-step causal chain

In addition to showing that people frequently infer unwarranted overarching causal relations from constituent covariations, the second goal of the current study is to examine whether the order effect is obtained when judging X to Z causal strength based on X-Y covariation and Y-Z covariation. In the positive-first condition, participants observed a bulk of positive evidence for X and Y, and for Y and Z, followed by a bulk of negative evidence for X and Y, and for Y and Z. In the negative-first condition, participants observed identical contingencies with the negative evidence preceding the positive evidence. It is hypothesized that compared to the negative-first condition, the positive-first condition will lead to more positive causal estimates for the relationship between X and Z.

## Method

### Overview of Methods

In general, there were four phases: instructions, a learning phase, a test phase, and a follow-up phase. In the learning phase participants observed a series of trials providing X-Y covariation, and Y-Z covariation. The test phase required that participants make judgments about the causal relationship between events X and Z. The main experimental manipulations which occurred during the learning phase were the order in which participants received a bulk of positive or negative evidence for X-Y and Y-Z covariation. This manipulation was a within-subject variable, so that each participant actually saw two sets of learning and test phases. In the follow-up phase, participants described their thought processes. Each phase is explained below.

The Instruction and the learning phases were presented on iMac computers, using Microsoft PowerPoint 98 ®. The test and follow-up phases were presented as a paper-and-pencil task. Participants were 39 undergraduates at Vanderbilt University.

### Procedure

**Instruction phase** In order to make participants get acquainted with the format of events, participants first received ten example learning trials with animations in which a person either does or does not eat a fictional plant called Ablex, and the same person subsequently does or does not exhibit a fictional physical reaction called Burlosis. The face of the person in each trial varied in order to have participants familiarized with the fact that each trial dealt with different cases.

Afterwards, participants were told to estimate "the extent to which Ablex plants cause Burlosis" on a scale from -100 (i.e., Ablex plants may prevent Burlosis) to 100 (i.e., Ablex may be a strong cause of Burlosis). Participants received instructions about the scale and examples of some of the scores. In addition, participants were instructed that

$XY$   $YZ$   $Y \sim Z$   $\sim X \sim Y$   $\sim Y \sim Z$   $XY$   $YZ$   $\sim X \sim Y$   $\sim Y \sim Z$   $X \sim Y$   $XY$   $YZ$   $\sim X \sim Y$   $\sim Y \sim Z$   
 $\sim XY$   $XY$   $YZ$   $\sim Y \sim Z$   $\sim X \sim Y$   $\sim Y \sim Z$   $XY$   $YZ$   $\sim X \sim Y$   $\sim Y \sim Z$  (Block2)  $Y \sim Z$   $X \sim Y$   $\sim XY$   
 $XY$   $YZ$   $Y \sim Z$   $\sim XY$   $\sim Y \sim Z$   $Y \sim Z$   $X \sim Y$   $\sim Y \sim Z$   $\sim XY$   $X \sim Y$   $\sim X \sim Y$   $\sim Y \sim Z$   $\sim Y \sim Z$

Figure 4. The sequence used for the positive-first condition. Note: The sequence should be read from left to right. The trials in outline are negative evidence. (Block 2) indicates where the positive block ends.

"You may also decide that you cannot determine an estimate, given the information presented. In this case, you should give an estimate of 'NA.'" Participants wrote down their estimate from the practice trials on the sheet provided.

**Learning Phase** Upon completing the practice trials, participants were told that in the actual experimental trials, they would see descriptions of three events; the possible application of a fictitious fertilizer, the possible increase in the level of a fictitious chemical in the soil, and the possible blooming of a fictitious flower. During each learning phase, they were told what these three events were; they were presented with animations that would accompany each event. (See the material section.)

Participants were specifically told that they will have only two pieces of information available during learning (e.g., "whether it [i.e., the plot] had increased levels of the chemical compound alizene and whether the plant Lanya subsequently bloomed on it, or whether it received the fertilizer Yerban and whether it subsequently had increased levels of alizene"). They were also explicitly told that they would never receive information about both the fertilizer and the plant. This instruction was added to prevent any false memory of having observed the covariation between the fertilizer and the plant. That is, if participants did not select "NA" in estimating the causal strength between the fertilizer and the plant, it cannot be due to the fact that they misremembered what covariation information they had seen. In addition, we attempted to reduce participants' cognitive load during the learning phase by instructing them what their task is in advance. Thus, participants were told that their task was, for instance, "to judge the causal relationship between Yerban and Lanya." After these instructions, participants were presented with 40 learning trials. (See the material section for more detail.)

**Test Phase** After observing the entire sequence of trials in a learning phase, participants provided causal strength ratings for the effect of the fertilizer on the plant's blooming. Following Wasserman, Elek, Chatlosh, and Baker (1993), participants were asked, for instance: "To what extent does the fertilizer Yerban cause the plant Lanya to bloom?" Participants wrote a number between -100 and 100. They were also reminded to write "NA" if they "cannot determine an answer from the evidence given."

**Follow-up Phase** When participants were done with the learning and test phases for two sets of materials, they were asked to rate how much thought they put into each judgment on a 5-point scale where 1 indicated "no" and 5 indi-

cated "very much." Finally, they were asked to write about their "thought process in performing the experimental task" such as "Were there any strategies in particular you used while observing the experimental trial? How did you interpret each type of evidence?"

## Design and Materials

During the learning phase, participants received 40 trials, in which 20 provided X-Y covariation information and 20 provided Y-Z covariation information. For both,  $\Delta P$  was 0.2 as in Figure 2.

Two experimental conditions were defined by the order in which covariation information was presented during the learning phase. In order to construct the experimental sequences, two different blocks (positive and negative blocks) were created. The positive block had 24 trials, 20 of which were positive evidence (i.e.,  $XY$ ,  $YZ$ ,  $\sim X \sim Y$ , or  $\sim Y \sim Z$ ). The negative block had 16 trials, 12 of which were negative evidence (i.e.,  $X \sim Y$ ,  $\sim X Y$ ,  $Y \sim Z$ , or  $\sim Y Z$ ). Within each block (positive or negative), the trials were randomly ordered except that  $XY$  was always followed by  $YZ$ , and  $\sim X \sim Y$  was always followed by  $\sim Y \sim Z$ <sup>2</sup>. This random order was fixed across participants.

The two different experimental conditions were constructed by manipulating the order of these two blocks, so that, in the positive-first order condition, the positive block came before the negative block. This pattern is shown in Figure 4 where the positive and the negative blocks are separated (Block 2). Although lines separate the positive and the negative blocks in this figure, the entire sequence was presented to the participants without any indication of blocks. In the negative-first order condition, the negative block came before the positive block, which can be seen by switching the two blocks in Figure 4. Each participant went through both experimental conditions; the order of conditions was counterbalanced across participants.

The actual events used for X was the application of a fertilizer called Yerban or Zertax, Y was a change in level of a chemical called Alizene or Banizon, and Z was the blooming of a plant called Lanya or Hyaeth. We used animations to show spraying of fertilizer, increasing chemical level, and blooming plant. These animations were intended to keep participants' attention and to reduce their cognitive load by visualizing the events, so that participants would not make NA responses simply because they were over-

<sup>2</sup> This constraint could not have limited our interpretation of the effect of order because the same constraint was used for both order conditions. In another experiment where this constraint was not imposed, the number of NA responses was approximately the same as in the current experiment.

whelmed with too many combinations of presence and absence of events. Finally, each trial had a unique plot number displayed at the top of the screen, so that it was clear that each observation was separate.

To summarize, after receiving general instructions, each participant observed a series of trials about covariation between X and Y, and covariation between Y and Z in either the positive-first or the negative-first order, and made a causal strength judgment about X and Z. Afterwards, they observed another series of trials about three new events in the other condition, and then made a second judgment. Finally, they wrote about their thought processes.

## Results

We first examined the number of NA responses. In order to be truly valid NA responses, a participant should have given NA responses in both the positive-first and the negative-first conditions. Only one out of 39 participants did so. This participant's explanation also agreed with the true justification for doing so, "It was very difficult to reason without seeing all three factors together..."

Overall, 20.5% of responses across the two conditions were NA responses. There are a number of reasons to believe that these NA responses were unlikely to indicate a response of "indeterminate," but rather were a way to indicate a lack of causal relation between the two events. First, as reported earlier, only one of these subjects gave NA responses in both conditions. Second, the other participants' reasons for giving an NA response are consistent with this interpretation. For instance, one participant stated, "... no causal relationship, or lack thereof, could be estimated because every relationship that was shown had another that contradicted it..."; another participant stated, "...There seemed to be no relationship between any..." Third, most interestingly, there were more NA responses from the negative-first condition (35.9% of participants) than from the positive-first condition (7.7% of participants),  $\chi^2(1, N=39) = 8.1, p < .01$ , McNemar's test (McNemar, 1947). As we shall see below, those who did not give NA responses gave lower estimates in the negative-first condition than in the positive-first condition. Thus, more NA responses in the negative-first condition seemed to reflect participants' belief in weaker causal strengths.

Finally, we examined the mean estimates for each condition. The mean rating in the positive-first condition were 32.5 whereas that in the negative-first condition was only 5.8. For a statistical analysis, we excluded data of those who gave at least one NA response in either condition. With the remaining 23 participants, a dependent *t*-test showed that the mean rating in the positive-first condition (22.6) was reliably higher than that in the negative-first condition (4.1),  $t(22) = 3.71, p = .001$ . Thus, although participants saw identical contingencies between X and Y, and between Y and Z, their estimated causal strength between X and Z was stronger when they first saw positive evidence

for these two contingencies than when they first saw negative evidence.

## Discussion

The experiment reported here suggests that people are willing to make overarching causal inductions from constituent covariations. The bulk of participants in our experiment were willing to infer a causal relationship between the two terminal events in a proposed causal chain, even though they did not see the actual covariation between the two events. Of those people who were not so willing to make that inference, the majority seemed not to understand the normatively correct reason for a response of "indeterminate," instead using such a response as a proxy for a perceived lack of causal relationship. This willingness to make overarching inductions seems to be a sensible thing to do, given that people rarely have the luxury in the real world of observing a complete set of covariation patterns between multiple events.

When people make these overarching inductions, they seem to first infer that X causes Y and Y causes Z. Based on these inferences, they conclude that X causes Z. Some participants' explanations for their responses supports this. For example, one participant wrote, "I tried to find the patterns; for example, that A caused B, and B caused C, so A probably causes C" Another wrote, "I tried to see the relationship between the plant and the compound, and compound and fertilizer separately first. From there I tried to determine whether or not the presence or absence of fertilizer yielded the presence or absence of plant..." In other words, it appears that people may try to integrate the relative strengths of the intermediate relationships to estimate the strength of the relationship between the terminal events.

In this study, we used three events that may have reflected prior knowledge about the function of chemical fertilizers. Participants could have judged the strength of the causal relationships based solely on this prior knowledge. But such an interpretation is unlikely, given that the events we used were fictitious ones (and thus, there could not have been prior knowledge about causal strengths among these events), and furthermore, people's causal strength estimates were susceptible to manipulation of the order of evidence. Finally, preliminary results from a new study show that the same effects occur using very abstract events (e.g. squares changing shape or triangles changing color).

Extending Dennis and Ahn (in press), we found an order effect in situations involving three events. As we suggested, we think this order effect occurs because of an anchoring-and-adjustment process. One participant's description precisely illustrates this process: "If a particular pattern kept coming, but one or two trials deviated from the pattern, I would excuse them as flukes." In this case, the adjustments to the initial anchor was not strong enough, leading to biased final estimates of causal strength.

These results also have implications for current, nor-

mative theories of causal learning (Cheng, 1997; Glymour, 1998; Glymour & Cheng, 1998). These theories propose that people's estimates of causal power match those predicted by contingency indices calculated from observed conditional probabilities. However, in the current experiment no such index can be calculated, given the lack of observed co-occurrence between the terminal events. Yet people still were willing to provide judgments of causal strength, suggesting that the normative contingency-based theories are inadequate descriptions of human causal learning.

In contrast, the results are consistent with a causal power view of causal learning. According to this view, people infer causal relationships based on the proposed transfer of some sort of causal force or energy between one object and another. Specifically, the mechanism by which one event brings about another is proposed to be the main focus of causal reasoning (Ahn, et al., 1995; Bullock, Gelman, & Baillargeon, 1982; Harré, 1988). In the case of our experimental results, the presence of a putative mechanism (i.e. the change in soil chemistry) seems to outweigh the absence of the covariation information necessary to draw accurate causal inferences. Furthermore, the perceived strength of the target relationship was tied to the perceived strength of the mechanism, as evidenced by the primacy effect obtained. That is, the current results demonstrate people's reliance on mechanism information in the acquisition of new causal learning.

### Acknowledgements

Support for this research was provided by a National Institute of Health Grant (NIH R01-MH57737) to Woo-kyoung Ahn.

### References

- Ahn, W., & Kalish, C. W. (in press). The role of mechanism beliefs in causal reasoning. In F. Keil & R. Wilson (Eds.), *Explanation and cognition*. MIT Press.
- Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*, 299-352.
- Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W.J. Friedman (Ed.), *The developmental psychology of time* (pp. 209-254). New York: Academic Press.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.
- Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*, 365-382.
- Dennis, M. J., & Ahn, W. (in press). Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. *Memory & Cognition*.
- Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines*, *8*, 39-60.
- Glymour, & Cheng, P. W. (1998). Causal mechanism and probability: A normative approach. In M. Oaksford & N. Chater (Eds.) *Rational models of cognition*. Oxford University Press.
- Goldvarg, Y., & Johnson-Laird, P. N. (1999). Naive causality: a mental model theory of causal meaning and reasoning. *Proceedings for 21st annual meeting of Cognitive Science Society*, Vancouver, Canada.
- Harré, R. (1988). Modes of explanation. In D.J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 129-144). Brighton, Sussex, UK: Harvester Press.
- Hashem, A. I., & Cooper, G. F. (1996). Human causal discovery from observational data. *Proceedings of the 1996 Symposium of the American Medical Informatics Association*.
- Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1-55.
- Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs*, *79*(1, Whole No. 594).
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*, 153-157.
- Pearl, J. (1996). Structural and probabilistic causality. In D. R. Shanks, D. L. Medin, & K. J. Holyoak (Eds.), *Psychology of learning and motivation, Vol. 34: Causal learning*. San Diego, CA: Academic Press.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research*. New York: Appleton-Century-Crofts.
- Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York: Springer-Verlag.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *211*, 453-458.
- Waldmann, M. R., & Martignon, L. (1998). A Bayesian Network Model of Causal Learning. *Proceeding of the 20th Cognitive Science Conference*, Hillsdale, NJ: Erlbaum.
- Wasserman, E. A., Chatlosh, D. L., & Neunaber, D. J. (1983). Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. *Learning and Motivation*, *14*, 406-432.
- Wasserman, E. A., Elek, S. M., Chatlosh, D. L., & Baker, A. G. (1993). Rating causal relations: Role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 174-188.

# Evaluating the Effectiveness of a Cognitive Tutor for Fundamental Physics Concepts

**Patricia L. Albacete** (albacete@isp.pitt.edu)  
Intelligent Systems Program; 607 Dixie Drive  
Pittsburgh, PA 15235 USA

**Kurt A. VanLehn** (Vanlehn@cs.pitt.edu)  
Learning, Research and Development Center; University of Pittsburgh  
Pittsburgh, PA 15260 USA

## Abstract

In this article we describe and analyze the evaluation of the Conceptual Helper, an intelligent tutoring system that uses a unique cognitive approach to teaching qualitative physics. The results of the evaluation are encouraging and suggest that the proposed methodology can be effective in performing its task.

## Introduction

Several studies (e.g. Hake, 1998; Halloun & Hestenes, 1985a, 1985b) have revealed that solving physics problems of a qualitative nature, such as the one presented in figure 1, pose a great cognitive challenge for most students taking elementary mechanics classes. They uncover naïve conceptions that are seldom removed or modified while completing their courses. Several attempts have been made to improve this situation though none has met with great success (Hake, 1998). Given that mechanics is a required course for most science majors, there is a clear need to improve its instruction. Toward this end we developed an intelligent tutoring system called the Conceptual Helper that follows a cognitive teaching strategy which is deployed emulating effective human tutoring techniques as well as successful pedagogical techniques and less cognitive demanding methods (Albacete, 1999; Albacete & VanLehn, 2000). In this article we describe the evaluation of the system and discuss its implications.

Two steel balls, one of which weights twice as much as the other, roll off of a horizontal table with the same speeds. In this situation:

- both balls impact the floor at approximately the same horizontal distance from the base of the table.
- the heavier ball impacts the floor closer to the base of the table than does the lighter.
- the lighter ball impacts the floor closer to the base of the table than does the heavier.

Figure 1. Example of a qualitative problem

## Brief description of the Conceptual Helper

The Conceptual Helper is an intelligent tutoring system (ITS) designed to coach students through physics homework problem solving of a qualitative nature, i.e., those problems that do not require the use algebraic manipulation to be solved but so require the application of conceptual knowledge. The tutor is basically a model-tracing ITS enhanced by the use of probabilistic assessment to guide the remediation. As a model-tracing ITS it contains a cognitive model that is capable of correctly solving any problem assigned to the student. Model tracing consists of matching every problem-solving action taken by the student with the steps of the expert's solution model of the problem being solved. This matching is used as the basis for providing immediate feedback to students as they progress through the problem. The system also has a student model which is represented by a Bayesian network. Each node in the network represents a piece of conceptual knowledge that the student is expected to learn or a misconception that the tutor can help remedy. Each node has a number attached to it that indicates the probability that the student will apply the piece of knowledge when it is applicable. As the student solves a problem, the probabilities are updated according to the actions taken by the student.

The challenge for the tutor is to decide when to intervene and what to say when it does so. This task is particularly challenging in this domain because tutoring of qualitative knowledge usually takes the form of verbal discussions, which given the state of the art of natural language processing is not an option for the computer tutor. To take care of the issue of when to intervene, we emulated human tutors in two ways: first, by giving immediate feedback (red for incorrect; green for correct) on each student entry (Merrill et al., 1992) and second, by helping the student with post-problem reflection (Katz & Lesgold, 1994; Katz et al., 1996). However, most of our work went into the second issue—deciding what to say when intervening. Novel approaches were developed in three areas: 1) the teaching strategy, 2) the manner in which the knowledge is deployed, and 3) the way in which misconceptions are handled.



## The Conceptual Helper's teaching strategy

Several studies (e.g. Van Heuvelen, 1991) have characterized students' knowledge of conceptual physics as a collection of ill-structured, unconnected facts and concepts which remain almost the same after completion of their physics classes. In contrast, cognitive science theory describes experts' knowledge bases as being well structured and *highly connected* (e.g. Chi & Koeske, 1983). Based on these findings, the teaching strategy embedded in the Conceptual Helper tries to make students' knowledge bases akin to the experts' by concentrating on teaching students the *links* that connect the domain's concepts of interest rather than the concepts in themselves.

The word "links" has been traditionally used in Semantic Networks to describe two-place predicates such as "is-a" or "part-of". However, we use the word "links" to describe rich qualitative rules that integrate pieces of knowledge. The links that the Conceptual Helper focuses can be inferred from the principles or from the definitions of the concepts of the domain. For example, one of the target links is "the direction of the net force applied to an object is the same as the direction of the object's acceleration." This connection between the concept of acceleration and the concept of net force can be inferred from Newton's second law. Likewise, the link "if the acceleration of an object is zero, then the object's velocity is constant" can be inferred from the definition of the concept of acceleration. These types of links are not evident to the students, in the sense that, even if students can repeat without hesitation the definition of acceleration and Newton's second law, by and large, they are generally not able to assert the links between concepts that follow from those definitions (Reif, 1995). However, these types of links are essential for reasoning qualitatively about the motion of objects and for solving the qualitative problems.

## How is the target knowledge taught?

The knowledge presented by the teaching strategy is deployed using a combination of: a) effective tutoring techniques, such as hinting through dialogues (Fox, 1993; Lepper et al., 1990), b) successful pedagogical techniques, like the use of molecular view of matter (Murray et al., 1990), and c) less cognitive demanding methods, such as using anthropomorphism (diSessa, 1993; Roschelle, 1992) and objects belonging to the material ontology (Chi, 1992) to reify abstract physics concepts. Figure 2 describes a mini-lesson that the tutor would present to the student when explaining the link "if (in a linear motion) the velocity of an object is decreasing, then the object's velocity and its acceleration have opposite directions." It exemplifies some of the techniques used by the tutor.

## The manner in which misconceptions are handled

To help students replace their misconceptions with scientifically correct knowledge, the Conceptual Helper presents students with the basic line of reasoning underlying the correct interpretation of the phenomena that are the base of the misconception. This is as opposed to using discovery environments or computer-simulated experiments, which

are two common ways in which teachers have tried to correct misconceptions (Hake, 1998). We believe that it is not setting up the (simulated) equipment, making the runs, recording the data, and inducing a pattern that convinces a student of a certain piece of knowledge, but rather the line of argument itself. Knowing the correct line of reasoning enables the student to self-explain the phenomenon, which has been argued (Chi, 1996) to be an effective means for learning.

## Evaluation of the Conceptual Helper

Forty-two students taking Introductory Mechanics classes were recruited and randomly divided into a Control group and an Experimental group. Both groups took a paper-and-pencil pre-test that consisted of 29 qualitative problems, 15 of which belonged to the Force Concept Inventory test<sup>1</sup>. Then they solved some problems with the Andes system receiving appropriate feedback according to the group they belonged to. The students in the Control Group had their input turned green or red depending on the correctness of the entry. Then, in the case of an incorrect action, the students could ask for help by making a choice from a help menu. The kind of help they received consisted of simple hints such as "the direction of the vector is incorrect." If the student asked for more help, they would just be told the correct answer. On the other hand the students in the experimental group received the green/red feedback depending on whether their action was correct but when the input was incorrect the Conceptual Helper intervened as explained above. After the students finished solving the problems with the system they took a post-test which was the same as the pre-test with the exception of a few changes in the cover stories of some problems. Among the problems included in the pre-test, post-test, and Andes there were multiple-choice questions and problems that required an explicit solution. Finally the students were asked to complete a questionnaire expressing their evaluation of the system.

## Results and their interpretation

The data gathered in such a way was analyzed in different ways.

### 1. T-test using the gain scores from pre-test to post-test as the dependent measure

Before comparing the gains of the two groups, we first checked whether their initial competencies were equivalent. The mean pretest score of the control group was 33.7 with standard deviation of 7.47. The mean pretest score of the experimental group was 31.36 with a standard deviation of 8.14. No reliable difference was found between the two groups ( $t(40)=0.965$ ,  $p=0.34$ ). Next, the gain scores from pre-test to post-test were compared. The mean of the control group was 4.12 with a standard deviation of 5.33.

---

<sup>1</sup> The Force Concept Inventory Test has become the standard test across the US to measure conceptual understanding of elementary mechanics (Hakes, 1998).

General definition of acceleration which constitutes the theoretical basis for the link (Reif and Allen, 1992)

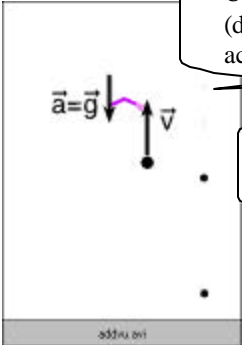
Anthropomorphizing the acceleration

General definition of the link (VanLehn et al., 1998)

Acceleration is a vector defined as the rate of change in velocity with time. You can think of the acceleration vector as what changes the velocity vector. Acceleration can change the velocity's magnitude, its direction, or both.

In this case, the magnitude of the velocity of the coin, i.e., its speed, is decreasing. The acceleration is making it shorter. For that to happen in a linear motion, the velocity vector and the acceleration, have to have opposite directions.

In the animation below, you can see the acceleration vector, with an imaginary arm, making the velocity vector shorter. Notice that the velocity and the acceleration have opposite directions.



Use of anthropomorphism to reduce cognitive demands (diSessa, 1993; Roschelle, 1992). Imaginary acceleration's arm shortening the velocity.

Use of vectors as the material representation of abstract concepts (Chi, 1992)

Why is the speed of the coin decreasing?

Figure 2: example of a mini-lessons

The mean of the experimental group was 7.47 with a standard deviation of 5.03. A reliable difference was found ( $t(40)=2.094$ ,  $p=0.043$ , two-tailed). This statistically significant difference suggests that the intervention of the Conceptual Helper had a positive impact on the students' understanding of the concepts as well as on their ability to abandon common misconceptions.

## 2. Effect size

Effect size is a standard way to compare the results of one pedagogical experiment to another. One way to calculate effect size, used in Bloom (1984) and many other studies, is to subtract the mean of the gain scores of the control group from the mean of the gain scores of the experimental group, and divide by the standard deviation of the gain scores of the control condition. That calculation yields  $(7.47 - 4.12) / 5.33 = 0.63$ . This result was comparable with peer and cross-age remedial tutoring (effect size of 0.4 according to Cohen, Kulik and Kulik, 1982). Some better results have been obtained with interventions that lasted a whole semester or academic year. For example, Bloom (1984) found an effect size of 2.0 for adult tutoring in replacement of classroom instruction and Anderson et al. (1995) reported an effect size of 1.0 for their tutoring systems. However,

our results were achieved with only two hours of instruction.

## 3. The fraction of the maximum possible gain realized (G)

Another measure that is used in the literature to compare the results of the FCI test is  $G = (S_f - S_i) / (100 - S_i)$ , where  $S_i$  and  $S_f$  are the pre- and post-test scores in percent (Hake, 1998). The nationwide score on the FCI test for traditionally taught classes is  $G = 0.25$ . For classes that are taught in a more interactive manner,  $G$  is between 0.36 and 0.68 (Mazur, 1997). The results obtained considering all the problems were the following: The mean of the control group was 0.26 with a standard deviation of 0.36. The mean of the experimental group was 0.43 with a standard deviation of 0.25<sup>2</sup>. The mean  $G$  for the control group matches that for traditionally taught classes. However, the  $G$  for the experimental group, 0.43, places it with the classes that are taught in a more interactive manner.

<sup>2</sup> Even though in the literature  $G$  is reported for each particular classroom in which a teaching method is applied and no statistical comparisons are made, we performed a two-tailed t-test to compare the  $G$  of the control and experimental group. The results were  $t(40)=1.84$ ,  $p=0.073$ .

#### 4. Existence of an aptitude-treatment interaction (ATI)

Innovative interventions sometimes cause higher gains for students with higher pre-test scores. What we want to find, of course, is that students with lower pre-test scores improved more, as they are the students who need more help. In order to see whether there was an aptitude-treatment interaction (ATI) and which way it would go, the experimental group was divided into two groups according to whether the student's pre-test score was above or below the median. The mean gain of the low pre-test score group was 10.68 with a standard deviation of 5.00. The mean gain of the high pre-test score group was 4.27 with a standard deviation of 2.39. A statistically significant difference between the gain scores was found ( $t(20)=3.83$ ,  $p=0.002$ , two-tailed).

A similar analysis was done with the control group. The results obtained were as follows: The mean gain of the low pre-test score group was 5.90 with a standard deviation of 5.87. The mean gain of the high pre-test score group was 2.35 with a standard deviation of 4.31. No statistically significant difference between the mean of the gain scores was found ( $t(18)=1.54$ ,  $p=0.14$ , two-tailed). Figure 3 illustrates the results for the experimental and control groups.

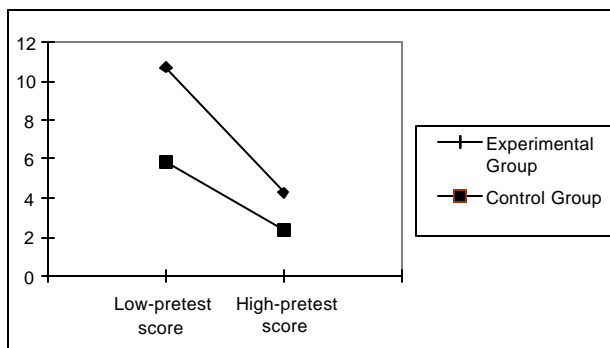


Figure 3. Mean gain of the low- and high- pretest score groups in the experimental and control groups.

It was encouraging to find that in the experimental group the poorer subjects' knowledge gains were significantly higher than those of good students, revealing that there was a desirable ATI. Additionally, it should be noted that the lower gain score in the high pre-test score group was not a consequence of a ceiling effect. The mean pre-test score of the high-pretest group was 38.04 with a standard deviation of 4.32. Since the maximum score is 49 there was an opportunity for this group to have a gain score very close to that achieved by the group of poorer students. Moreover, one student got a post-test score of 49, which indicates that the post-test did not require unlearnable knowledge.

#### 5. Detailed analysis of the individual pieces of knowledge and the effectiveness of each mini-lesson

A more detailed analysis was performed with the objective of determining the effectiveness of each mini-lesson in conveying the appropriate pieces of knowledge and in fostering their transfer. The method used basically consisted

of comparing whether receiving a mini-lesson had an effect on gaining versus not gaining the knowledge. Gaining the knowledge means giving an incorrect answer in the pre-test and a correct one in the post-test. Not gaining the knowledge means giving an incorrect answer in both pre- and post- tests. In the case where the target knowledge was addressed during explicit problem solving (e.g. for the rule "if an object's velocity is constant then its acceleration is zero") only students from the experimental group were considered, because only they could receive the mini-lessons. In the case where the target knowledge was addressed only through multiple-choice questions (e.g. the rule "heavier/lighter objects fall faster"), we compared the gains of the experimental group to the gains of the control group. The reason for doing this is that all the students in the experimental group received these mini-lessons, which were presented whenever the student answered a multiple-choice question (whether correctly or incorrectly). In the cases where the knowledge was addressed in both explicit and multiple-choice questions (e.g. the rule "force that continues to act after no contact"), we investigated whether receiving a mini-lesson during explicit problem solving would have any effect on gaining the rule. Hence only students from the experimental group were considered.

Statistical power problems prevented the analysis of most of the 18 target rules from showing a reliable relationship between receiving a mini-lesson and gaining the piece of knowledge. For some rules, almost all students received the corresponding mini-lesson, whereas for other rules, too few students received the mini-lesson. Nonetheless, there were a few rules where the relationship between mini-lessons and gain could be tested. They are described in Table 1. In all cases a Fisher's exact test (Hayes, 1994) was performed.

In most cases shown in Table 1 the number of students in each group was not large enough to provide statistical power, even if all those who received the mini-lesson gained and all those who did not receive the mini-lesson failed to gain (see third row of Table 1). Nonetheless, the data suggest a positive relationship between receiving a mini-lesson and gaining the corresponding knowledge.

#### 6. Summary of students' comments about the system

Students were asked to fill out a short questionnaire to express their opinion about the system. The rating of the different aspects of the system was done on a scale ranging from 1 to 5 where 5 was the best possible score. Students gave a score of 4 or above to all different aspects of the system (e.g., explanations that are clear to understand) which show a favorable acceptance of the system as well as a fairly high degree of liking of the mini-lessons.

## Discussion

The evaluation of the tutor suggests that the teaching strategy followed by the Conceptual Helper along with its methodology for deploying the target knowledge and handling misconceptions, is effective in accomplishing the task it was designed to perform. The experimental group surpassed the control group in every statistical test performed. Moreover, a detailed examination of the

Table 1. Relationship between receiving a mini-lesson and gaining knowledge for selected rules

Rule name	Group	Gainers	Non-gainers	Total	P	P for most extreme cases
Influence of weight on horizontal motion	<b>Experimental</b>	12 (.44)	2 (.07)	14 (.52)	0.005	<0.005
	<b>Control</b>	4 (.15)	9 (.33)	13 (.48)		
	<b>Total</b>	16 (.59)	11 (.41)	27		
When the velocity is constant the acceleration is zero	<b>Got mini-lesson</b>	5 (.5)	0 (0)	5 (.5)	0.08	0.08
	<b>Did not get mini-lesson</b>	2 (.2)	3 (.3)	5 (.5)		
	<b>Total</b>	7 (.7)	3 (.3)	10 (1)		
Heavier/lighter objects fall faster	<b>Experimental</b>	3 (.6)	0	3 (.6)	0.1	0.1
	<b>Control</b>	0	2 (.4)	2 (.4)		
	<b>Total</b>	3 (.6)	2 (.4)	5		
Vertical motion takes over horizontal motion	<b>Experimental</b>	9 (.45)	2 (.1)	11 (.55)	0.38	P<0.05 if all students in Exp. group are gainers
	<b>Control</b>	6 (.3)	3 (.15)	9 (.45)		
	<b>Total</b>	15 (.75)	5 (.25)	20		
Force that continues to act after no contact	<b>Got mini-lesson</b>	5 (.42)	3 (.25)	8 (.67)	0.24	P<0.05 all students that did not get mini-lesson were non-gainers
	<b>Did not get mini-lesson</b>	1 (.08)	3 (.25)	4 (.33)		
	<b>Total</b>	6 (.5)	6 (.5)	12 (1)		

The numbers in parenthesis represent the proportions with respect to the grand population

effectiveness of each individual mini-lesson showed a trend in favor of using the lesson.

Several studies (e.g. Halloun & Hestenes, 1985a) suggest that practice on solving quantitative problems does not transfer to conceptual problem solving. For instance, student who get full marks in their physics course still score poorly on the FCI test. On the other hand, elaborate confrontation-based, interactive instruction (e.g. see Hake, 1998) does raise scores on the FCI test, and by approximately the same amount as the Conceptual Helper. We believe that both forms of instruction are successful, at least in part, for the same reasons.

First, both handle misconceptions and errors with a form of confrontation. Both present students with situations (problems) and ask them to express their reasoning while solving them. In the Conceptual Helper, they do that by either taking an action, such as drawing a force, in an explicit solution problem or by choosing an answer in a multiple-choice question. If the action taken is incorrect, they are confronted with their erroneous knowledge by getting a mini-lesson. In interactive instruction the confrontations are quite elaborate and often involve doing experiments (e.g., Hake, 1992, McDermott, Shaffer & Somers, 1994, or White, 1993). What is interesting is that the evaluation of our system suggests that, in the case of misconceptions, confrontation based on *simply showing the*

*correct line of reasoning* to describe the phenomena under consideration can be just as effective in remediating misconceptions as the more elaborate, time-consuming kinds traditionally used to teach conceptual physics. Additionally, the evaluation suggests that, for correcting conceptual errors (or lack of knowledge), confrontation based on teaching the links that connect the concepts of the domain in the manner presented by the Conceptual Helper, may help the students build a more organized and better connected knowledge base, which in turn may facilitate qualitative reasoning.

A second factor underlying the success of both forms of instruction is that they both use conceptual problems instead of quantitative problems. This facilitates transfer, but it does not make it trivial. In particular, the Conceptual Helper does not “teach to the test” i.e., it does not teach exactly what the students are tested on. For example, the last rule in Table 1, which corresponds to the common misconception that there exists a force in the direction of the motion that continues to act after an object has been set in motion, shows a trend in favor of receiving a mini-lesson. The mini-lesson was received by students when they made a mistake in solving a problem that dealt with describing the motion of a box sliding on a frictionless surface after it has been pushed. On the other hand, the post-test problem analyzed in Table 1 involved describing the forces acting on

a ball thrown up in the air. Hence the situations presented in both problems were quite different even if the underlying misconception involved was the same.

In summary, it seems that the Conceptual Helper is just as effective but more efficient than other forms of qualitative physics instruction, in part, possibly because both forms of instruction use conceptual problems and confrontation. The next step in this line of research is to develop efficient and effective methods for *integrating* conceptual and quantitative learning.

## References

- Albacete, P.L. (1999). An Intelligent Tutoring System for teaching fundamental physics concepts. *Unpublished doctoral dissertation*. Intelligent Systems Program, University of Pittsburgh. Pittsburgh, Pennsylvania.
- Albacete, P.L. & VanLehn, K.A. (2000). (in press). *Fifth International Conference on Intelligent Tutoring Systems*. ITS'2000, Montreal, Canada.
- Anderson, J.R., Corbett, A.T., Koedinger, K.R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning sciences*, 4(2) 167-207.
- Bloom, B.S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13, 4-16.
- Chi, M.T.H. (1992). Conceptual change within and across ontological categories. In Gier, R. (Ed.) *Cognitive models of science: Minnesota studies in the philosophy of science*. University of Minnesota Press, Minneapolis, MN.
- Chi, M.T.H. (1996). Constructing Self-Explanations and Scaffolded Explanations in Tutoring. *Applied Cognitive Psychology*, 10, S33-S49.
- Chi, M.T.H. & Koeske, R.D. (1983). Network Representation of a Child's Dinosaur Knowledge. *Developmental Psychology* 19(1), 29-39.
- Cohen, P.A., Kulik, J.A., & Kulik, C.C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19, 237-248.
- diSessa, A.A. (1993). Toward an Epistemology of Physics. *Cognition and Instruction*, 1993, 10(2&3), 105-225.
- Fox, B.A. (1993). *The Human Tutorial Dialogue Project: Issues in the Design of Instructional Systems*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Hake, R. R. (1992). Socratic Pedagogy in the Introductory Physics Lab, *Phys. Teach.* 30, 546 (1992).
- Hake, R.R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(64).
- Hayes, W. (1994). *Statistics*. Holt, Reinhart & Winston, Inc. 5th edition.
- Halloun, I.A., & Hestenes, D. (1985a). The initial knowledge state of college physics students. *American journal of Physics* 53 (11) 1043-1055.
- Halloun, I.A., & Hestenes, D. (1985b). Common sense knowledge about motion. *American journal of Physics* 53 (11) 1056-1065.
- Katz S. & Lesgold A. (1994). Implementing Post-problem Reflection within Coached Practice Environments. In *Proceedings of the East-West International Conference on Computer Technologies in Education* (Part I; pp. 125-130). P. Brusilovsdy, S. Dikareva, J. Greer, V. Petrushin (Eds.). Crimea, Ukraine.
- Katz, S., Lesgold, A., Eggan, G., Greenberg, L. (1996). Towards the Design of a More Effective Advisors for Learning by Doing Systems. *Proceedings of the Third International Conference on Intelligent Tutoring Systems*, ITS'96. Montreal, Canada. Springer-Verlag.
- Mazur, E. (1997). *Peer Instruction*. Prentice Hall series in educational innovation. Upper Saddle River, NJ.
- McDermott, L. C., Shaffer, P. S., & Somers M. D. (1994). Research as guide for teaching introductory mechanics: An illustration in the context of the Atwood's machine. *American Journal of Physics* 62 (1) 46-55.
- Merrill, D.C., Reiser, B J., Ranney, M., & Trafton, J.G. (1992). Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *Journal of the Learning Sciences*, 2,2 77-306.
- Murray, T., Schultz, K., Brown, D., & Clement, J. (1990). An Analogy-Based Computer Tutor for Remediating Physics Misconception. *Interactive Learning Environments* 1(2), 79-101.
- Reif, F. (1995). Understanding and Teaching Important Scientific Thought Processes. *American Journal of Physics*, January 1995.
- Reif, F. & Allen S. (1992). Cognition for Interpreting Scientific Concepts: A study of Acceleration. *Cognition and Instruction*, 9(1), 1-44.
- Roschelle, J. (1992). Learning by Collaborating: Convergent Conceptual Change. *The Journal of the Learning Sciences*, 2(3), 235-276.
- Lepper, M.R., Aspinwall, L., Mumme, D., & Chabay, R.W. (1990). Self-perception and social perception processes in tutoring: subtle social control strategies of expert tutors. In J. M. Olson & M. P. Zanna (Eds.), *Self-inference processes: The Sixth Ontario Symposium in Social Psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Van Heuvelen, A. (1991). Learning to think like a physicist: A review of research-based instructional strategies. *American Journal of Physics*, 59(10), 891-896.
- VanLehn, K., Siler, S., Murray, C, & Bagget, W.B. (1998). What Makes a Tutorial Event Effective? In *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum.
- White, B.Y. (1993) ThinkerTools: Causal Models, Conceptual Change, and Science education. *Cognition and instruction*, 10(1), 1-100.

## Acknowledgments

This research was supported by a grant from the Cognitive Science Division of ONR, N00014-96-1-0260.

# Memory in Chains: Modeling Primacy and Recency Effects in Memory for Order

Erik M. Altmann (altmann@gmu.edu)  
Human Factors & Applied Cognition  
George Mason University  
Fairfax, VA 22030

## Abstract

Memory for order is fundamental in everyday cognition, supporting basic processes like causal inference. However, theories of order memory are narrower, if anything, than theories of memory generally. The memory-in-chains (MIC) model improves on existing theories by explaining a family of order memory effects, by explaining more processes, and by making strong predictions. This paper examines the MIC model's explanation of primacy and recency effects, and the prediction that primacy should dominate recency. This prediction is supported by existing data sets, suggesting that Estes's (1997) perturbation model, dominant among theories of order memory, is incorrect. Fits to data are presented and compared with fits of other models.

## Introduction

When EgyptAir Flight 990 crashed off the coast of Massachusetts last year, the co-pilot had been recorded commending his life to God shortly before the plane went down. Did he do this because he had decided to crash the plane? Or did he do this because the plane was already crashing? The correct causal inference depends on knowing more than the key events — it also depends on knowing the *order* in which they occurred. If there were a living eyewitness, that person's memory for order would be immensely valuable, assuming it were correct. A theory that would help to predict the accuracy of order memory would thus be important in many applied domains.

Despite the importance of order memory, current theories are, if anything, narrower than is typical of memory theories generally. For one thing, they are only descriptive, in that they reproduce empirical phenomena once the analyst has encoded the appropriate underlying memory representation. For example, a widely cited model of order memory is the perturbation model (Estes, 1997). This model takes as input an array of items indexed by the dimension along which order confusion can occur (in the example above, time). Every so often, two cells in this array have some chance of swapping with one another. Over time, elements drift away from their original position, producing an "uncertainty gradient". However, the assumption that memory is organized as an array suggests that memory is an immense multi-dimensional array, with a dimension for each different kind of confusion. A representation this complex would place a heavy burden on the encoding process that creates it, and yet the perturbation model fails to address encoding at all. Two other models of order memory, the primacy model (Henson, Norris, Page, & Baddeley, 1996; Page & Norris, 1998) and the partial matching model (Anderson & Matessa, 1997), fail to address the encoding question, as well.

This paper presents a model of order memory that not only explains the underlying encoding processes, but also fits existing data better than the other models cited above.<sup>1</sup> The memory-in-chains (MIC) model accounts for a family of effects, but the focus here is on the theoretical prediction that primacy should dominate recency in memory for order.

## Encoding Memory for Order

The model presented here is built on the ACT-R/PM cognitive theory, which combines perceptual-motor constraints (Byrne, 1998) with an analysis of memory as adapted to the structure of the environment (Anderson, 1990). The three theoretical mechanisms underlying the MIC model are a dual-code representation of attended objects, associative learning, and noisy communication between cognition and attention.

## Dual-Code Representation

The main assumption shaping the representation of items in the MIC model is that cognition and attention are different processes that must communicate.<sup>2</sup> This assumption is fleshed out by what we know about the functional roles of the two processes. For example, we know that cognition can program attention in a top-down manner, and we know that attention communicates relatively low-level information to cognition for complex processing.

This analytical framework allows us to specify generic processes involved in processing sequential stimuli. For a given stimulus, cognition must first tell attention to attend to the stimulus. Then, attention must send the attended object back to cognition for further task-related processing. Thus, processing one stimulus requires two acts of communication — one to direct attention and one to receive the contents of the attended location.

In terms of representation, this communication model implies that processing a single stimulus involves two codes. One code, representing the item's location or position, is passed from cognition to attention. Another code representing the item's semantic or post-categorical identity, is passed back from attention. This need for two codes per item converges with broad support in the literature for dual-

---

<sup>1</sup> Executable and documented code for the model is available at <http://hfac.gmu.edu/people/altmann/nairne-rpm.txt>

<sup>2</sup> I use "attention" here to mean attention to external stimuli, and will use "the focus of mental attention" to refer to ACT-R's internal goal focus. The latter maps roughly to the task-related contents of the central executive (reviewed in Baddeley, 1992).

code representations (e.g., Logan, 1996; Paivio, 1971; Whiteman, Nairne, & Serra, 1994).

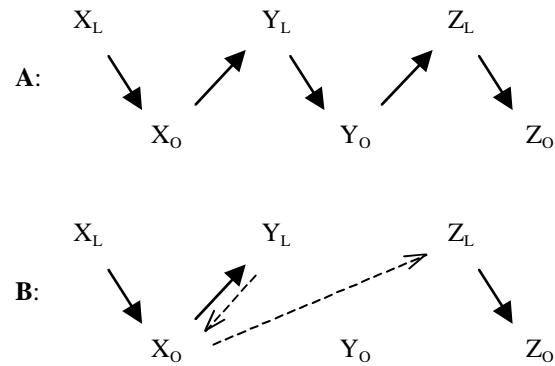
The communication model is illustrated in Figure 1A, which shows codes for three hypothetical items (X, Y, and Z). Time moves from left to right, and arrows mark the sequence in which codes appear in the focus of mental attention within cognition. (This interpretation of the arrows is elaborated below.) To process stimulus X, cognition sends a location code ( $X_L$ ) to attention, from which it receives an object code ( $X_O$ ). This is followed by whatever further task-related processing (not shown in the figure) might be required of the stimulus. The cycle then repeats for the next stimulus, Y.

An additional constraint on the model is that the channel through which cognition and attention communicate is the memory system itself. That is, when cognition sends a message to attention, it places a location code in memory for attention to retrieve. Similarly, attention sends a message back by placing an object code in memory for cognition to retrieve. This implementation of the communication channel is specified by the underlying theory, ACT-R/PM, but the tight functional integration of communication and memory can be traced to the earliest information-processing models of the cognitive system (e.g., Broadbent, 1958). The general implication is that functional descriptions of memory can also serve as functional descriptions of communication within the cognitive system as a whole. Two specific implications for the MIC model, concerning associative learning and noisy communication, are addressed in the next two subsections.

### Associative Learning

Evidence suggests that associative links between temporally proximal codes are acquired incidentally by the cognitive system (e.g., Altmann & John, 1999; Crowder, 1976; Hasher & Zacks, 1979; Mandler & Mandler, 1964; Nairne, 1983). Like other unified cognitive theories, ACT-R contains an *associative learning* mechanism to explain and predict the corresponding behavioral phenomena (Anderson & Lebiere, 1998). Associative learning in ACT-R creates a link between two codes if one (the target) is retrieved from memory while the other (the cue) is already in the focus of mental attention within the cognitive system. As in Soar (Newell, 1990), this association is a new, permanent element of long-term memory. In the future, if the cue again enters the focus of mental attention it will prime (spread activation to) the target, increasing the chance that the target will be the next item retrieved to the focus of attention. Associative links therefore allow chained retrieval, in which each retrieved item cues retrieval of the next item.

Applied to the memory-based communication protocol described above, associative learning produces a linked structure in which location codes are interleaved with object codes. Figure 1A illustrates such a structure after the model has studied and encoded the three hypothetical items (X, Y, and Z) introduced earlier. An important assumption in the model, based on standard associationist principles, is that each code remains in the focus of attention long enough to still be there when the next code is retrieved. The consequence is that the first code becomes the cue for the



**Figure 1:** Memory representations encoded by the MIC model after studying items X, Y, and Z. An item has a location code (subscript L) and an object code (subscript O). Panel A: Error-free representation. Panel B: Representation with two branches (incorrect links), in dashed ink, caused by noisy processing at study time.

second code, and the associative-learning mechanism links the two codes permanently in memory. In Figure 1 (and later figures), links created by associative learning are represented by arrows.

### Noisy Communication

If communication between cognition and attention were free of noise, then, subject to associative learning, it would produce a memory structure that allowed perfect sequential retrieval of items (Figure 1A). However, a memory system without noise would be unrealistic, and indeed sub-optimal (Anderson & Lebiere, 1998). In ACT-R as in other theories, items in memory have activation levels that determine their availability — items high in activation are less vulnerable to interference from other items. Noise in the memory system is expressed as transient fluctuations in individual activation levels, introducing the possibility of memory-retrieval error.

In the MIC model, noise can critically affect communication between attention and cognition at study time and produce incorrect links between codes. For each item processed, two memory retrievals are involved, one of a location code and one of an object code. Both retrievals are subject to activation noise. Specifically, when attention attempts to retrieve the location code most recently placed in memory, it may retrieve an old location code instead. Similarly, when cognition attempts to retrieve the object code most recently placed in memory, it may retrieve an old object code instead.<sup>3</sup> In terms of an everyday example, suppose that a newcomer is being introduced to a number of people, one at a time but perhaps too rapidly. While looking at the current person, the newcomer might “fall behind” and retrieve a previous, incorrect name. The result of such an

<sup>3</sup> I assume that errors occur within a code type only, and that a retrieval attempt always produces an item. These assumptions imply, for example, that an attempt to retrieve a location code will always produce a location code, though it may produce the wrong location code.

error would be that the newcomer might associate the wrong name with the wrong face. This kind of associative error is what the MIC model can encode at study time when there is noisy communication between attention and cognition.

Associative learning implies that a retrieval error during encoding produces an incorrect link in memory. I will refer to an incorrect link as a *branch*, because it branches off the correct temporal path through the codes of the list. The creation of a branch is illustrated in Figure 1B. There, a retrieval error occurs as  $Y_L$  is in the focus of mental attention and cognition tries to retrieve  $Y_O$ . This code was just placed in memory by attention, but due to noise in activation levels,  $X_O$  is transiently more active and hence is retrieved instead. This incorrect retrieval causes an association to be encoded between  $Y_L$  and  $X_O$ . This branch, shown as a dashed arrow, means that X could be mistakenly placed in Y's position at test time, producing an order error. This possibility is explored below in a discussion of the model's order-reconstruction process.

A second branch is also created in the scenario in Figure 3. When the model is presented with Z, it correctly retrieves  $Z_L$ , but  $X_O$  is still in the focus of mental attention (because of the retrieval error that occurred while processing Y). Therefore, associative learning creates a branch from  $X_O$  to  $Z_L$ , bypassing  $Y_O$ . This branch, however, need not produce an order error at test time, a possibility I also explore below.

A critical constraint on the communication model is the *near-miss* constraint, which is that incorrect codes temporally proximal to the correct code are more likely to intrude (and cause a branch). This constraint follows directly from the dynamics of activation in ACT-R. A code's activation depends on the lag since it was last retrieved — the longer the lag, the lower the activation. Therefore, a presented item will be more active than its predecessor (more precisely, the item's codes will be more active than its predecessor's codes), because the lag since presentation is smaller. The implication is that most branches created at study will be like those in Figure 1B — near misses, rather than far misses. This explains the uncertainty gradient, as I describe next.

## Reconstructing Memory for Order

In order-memory experiments, items themselves are usually shown at test as well as at study — participants are asked simply to reconstruct their original order. Because items and positions are available at test, an assumption I represent in the model is that people randomly choose an initial item or position to start the reconstruction process. This assumption means that the model can take many paths through the representation in Figure 1B. In particular, one of these paths produces a positional swap of the kind that underlies uncertainty gradients (Nairne, 1992), and a second path produces a correct reconstruction.

The model will make an order error if the first cue it uses is location code  $Y_L$ . This code was linked incorrectly to  $X_O$  at encoding time, because of a retrieval error then. The result now is that the model will infer that  $X_O$  was the object that originally appeared in location  $Y_L$ , producing an order error. Next, the model might use  $X_O$  as a cue for which location to focus on next, in which case it would focus on

location  $Z_L$ . Using  $Z_L$  as a cue, the model would most likely retrieve  $Z_O$ , which is correct. Thus, of two items placed, one was placed incorrectly and one correctly. The environment now indicates one remaining position and one remaining item. (Participants are typically instructed in the one-to-one nature of the reconstruction task, namely that every item maps to one position, with no items or positions left over.) The model will therefore infer that object  $Y_O$  occurred at location  $X_L$ . That is, the model will have swapped the order of the neighboring items X and Y. This is precisely the swap assumed (but not explained) by the perturbation model (Estes, 1997; Nairne, 1992).

Despite the encoding error, the structure in Figure 1B can also produce a correct reconstruction. If the model begins with location code  $X_L$ , for example, then it will most likely retrieve  $X_O$ , which is correct. Used as a cue,  $X_O$  will then prime two location codes,  $Y_L$  and  $Z_L$ . Suppose, first, that  $Z_L$  is retrieved. Used as a cue,  $Z_L$  will likely retrieve  $Z_O$ , which is correct. At this point, because only one item and one position remain, the model can place  $Y_O$  at  $Y_L$ , and the reconstruction will be correct. Suppose, instead, that when  $X_O$  is the cue,  $Y_L$  is retrieved. Used as a cue,  $Y_L$  will likely retrieve  $X_O$ , but this is now a dead end —  $X_O$  has been placed already. The model might now decide to place  $X_O$  elsewhere, but it might also decide simply to abandon  $Y_L$  as a cue and use  $Z_L$  instead. This would also produce a correct reconstruction.

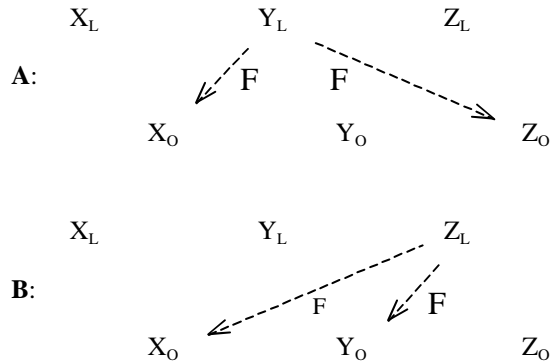
## How Primacy and Recency Arise

A standard empirical finding is that items at either end of a list are remembered more accurately than items in the middle. To explain these primacy and recency effects in order memory, we first need to revisit how the model generates order errors from an incorrect representation like the one in Figure 1B. Suppose, again, that the model initially focuses on  $Y_L$  at test time (essentially asking itself, "What item was in the second location?"). This cue will prime retrieval of  $X_O$ , causing the model to place X second instead of first. In contrast, given the correct representation of Figure 1A,  $Y_L$  would correctly prime  $Y_O$ . Thus, the frequency of branches, in aggregate data, is an important factor in determining the frequency of order errors. This relationship between branches and order errors means that we can examine branching patterns in the representation created at study time to predict error patterns at test time.

Primacy and recency effects arise in the MIC model because branch frequency is higher for middle items than for end items. Support for this claim comes from analyzing the interaction of branch frequency, branch *length*, and the distribution of branch lengths across a list. The notion of branch length is illustrated in Figure 2. Panel A shows two branches out of  $Y_L$ . Each branch is of length 1, meaning that the code at the head of the branch is temporally off by one from the correct code. Panel B shows two branches out of  $Z_L$ . One branch is of length 1, but the other is of length 2 because the code at the head of the branch is off by two from the correct code.

Two important points are illustrated in Figure 2. First, branch frequency varies inversely with branch length. That is, in aggregate data, branches to nearby codes are more





**Figure 2:** Middle codes have greater branch frequency than end codes. Panel A: A middle code with two short branches. Panel B: An end code with one short branch (bigger F) and one long branch (smaller F).

frequent than branches to far-away codes. This relationship follows directly from the near-miss constraint at encoding time: Temporally near codes are more likely than temporally remote codes to intrude on communications between cognition and attention and thereby cause branches. In Figure 2, branch frequency is indicated by the size of the “F” label. The branch of length 2 has a smaller F, meaning that it occurs less frequently in aggregate data.

The second point is that branch lengths are distributed unevenly across a list: Middle items have more short branches than end items. This distribution is also illustrated in Figure 2. Panel A shows all possible branches out of a middle code, where by “all possible” I mean that there is one branch to each possible incorrect code in the list. Similarly, Panel B shows all possible branches out of an end code. The middle code in Panel A has two short branches, whereas the end code in Panel B has only one. Because short branches are more frequent in aggregate data, the middle code will produce more order errors at test time.

In sum, primacy and recency effects in the MIC model reflect error patterns during encoding, in that middle items suffer branches more frequently than end items. At test, these extra branches produce more order errors.

### Prediction: Primacy Dominates Recency

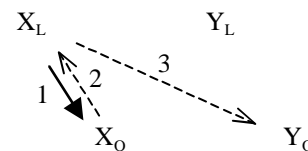
Models of order memory make conflicting predictions about the relationship between primacy and recency. The perturbation model, for example, predicts that primacy and recency should be symmetrical. In contrast, the primacy model was constructed to account for the common result that primacy is greater than recency (Henson et al., 1996).

The MIC model predicts that primacy should be greater than recency, an effect I refer to as *primacy dominance*. This prediction is a logical consequence of interactions between the task and constraints on the cognitive system (as specified by ACT-R/PM). In contrast, the primacy model (Henson et al., 1996; Page & Norris, 1998) accounts for primacy dominance with ad hoc mechanisms that are not constrained by task structure or independent theory.

Primacy dominance in the MIC model is a consequence of three interacting constraints. The first constraint is sequential processing at study — participants see one item at a time. The second constraint is related to branch *direction*. Every branch has a direction in that it points either forward or backward in time. A forward branch points to a code newer than the correct one (in Figure 3, from  $X_L$  to  $Y_O$ ). A backward branch points to a code older than the correct one (in Figure 3, from  $X_O$  to  $X_L$  instead of to  $Y_L$ ). As I elaborate below, branch direction interacts with sequential processing to make forward branches less likely to be taken at test time as the model is reconstructing order. The third constraint is the distribution of branch directions across a list. The early (not-recent) end of the list systematically involves more forward branches than the late (recent) end. Because forward branches are less likely to be taken at test time, early items suffer fewer order errors.

To see why forward branches are less likely to be taken at test time than backward branches, we need to consider the contingent nature in which forward branches are encoded at study. The encoding of forward and backward branches is illustrated in Figure 3. In that scenario, the model correctly processes  $X_L$  and transitions to  $X_O$  (creating link 1). A retrieval error then occurs — with  $X_O$  still in the focus of attention, the model retrieves  $X_L$  instead of  $Y_L$ . This creates a backward branch from  $X_O$  to  $X_L$  (link 2). The next step (assuming no further retrieval errors) creates a forward branch from  $X_L$  to  $Y_O$  (link 3). Thus, one retrieval error has produced two branches, one backward and one forward.

Two important points are illustrated in Figure 3. First, link 3 (the forward branch) is *contingent* on link 1 (the correct link). That is, a forward branch can only occur if a correct link out of the same code already exists. This contingency simply reflects sequential processing — X is already linked into the chain when Y is processed. The effect of this contingency is that at test time, if the model uses  $X_L$  as a cue, link 3 and link 1 prime competing targets. Thus the potential for taking a forward branch (link 3) is mitigated by the existence of the correct alternative (link 1). (By “taking a branch” I mean that the code at the tail end successfully primes the code at the head end, causing the latter code to be retrieved next.) The second important point in Figure 3 is that no such contingency accompanies a backward branch. Link 2 is the only link leading from  $X_O$ . At test time, if the model uses  $X_O$  as a cue, the backward branch will prime only  $X_L$ , with no correct alternative. Thus,



**Figure 3:** Forward branches are contingent on correct links, but backward branches are not. (1) Cognition retrieves  $X_O$ , creating a correct link. (2) Cognition retrieves  $X_L$  instead of  $Y_L$ , creating a backward branch. (3) Cognition retrieves  $Y_O$ , creating a forward branch.

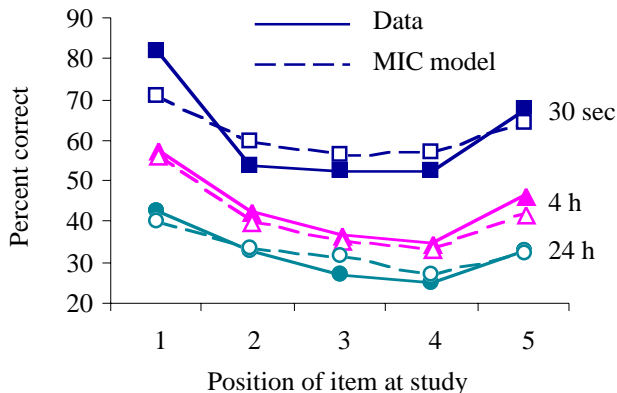
backward branches are more likely than forward branches to be taken at test time, in the sense that they prime only incorrect target codes. Put another way, backward branches have a higher *effective* branch frequency than forward branches. If a given forward branch and a given backward branch have the same frequency over multiple trials, the backward branch will be taken more often, making it effectively more frequent.

The third constraint leading to primacy dominance is that forward and backward branches are distributed unevenly across a list. Both kinds of branch occur with equal frequency overall, because a single retrieval error at study produces one branch in each direction. However, earlier items have more forward branches than later items. In the extreme cases, the first item can have only forward branches, and the last item can have only backward branches. Thus, earlier items have a lower effective branch frequency. That is, branches from earlier items, though as frequent as branches from late items, are effectively less frequent because they are less likely to be taken during order reconstruction.<sup>4</sup>

In sum, primacy dominates recency as a natural consequence of task structure interacting with cognitive constraints. Sequential processing makes forward branches contingent on correct links, and because forward branches are more frequent for early items, these items suffer fewer order errors. In graphical terms, the serial position curve in order memory is rotated slightly clockwise.

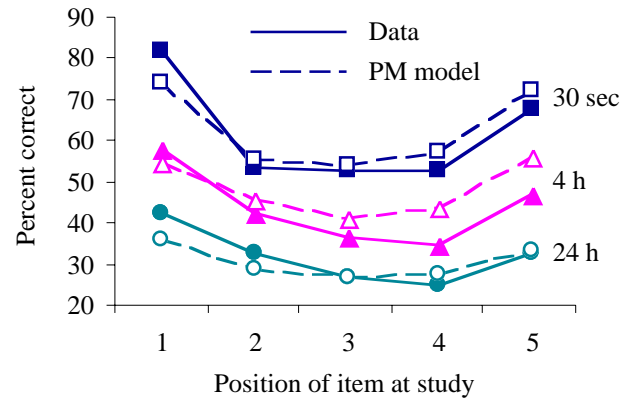
### Comparing Model to Data

To test whether the model reproduces the serial position effects predicted by the analysis above, I simulated data from Nairne (1992). In that study, memory for order was tested implicitly. Participants were asked to give pleasantness ratings of words, with words presented in lists of five for three seconds a word. In a between-subjects manipulation, participants were given a surprise order-reconstruction test after 30 seconds of distraction, after 4 hours, or after 24 hours.



**Figure 4:** Accuracy data for order memory (Nairne, 1992) and fits of the MIC model.

<sup>4</sup> Specifically, the first item has a lower effective branch frequency than the last item, the second item has a lower one than the second-last, and so on.



**Figure 5:** Accuracy data for order memory (Nairne, 1992) and fits of the partial-matching (PM) model.

Data from Nairne (1992) are shown in Figure 4, fit to data from the MIC model. In all three conditions, primacy appears to dominate recency, and the model captures this pattern, accounting for 93% of the variance over 15 data points (RMSE = 4.2%). The close fit of the MIC model to complex data is strong support for its assumptions.

Moreover, the fit of the MIC model improves slightly on that of the perturbation and partial-matching models of the same data. The partial matching model, which fits better than the perturbation model (Anderson & Matessa, 1997), accounts for 90% of the variance over the same 15 data points (RMSE = 5.0%).<sup>5</sup> These fits are close, but Figure 5 shows that in all three conditions the model under-predicts primacy and over-predicts recency. This mis-alignment is systematic, according to the MIC model, because the partial matching model (like the perturbation model) mistakenly predicts that primacy and recency should be the same.

Many important details about the MIC model are omitted here. For example, only 15 data points, or those for correct responses, are shown in Figure 4; the total number of points fit by the model is 75. In addition, I have not described the time parameter that causes the model's serial position curve to shift downwards with longer retention intervals. These issues will be addressed in a subsequent report.

### Discussion

The MIC model explains a family of phenomena in memory for order. This paper has described how the model explains primacy and recency effects — why they occur, and how they are related. Primacy and recency effects occur because middle items suffer more branches (incorrect links) than end items and thus are more vulnerable to order errors. In addition, primacy should dominate recency because early items suffer fewer backward branches than early items. Backward branches cause more order errors than forward branches, offsetting the benefits of recency and rotating the

<sup>5</sup> The 15 data points given here are a subset of the 75 data points found in Nairne (1992). Fits of the perturbation and partial matching models to the complete data set are given in Anderson and Matessa (1997). The fit of the partial matching model to the 15 data points used here was determined by running the model available on the Web at <http://act.psy.cmu.edu>.

bowed serial-position curve slightly clockwise. In addition to these serial position effects, the MIC model also explains positional uncertainty (Altmann, 2000), and thus is a step toward an integrated and executable theory of memory for serial order.

The MIC model is important for several reasons. First, it extends an existing cognitive theory to incorporate an additional set of effects. The model inherits a representation, a learning mechanism, and a communication channel from ACT-R/PM. The model's explanations follow directly from the integration of these mechanisms, illustrating (again) the explanatory power of unified theories (Newell, 1973; 1990).

Second, the MIC model goes beyond existing models of order memory to explain study-time processes as well as test-time processes. Of existing models, the perturbation model is the best known, and has been advanced as a generalized model of memory loss and distortion (Estes, 1997). However, the perturbation model has nothing to say about how memory for order is encoded at study time, begging the question of how the information-rich, array-like memory representation input to the perturbation model comes about in the first place.

Third, the MIC model is behaviorally distinguishable from the perturbation and partial-matching models. Both models predict that primacy and recency should be symmetrical, but several data sets suggest otherwise. The primacy model (Henson et al., 1996; Page & Norris, 1998) accommodates this primacy dominance, but like the others fails to explain how order information is encoded in the first place. The MIC model, in which primacy dominance is a logical consequence of the underlying memory theory, may also be the most accurate and complete explanation, as well.

Rigorously testing the prediction of primacy dominance will be the next important step in this research. Because this prediction flows from architecture-level premises (about representation, learning, and cognitive noise), primacy dominance should be found pervasively in empirical studies. A second important step will be to extend the model to account for the "sawtooth" pattern arising when confusable and non-confusable items are interleaved (Henson et al., 1996). Finally, order memory is a strong constraint on memory theory generally. As we build toward unified theories of cognition, it will be important to integrate order memory with related models (e.g., Anderson & Matessa, 1997; Burgess & Hitch, 1999) and with the rich theoretical history of serial learning (see, for example, Crowder, 1976).

### Acknowledgments

This work was supported by grant F49620-97-1-0353 from the US Air Force Office of Scientific Research. Thanks to Melanie Diez, Wai-Tat Fu, Wayne Gray, Margaret Peterson, Lelyn Saner, Wolfgang Schoppek, Chris Schunn, Greg Trafton, Richard Young, and an anonymous reviewer for their insights and comment.

### References

Altmann, E. M. (2000). Memory in chains: A dual-code associative model of positional uncertainty. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the 3rd international*

- conference on cognitive modeling*. Groningen, The Netherlands.
- Altmann, E. M., & John, B. E. (1999). Episodic indexing: A model of memory for attention events. *Cognitive Science*, 23(2), 117-146.
- Anderson, J. R., & Lebiere, C. (Eds.). (1998). *The atomic components of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review*, 104(4), 728-748.
- Baddeley, A. D. (1992). Is working memory working? The fifteenth Bartlett lecture. *Quarterly Journal of Experimental Psychology*, 44A, 1-31.
- Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon Press.
- Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, 106(3), 551-581.
- Byrne, M. D. (1998). Perception and action. In J. R. Anderson & C. Lebiere (Eds.), *Atomic components of thought* (pp. 167-200). Hillsdale, NJ: Erlbaum.
- Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.
- Estes, W. K. (1997). Processes of memory loss, recovery, and distortion. *Psychological Review*, 104, 148-169.
- Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, 108(356-388).
- Henson, R. N. A., Norris, D. G., Page, M. P. A., & Baddeley, A. D. (1996). Unchained memory: Error patterns rule out chaining models of immediate serial recall. *Quarterly Journal of Experimental Psychology*, 49A(1), 80-115.
- Logan, G. D. (1996). The CODE theory of visual attention: An integration of space-based and object-based attention. *Psychological Review*, 103(4), 603-649.
- Mandler, J. M. & Mandler, G. (1964). *Thinking: From association to Gestalt*. New York: John Wiley & Sons.
- Nairne, J. S. (1983). Associative processing during rote rehearsal. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(1), 3-20.
- Nairne, J. S. (1992). The loss of positional certainty in long-term memory. *Psychological Science*, 3(3), 199-202.
- Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283-308). New York: Academic Press.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review*, 105(4), 761-781.
- Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Reinhart, and Winston.
- Whiteman, H. L., Nairne, J. S., & Serra, M. (1994). Recognition and recall-like processes in the long-term reconstruction of order. *Memory*, 2(3), 275-294.

# Computational Explorations of the Irrelevant Sound Effect in Serial Short-Term Memory.

C. Philip Beaman (c.p.beaman@reading.ac.uk)

Department of Psychology  
University of Reading  
Earley Gate, Whiteknights  
Reading RG6 6AL

## Abstract

Although a number of current models of immediate serial recall exist, only one published model (Neath, 1999, 2000) incorporates simulations of the disruption of immediate serial recall by irrelevant background sound. This paper explores a possible model of irrelevant sound effects derived from Neath (1999) and applies the results of the model to previously unconsidered data sets. Studies by Neath (1999, 2000) apply the feature model, a mathematical model of short-term memory (Nairne, 1990), to some basic data regarding the irrelevant sound effect but this approach is ultimately limited by implicit assumptions regarding the nature of interference in short-term memory. Relaxing these assumptions allows for a wider application of a model of the irrelevant sound effect derived from that of Neath but not tied to the implementational detail of the feature model. The new model fits not only the original data considered by Neath (1999, 2000) but also empirical results concerning the effects of word-dose (Bridges & Jones, 1996) and token set size (Tremblay & Jones, 1998). It is concluded that the principles underlying the model provide a promising basis for further theoretical work.

## Introduction

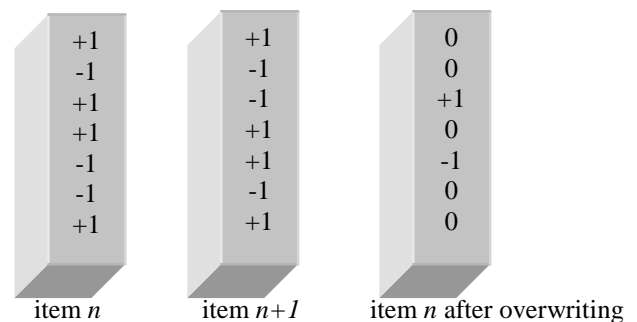
Theories of verbal short-term memory have frequently made use of the irrelevant sound effect, the disruption to serial recall of visually-presented verbal lists by background noise, to inform the proposed architecture of short-term memory (e.g., Salamé & Baddeley, 1982; Jones & Macken, 1993). Briefly, sound played to a participant during or immediately following the visual presentation of a to-be-recalled list impairs recall of the list even though the participant was explicitly told to ignore anything they might hear and participants are never tested on the contents of the "irrelevant" or "unattended" sound. It is well established that, although there are individual differences in the level of susceptibility to irrelevant sound disruption (Ellermeier & Zimmer, 1997), most participants show the effect.

There has been a paucity of formal simulations of irrelevant sound disruption, even though it has been claimed that a number of models of immediate serial recall can, in principle, account for the effect (Burgess & Hitch, 1999; Norris, Page & Baddeley, 1995). One model that has been applied to the effect is the feature model of Nairne (Nairne, 1990; Neath & Nairne, 1995), a mathematical model of

short-term memory based around the idea that the items most likely to be recalled from a list are those items which are most distinctive (Nairne, 1988). The model has been applied to a number of short-term memory phenomena including modality differences, interference from concurrent articulation and from post-list stimulus suffixes (Nairne, 1990), the word-length effect (Neath & Nairne, 1995), and latterly the irrelevant sound effect (Neath, 1999, 2000).

## The Feature Model

The feature model assumes that interference rather than decay accounts for loss from short-term or *primary* memory. Representations of items in the feature model are vectors that code for the "features" of an item using a binary system allowing features to assume the values of +1 or -1. Features may be modality dependent, coding information available only in a specific sensory modality, or modality independent, coding information that can be conveyed equally by two or more modalities. Interference occurs in the model through overwriting. If a feature takes the same value as its counterpart in the immediately preceding vector, the earlier feature value is overwritten. This is implemented by setting the value of the feature to 0 so that it is informationally uninformative. To give an example, if feature  $x$  of item  $n+1$  is the same as feature  $x$  of item  $n$ , then feature  $x$  of item  $n$  is overwritten and can play no part in determining whether or not item  $n$  is accurately recalled (see Figure 1). Feature values are generated randomly and independently for each feature vector.



**Figure 1.** Degradation of the representation of a list item in primary memory when a successive item shares some of the same feature values.

Retrieval consists of finding the best match to a degraded cue amongst a set of undegraded feature vectors assumed to reside in what is termed *secondary memory*. Two memory systems coding the same information is in many ways an unsatisfactory situation if retrieval depends on the degraded representation. Nevertheless, it is useful from the point of view of modelling the irrelevant sound effect since accurate recall of the degraded memory trace can be viewed as recall of the correct item in the correct order. In immediate serial recall the to-be-recalled stimuli are typically overlearned, lists of digits or consonants for example, so the task is essentially one of identifying which (known) item appeared in which serial position. As argued elsewhere (Beaman & Jones, 1997, 1998) the irrelevant sound effect consists primarily of a disruption of order information. The distance between the degraded item and its undegraded secondary memory representation is calculated by summing the number of mismatched features,  $M$ , and dividing by the total number of compared features,  $N$ , as described in Equation 1.

$$d_{ij} = a \sum \frac{b_k M_k}{N} \quad (1)$$

The value  $M_k$  is the number of times feature position  $x_{jk}$  does not equal feature position  $x_{ik}$ . The parameter  $a$  is a scaling parameter that is assumed to correspond to the overall level of attention, and  $b_k$  is used to weight particular comparisons if the task makes them more important than other comparisons. Distance,  $d$ , is then used to calculate the similarity between the degraded vector and the undegraded secondary memory representation according to Equation 2.

$$s(i, j) = e^{-d_{ij}} \quad (2)$$

The probability that a particular secondary memory trace,  $SM_j$  will be sampled as a potential recall response for a particular degraded memory vector  $PM_i$  is then given by Equation 3, where  $w_{ij}$  and  $w_{ik}$  are possible response bias weights.

$$P_s(SM_j | PM_i) = \frac{w_{ij}s(i, j)}{\sum_{k=1}^N w_{ik}s(i, k)} \quad (3)$$

This basic overwriting model was supplemented by Neath (1999, 2000) with two additional assumptions to account for the irrelevant sound effect<sup>1</sup>. The first assumption was that

<sup>1</sup> In fact, the full version of the feature model also includes a further recovery equation that produces the characteristic bow-shaped serial position curve. However here we are specifically concerned with the results of overwriting. Since there has been never been any suggestion that interactions between irrelevant sound and serial position might be of theoretical significance the recovery equation has been omitted here and performance averaged over serial position, a procedure also followed by Neath (Neath, 1999, 2000).

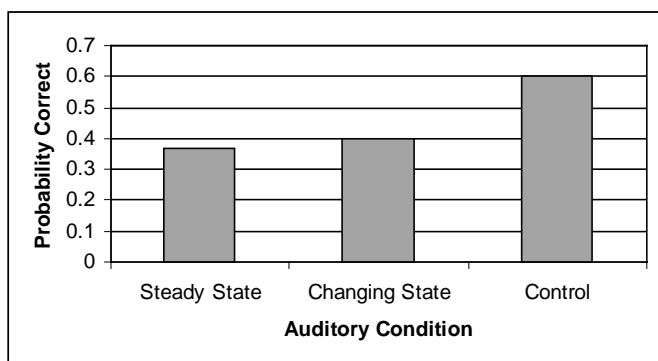
irrelevant sound will act like a concurrent articulation task, already accounted for by the model (Nairne, 1990) and overwrite a certain proportion of the modality independent features. The second assumption was that irrelevant sound and concurrent articulation manipulations differ in that effort is required to actively produce irrelevant noises in the concurrent articulation manipulation, which is not true of the irrelevant sound manipulation. Neath therefore proposed varying the attentional parameter  $a$  by a greater amount in simulations of concurrent articulation than in simulations of irrelevant sound.

With these amendments the feature model shows the correct qualitative pattern of results across a number of experimental studies altering only those parameters associated with the particular psychological process implicated (Neath, 1999, 2000). For example, the model shows correctly that the irrelevant sound manipulation impairs memory for lists of words, but less so than concurrent articulation. However, as with all simulation studies, there is experimental evidence not addressed by the model. Some of this evidence is directly relevant to the way in which irrelevant sound interferes with memory representations and cannot be accounted for by the feature model as it is currently formulated.

Two inconsistencies exist in the feature model account. Firstly, as described earlier, within-list interference results only in a lack of information about the overwritten item, not misinformation. Equation 1 ensures that these effects will be functionally identical, since only mismatches between the degraded vector and the undegraded secondary memory representation influence the similarity calculation (Equation 2) and both lack of information and misinformation are counted as mismatches. Nevertheless, the theory would be more parsimonious if all overwriting was implemented in an identical manner. The second inconsistency is more serious and concerns the difference between overwriting by irrelevant sound and overwriting by concurrent articulation. Concurrent articulation is implemented as setting half of the modality independent features to a constant value because participants are required to repeated the same utterance over and over "so the same information will overwrite the to-be-remembered items" (Neath, 2000). However in a simulation showing how varied speech (referred to in the literature as "changing-state" irrelevant sound) impairs recall performance more than repeated speech ("steady-state" irrelevant sound) (Jones, Madden & Miles, 1992) this logic was not used. Instead a variation in the attentional parameter is invoked, with variable speech assumed to attract more attentional resources.

The alteration in the attention parameter  $a$  is necessary as demonstrated by Figure 2, which shows the average sampling probabilities of a 9-item list in steady state, changing state and quiet control conditions across 200 simulations. The steady state condition comprised of setting half of the modality independent features to a constant value as described in previous simulation studies (Nairne, 1990). The changing-state condition comprised of overwriting half the modality dependent features with different random combinations of +1 and -1. The attentional parameter,  $a$ , was set to an identical value for all conditions. All other

weights were set to 1.0. Note that, provided all the other parameters remain unaltered, the same patterns of performance can be obtained at different overall recall levels by simple manipulation of the attentional parameter,  $\alpha$ . However, this would simply be an exercise in data-fitting and not of psychological interest. The important point to note is that without the adjustment of the attentional parameter no changing-state effect is observed. Parameter adjustment of this type is also perilously close to data-fitting.



**Figure 2.** Sampling probabilities of items in the feature model under steady state and changing state irrelevant sound conditions when the attentional parameter is kept constant.

### A Revised Model: The Changing-State & Token Set Size Effects

The problem of the changing-state effect can be viewed as a special case of what Tremblay and Jones (1998) termed "token set size". These authors argued that the essential cause of disruption by irrelevant sound was the presence of change in the irrelevant speech stream (Tremblay & Jones, 1998). The number of different changes, they argued, was irrelevant: disruption should markedly increase from one token (steady-state) to two tokens (changing-state) and there should be little or no further disruption beyond this token set size.

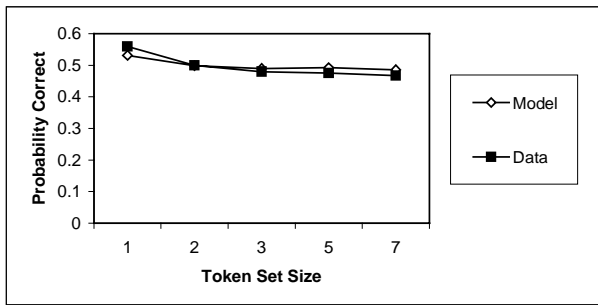
To give a concrete example, repetition of the utterance "A" in the irrelevant sound stream constitutes steady-state irrelevant sound and a token set size of one. According to Tremblay and Jones this should not cause discernible disruption to immediate recall. Repetition of the utterance "A-B" however, has a token set size of two and introduces change into the irrelevant sound stream and should therefore disrupt immediate recall. Repetition of the utterance "A-B-C" is also a changing-state stimulus (with a token set size of three) and should therefore also disrupt recall, but not necessarily to any greater degree than a token-set size of two since it is the number of changes, not the nature of the changes, which is important. Thus, changes from A to B then back to A are functionally equivalent to changes from A to B to C. Jones and Tremblay (Jones & Tremblay, 2000) argued that the increment in the attention parameter necessary to account for the changing-state effect did not

have a principled basis. If the increase in the attention parameter was necessary to account for increased attentional demands of changing-state stimuli, they argued, it should be increased in linearly as token set size increases, which would result in a linear increase in disruption not present in the experimental data.

A more realistic simulation than that attempted by Neath (Neath, 1999, 2000), and one that is not subject to these criticisms can be attempted by dropping the overwriting inconsistencies within the feature model. Closer examination of the experimental procedure employed in the Tremblay and Jones study reveals that over a 19 second presentation and retention interval Tremblay and Jones presented large numbers of repetitions of the same tokens. 38 separate occurrences of the same token in set size 1 condition, 19 repetitions each of 2 tokens in set size 2 condition, 13 repetitions of 3 tokens in set size 3, 8 repetitions of 5 tokens in set size 5 and 5 repetitions of 7 tokens in set size 7. If a conservative estimate of a covert rehearsal rate no faster than the slowest overt rehearsal rate of 2 items per second is assumed there could have been 38 rehearsals of a single item in this time period. The feature model has to assume that interference with the representations can occur at rehearsal as well as encoding since the experimental data demonstrate that the irrelevant sound effect can occur in an unfilled retention interval, after list presentation but before recall (Beaman & Jones, 1998). Therefore there will have been multiple opportunities for interference in this time period and the feature model's assumptions that changing-state irrelevant sound randomly overwrites half of each item's feature values once begins to look implausible.

Instead assume that each item was rehearsed once as it is presented- this is a standard assumption common to many models of immediate serial recall (e.g., Page & Norris, 1998). This leaves a 10 second retention interval which, with a slow rehearsal rate of 2 items per second and a 9 item list to rehearse gives time for only 2 complete rehearsals of the entire list. Thus altogether there is sufficient time for at least 3 rehearsals of the whole to-be-recalled list. During this time overwriting can occur. If, instead of the rather arbitrary random overwriting in Neath's version of the feature model, feature vectors are generated to represent the irrelevant sound utterances overwriting can then proceed according to the within-list overwriting principles specified by Nairne (1990). To simulate the token set size, the number of feature vectors representing the irrelevant sound was varied. Unlike the previous reported simulation, there was no adjustment of the attention parameter between set size 1 (steady-state) and set size 2 or above (changing-state).

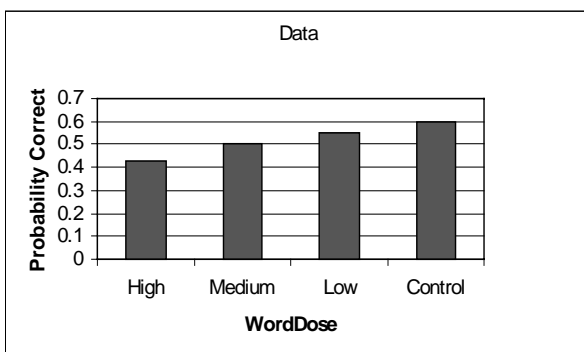
The data regarding the token set size effect are shown in Figure 3, together with a simulation study using the same procedure described here. The number of overwrites was set to 3 per item, and the items chosen to overwrite were randomly sampled from a set size of 2, 3, 5, or 7 randomly generated feature vectors. As Figure 3 clearly shows, this procedure produces a very good match between the performance of the model and the data from the experiment. Notably, the model actually provides a closer fit to the data than the predictions of Tremblay and Jones (1998).



**Figure 3.** The effects of token set size of irrelevant sound on sampling probabilities in the revised model. The crucial difference between steady state and changing state conditions is represented by the difference between set sizes 1 and 2, and unlike in the feature model, is here reproduced accurately.

### The Word-Dose Effect

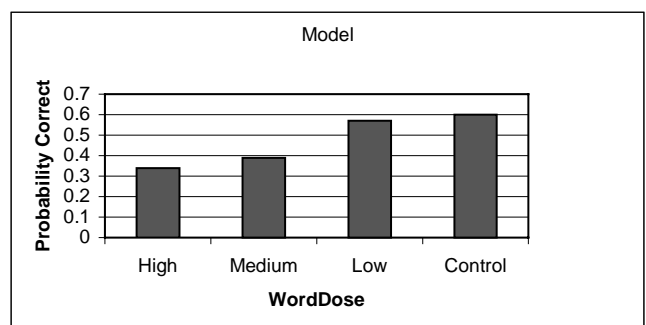
Further evidence not specifically considered by Neath (2000) which is impossible to fit into his account without amendment includes the so-called "dose" effect (Bridges & Jones, 1996). This refers to the finding that increasing the absolute number of words in the irrelevant sound stream increases the size of the effect. Dose differs from token set size in that, for example, "A-B-A-B" has a set size of two but a dose of four. The word dose manipulation introduced by Bridges and Jones (1996, Experiment 1) shows strongly linear effects (see Figure 4) when recall performance is collapsed across presentation position of the to-be-recalled lists. An attempt to fit a linear trend line to these data yielded an  $R^2$  value of .9978. The original feature model cannot account for these data because there is no mechanism within the model for relating probability of overwriting to number of irrelevant items presented. In the absence of this the model simply implements irrelevant sound interference of any type, regardless of the number of times each irrelevant item is presented as a single overwriting of each to-be-recalled item by a random combination of +1s and -1s. The problem presented by token set size effects is thus repeated, and the model cannot produce word dose effects.



**Figure 4.** Moderating effects of word dose on disruption of serial recall by irrelevant sound.

However, as before, reconsideration of the word dose data suggests an alternative modelling formulation. Bridges and Jones presented 5 different speech items repeatedly over the 9 second period of presentation of the to-be-recalled lists, a 10 second retention interval and a 15 second response time (Bridges & Jones, 1996). If the simplifying assumption is made to exclude the response time from the analysis then in the high dose condition participants heard 57 separate utterances, in the medium dose, 29 utterances, and in the low dose 19 utterances.

It is not clear how the timing of the rehearsal coincides with the presentation of the irrelevant sound material, however the data indicate that a linear relationship exists across high, medium and low word "dose". In the next simulation therefore I assume one overwrite per item for the low dose condition, and increment the number of overwrites by one for the medium-dose and two for the high-dose conditions. The item chosen to overwrite each time will be chosen at random from a set of 5 vectors representing the 5 irrelevant sound items generated in the same manner as the vectors representing the to-be-recalled list. Overwriting will then proceed in the same manner as within-list overwriting. It is clear that this procedure ensures not only that overwriting becomes internally more consistent but also allows for simulation studies of such effects as word dose that are more directly motivated by the experimental procedure and do not resort to altering free parameters. The results of the word dose simulation are shown in Figure 5. Comparison of this figure to the data displayed in Figure 4 reveals that a reasonable qualitative fit to the data has been obtained. There is a discernible effect of word dose, to which a linear trend line can be fit with  $R^2 = .9198$ , mirroring the linear trend observed in the experimental data.



**Figure 5.** Effects of word-dose on sampling probabilities of the revised model.

### General Discussion

Although the basic structure of the feature model was appropriated for this series of demonstrations, the intention was not to produce a simulation of irrelevant sound effects specific to the feature model. Instead, the intention was to investigate how some of the basic data regarding the irrelevant sound effect can emerge from an architecture in which items are represented in a distributed fashion and

presentation of irrelevant sound reduces the signal-to-noise ratio when recall of the order of those items is necessary. This investigation has succeeded in showing that increasing the noise in a distributed representation will reproduce many of the main findings in the irrelevant sound effect literature with relatively few assumptions. As such, there are three important points to note about this exercise.

The first point is to note that many of the feature model's assumptions, although implemented here, did not play any role in determining the outcome of the simulations. For example, although the assumption that overwriting occurs across features sharing the same value was implemented here, it is not necessary to make this assumption in order to obtain these results. Since each vector was constructed using random selections of binary values, the same result would be expected even if overwriting occurred across features with different values. It is possible to state with some confidence that reducing the signal-to-noise ratio by addition of noise to a distributed representation of the to-be-recalled item will therefore reproduce at least some of the key phenomena of irrelevant sound. The second point of note is that the simulations presented here reproduce many of the key characteristics predicted by Jones' changing-state hypothesis (Jones, Madden & Miles, 1992). These include: the changing-state effect itself, the specific disruption of order information, the word dose effect, and the lack of any great effect of token set size above 2 tokens. The simulations produce these effects, however, without the explicit representation of order cues assumed to be necessary by Jones.

The final point in favour of the current set of simulations is their relative parsimony and close correspondence to experimental procedure. Neath (2000) was criticized by Baddeley (2000) and Jones and Tremblay (2000) for the number of free parameters required in his simulations of irrelevant sound effects. The current set of simulations show that incrementing the attentional parameter is not necessary if the original (within-list) overwriting principles of the feature model are followed. This procedure provides a better fit to the data than the addition of the extra parameter. By explicitly matching the possibilities of overwriting to the rehearsal process it also proves possible to account for the word dose effect, which cannot otherwise be accounted for by the feature model. What is envisaged is an interference effect of discrete irrelevant sound elements on a continuous, serial, mental rehearsal process.

### Acknowledgements

Thanks to Tom Campbell, Dylan Jones and Philip Smith for comments and criticism. Much of this work was carried out while the author was employed as a postdoctoral researcher at the MRC Cognitive Development Unit, London.

### References

Baddeley, A. D. (2000). The phonological loop and the irrelevant speech effect: Some comments on Neath (2000). *Psychonomic Bulletin and Review*. In press.

- Beaman, C. P., & Jones, D. M. (1997). Role of serial order in the irrelevant speech effect: Tests of the changing-state hypothesis. *Journal of Experimental Psychology: Learning, Memory and Cognition* 23, 459-471.
- Beaman, C. P., & Jones, D. M. (1998). Irrelevant sound disrupts order information in free recall as in serial recall. *Quarterly Journal of Experimental Psychology* 51A, 615-636.
- Bridges, A. M., & Jones, D. M. (1996). Word dose in the disruption of serial recall by irrelevant speech: Phonological confusions or changing state? *Quarterly Journal of Experimental Psychology* 49A, 919-939.
- Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, 106, 551-581.
- Ellermeier, W., & Zimmer, K. (1997). Individual differences in susceptibility to the "irrelevant speech effect". *Journal of the Acoustical Society of America*, 102, 2191-2199.
- Jones, D. M., & Macken, W. J. (1993). Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19, 369-381.
- Jones, D. M., Madden, C., & Miles, C. (1992). Privileged access by irrelevant speech to short-term memory: The role of changing state. *Quarterly Journal of Experimental Psychology* 44A, 645-669.
- Jones, D. M., & Tremblay, S. (2000). Interference in memory by process or content? A reply to Neath (2000). *Psychonomic Bulletin and Review*. In press.
- Nairne, J. S. (1988). A framework for interpreting recency effects in immediate serial recall. *Memory and Cognition* 16, 343-352.
- Nairne, J. S. (1990). A feature model of immediate memory. *Memory and Cognition* 18, 251-269.
- Neath, I. (1999). Modelling the disruptive effects of irrelevant speech on order information. In: I. Neath, G. D. A. Brown, Poirier, M. & Fortin, C. (Ed.s). *Short-term/working memory*. Hove: Psychology Press.
- Neath, I. (2000). Modeling the effects of irrelevant speech on memory. *Psychonomic Bulletin and Review*. In press.
- Neath, I., & Nairne, J. S. (1995). Word-length effects in immediate memory: Overwriting trace decay theory. *Psychonomic Bulletin and Review* 2, 429-441.
- Norris, D., Page, M., & Baddeley, A. D. (1995). Connectionist modeling of short-term memory. *Language and Cognitive Processes*, 10, 407-409.
- Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of serial recall. *Psychological Review*, 105, 761-781.
- Salamé, P., & Baddeley, A. D. (1982). Disruptions of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior*, 21, 150-164.
- Tremblay, S., & Jones, D. M. (1998). The role of habituation in the irrelevant sound effect: Evidence from the role of token set size and rate of habituation. *Journal of Experimental Psychology: Learning, Memory and Cognition* 24, 659-671.



## Sex, Syntax, and Semantics

**Lera Boroditsky** (lera@psych.stanford.edu)

Department of Psychology; Jordan Hall, Bldg 420  
Stanford, CA 94305-2130 USA

**Lauren A. Schmidt** (lschmidt@stanford.edu)

Department of Psychology; Jordan Hall, Bldg 420  
Stanford, CA 94305-2130 USA

### Abstract

Many languages have a grammatical gender system whereby all nouns are assigned a gender (most commonly feminine, masculine, or neuter). Two studies examined whether (1) the assignment of genders to nouns is truly arbitrary (as has been claimed), and (2) whether the grammatical genders assigned to nouns have semantic consequences. In the first study, English speakers' intuitions about the genders of animals (but not artifacts) were found to correlate with the grammatical genders assigned to the names of these objects in Spanish and German. These findings suggest that the assignment of genders to nouns is not entirely arbitrary but may to some extent reflect the perceived masculine or feminine properties of the nouns' referents. Results of the second study suggested that people's ideas about the genders of objects are strongly influenced by the grammatical genders assigned to these objects in their native language. Spanish and German speakers' memory for object-name pairs (e.g., apple--Patricia) was better for pairs where the gender of the proper name was congruent with the grammatical gender of the object name (in their native language), than when the two genders were incongruent. This was true even though both groups performed the task in English. These results suggest that grammatical gender may not be as arbitrary or as purely grammatical as was previously thought.

### Introduction

Does the language you speak shape the way you understand the world? Linguists, philosophers, anthropologists, and psychologists have long been interested in this question. This interest has been fueled in large part by the observation that different languages talk about the world differently. However, despite the interest and controversy, definitive answers are scarce. This paper briefly reviews the empirical history of this question and describes two new studies that demonstrate both the role of semantic constraints in shaping language, and the role of language in shaping habitual thought.

The doctrine of Linguistic Determinism—the idea that thought is determined by language—is most commonly associated with the writings of Benjamin Lee Whorf. Whorf proposed that in so far as languages differ, their speakers too may differ in how they perceive and act in objectively similar situations (Whorf, 1956). What has been called the strong Whorfian view—the idea that thought and action are *entirely* determined by language—has long been abandoned in the field. Particularly effective in undermining the strong

view was work showing striking similarity in color memory despite wide variation in color language (Heider, 1972; but see Lucy & Shweder, 1979; Kay & Kempton, 1984).

Although the strong linguistic determinism view seems untenable, many weaker but still interesting formulations can be entertained. Several lines of research that have looked at domains other than color, have found cross-linguistic differences in thought. Unlike English speakers, speakers of classifier languages like Yucatec Mayan and Japanese were found to attend to the substance of an object more so than to its shape, and were also more likely to extend novel labels based on the substance than on the shape of a given example (e.g., Imai & Gentner, 1997; Lucy, 1992). When asked to reconstruct an array of objects, speakers of Tzeltal (a Mayan language that relies primarily on an absolute framework for describing spatial relations) were likely to preserve the positions of objects with respect to cardinal directions (so that the Northern-most object was still the Northern-most), while English speakers (who rely heavily on relative spatial descriptions) tended to preserve the objects' positions relative to themselves (so that the left-most object was still left-most) (Levinson, 1996).

Studies of conceptions of time have also revealed cross-linguistic differences (Boroditsky, 1999). English and Mandarin speakers talk about time differently—English speakers predominantly talk about time as if it were horizontal, while Mandarin speakers commonly use both horizontal and vertical metaphors to talk about time. This difference between the two languages is reflected in the way their speakers think about time. A collection of studies showed that Mandarin speakers tend to think about time vertically even when they are thinking for English (Mandarin speakers were faster to confirm that March comes earlier than April if they had just seen a vertical array of objects than if they had just seen a horizontal array, and the reverse was true for English speakers). Another study showed that the extent to which Mandarin-English bilinguals think about time vertically is related to how old they were when they first began to learn English. In another experiment native English speakers were taught to talk about time using vertical spatial terms in a way similar to Mandarin. On a subsequent test, this group of English speakers showed the same bias to think about time vertically as was observed with Mandarin speakers. This last result suggests two things: (1) language is a powerful tool in shaping thought, and (2) one's native language plays a role in shaping habitual

thought (how we tend to think about time, for example) but does not completely determine thought in the strong Whorfian sense.

There is an interesting discrepancy between these later findings, and those on color perception. Why would there be such strong evidence for universality in color perception, but quite the opposite for spatial relations or thinking about time? One possibility is that language is most powerful in influencing thought for more abstract domains, that is, ones not so reliant on sensory experience (Boroditsky, 1999). This paper considers an extreme point along this concrete-abstract continuum—the influence of grammatical gender on the way people think about inanimate objects. We will first characterize the ways in which people’s ideas about the genders of objects may be similar across cultures, and then go on to explore whether there may also be systematic language-driven differences in how people conceive of objects.

## Grammatical Gender

Forks and frying pans do not (by virtue of being inanimate) have a biological gender. The perceptual information available for most objects does not provide conclusive evidence as to their gender—conclusive gender information is only available in language (and only in those languages that have grammatical gender). The present paper examines whether (1) there are any correspondences in the assignment of grammatical gender between languages, (2) whether people include gender in their conceptual representations of objects (despite the fact that objects don’t actually have gender), and (3) whether people’s ideas about the genders of objects (if they have any at all) are influenced by the grammatical genders assigned to these objects in their native language.

Unlike English, many languages have a grammatical gender system whereby all objects (e.g., penguins, pockets, and toasters) are assigned a gender. Many languages only have masculine and feminine genders, but some also assign neuter, vegetative, and other more obscure genders. It has long been claimed that the assignment of grammatical gender to object names is semantically arbitrary, and has nothing to do with the conceptual properties of the referent (e.g., Bowers, Vigliocco, Stadthagen-Gonzalez & Vinson 1999). At first glance, this does appear to be the case. As Mark Twain noted, “In German, a young lady has no sex, while a turnip has, ....a tree is male, its buds are female, its leaves are neuter; horses are sexless, dogs are male, cats are female—tomcats included.” Further, the grammatical genders assigned to names of particular objects vary greatly across languages (Braine, 1987). For example, the sun is feminine in German, but masculine in Spanish, and neuter in Russian. The moon, on the other hand, is feminine in Spanish and Russian, but masculine in German.

Despite wide variation in the assignment of grammatical genders, speakers across languages do share some common beliefs about the genders of objects. For example, when asked to classify names or pictures of objects into masculine and feminine, English and Spanish speakers tend judge natural objects as feminine and artifacts as masculine (Mullen, 1990; Sera et al., 1994). It is also interesting that English speakers make consistent judgments about the genders

of objects, despite the lack of a grammatical gender system in English (Sera et al., 1994).

So are people’s shared beliefs about the genders of objects reflected in the assignment of grammatical gender, or is grammatical gender entirely arbitrary? If the assignment of grammatical gender is not entirely arbitrary, then there may be some correspondences across languages. For example, animals or things that are easy to anthropomorphize may have stereotypically feminine or masculine qualities and so may be more likely to have consistent grammatical genders across languages. The names of animals that are beautiful and graceful may tend to be grammatically feminine, while those of aggressive and strong animals may tend to be masculine. It is possible then, that the grammatical genders of nouns may correspond across languages. Further, we should see more correspondence for nouns whose referents are easy to anthropomorphize (and are likely to have stereotypically masculine or feminine properties) than for nouns whose referents are more abstract or less human-like.

To test these predictions, we compared the grammatical genders assigned to objects in Spanish and German to the intuitions of English speakers regarding the gender of the same objects. Since English does not use grammatical gender, English speakers’ untrained intuitions about the genders of objects provide a nice comparison group. If the assignment of grammatical gender is truly arbitrary, then we should see no correspondence between the intuitions of English speakers about the genders of objects and the genders assigned to those objects in Spanish and German. If, on the other hand, the grammatical genders of nouns do in part reflect the properties of their referents, then we should see a correspondence in the assignment of genders across languages, and also a correspondence between Spanish and German genders and English speakers’ naive intuitions.

## Experiment 1

### Methods

#### Participants

Fifteen native English speakers (none of whom were familiar with either Spanish or German) participated in this study in exchange for payment.

#### Materials

We constructed a list of 50 animal names and 85 names of artifacts (including vehicles, articles of clothing, and household items). Only words that had a single dominant translation (as determined by two native Spanish and two native German speakers) into both Spanish and German were included on the list.

#### Procedure

English speakers were asked to classify each object and animal on our list as either masculine or feminine. Participants were required to provide a single answer for each item.

## Results

Overall, there was appreciable agreement on the assignment of grammatical genders between Spanish and German ( $r=.21$ ,  $p<.05$ ). As we predicted, the two languages agreed more on the genders of animals ( $r=.39$ ,  $p<.01$ ), then on the genders of artifacts ( $r=.10$ ,  $p=.35$ ). Interestingly, English speakers' ratings of these objects showed the same pattern of correspondence. Spanish and German grammatical genders corresponded well with English speakers' intuitions about the genders of animals ( $r=.29$ ,  $p<.05$ , and  $r=.43$ ,  $p<.01$  respectively), but not the genders of artifacts ( $r=.04$ ,  $p=.73$ , and  $r=.11$ ,  $p=.32$  respectively). It is striking that despite a lack of grammatical gender in English, English speakers intuitions about the genders of animals corresponded well with the grammatical genders assigned to those animals in Spanish and German. These findings suggest that the grammatical genders assigned to animals may not have been entirely arbitrary, but rather may have reflected people's perceptions of the particular animals as having stereotypically masculine or feminine properties.

It appears that the assignment of grammatical genders to nouns (or at least to animal names) may not be entirely arbitrary, and may have been influenced in part by people's perceptions of the nouns' referents. But what happens once grammatical genders are assigned? Could they in turn influence people's mental representations of objects? If so, then there may be striking cross-linguistic differences in how people think about objects.

How might people's representations of objects be affected by the grammatical gender of their labels? One possibility is that in order to efficiently learn the grammatical gender of a noun to begin with, people focus on some property of that noun's referent that may pick it out as masculine or feminine. For example, if the word for "sun" is masculine in one's language, one might try to remember this by conceiving of the sun in terms of what are perceived as stereotypically masculine properties like powerful and threatening. If the word for "sun" is feminine, on the other hand, one might focus on its warming and nourishing qualities.

Even after the grammatical genders of nouns are learned, language may influence thought during "thinking for speaking" (Slobin, 1996). Languages can force their speakers to attend to the genders associated with objects by making them grammatically obligatory. When speaking a language with grammatical gender, speakers often need to mark objects as gendered through definite articles (e.g., "le" and "la" in French), refer to objects using gendered pronouns (e.g., if the word for "fork" is masculine, a speaker might say, "he is sharp"), and alter adjectives or even verbs to agree in gender with the nouns (e.g., in Russian, verbs in the past tense must agree in gender with their subject nouns). Needing to refer to an object as masculine or feminine may lead people to selectively attend to that object's masculine or feminine qualities thus making them more salient in the representation.

So, does talking about inanimate objects as if they were masculine or feminine lead people to think of inanimate objects as masculine or feminine? Some preliminary evidence suggests that it may (Jakobson, 1966; Konishi, 1993; Sera, Berge, & del Castillo, 1994). In one early study, Rus-

sian speakers were asked to personify days of the week (reported in Jakobson, 1966). Subjects consistently personified the grammatically masculine days of the week (Monday, Tuesday, and Thursday) as males, and the grammatically feminine days of the week (Wednesday, Friday, and Saturday) as females, though they could not explicitly say why they did so.

In another study, German and Spanish speakers rated a set of nouns on the dimension of potency (a dimension highly associated with masculinity) (Konishi, 1993). Half of the nouns were grammatically masculine in German and feminine in Spanish, and the other half were masculine in Spanish and feminine in German. Both German and Spanish speakers judged the word "man" to be more potent than "woman". Interestingly, they also judged nouns that were grammatically masculine in their native language to be more potent than nouns that were grammatically feminine. This was true even though all of the test nouns referred to objects or entities that had no biological gender (including names of inanimate objects, places, events, and abstract entities).

Converging evidence comes from a series of studies in which Spanish speakers were asked to rate pictures of objects as masculine or feminine (Sera et al., 1994). Spanish speakers consistently classified objects in accordance with their grammatical gender in Spanish. The effect was more pronounced when the pictures were accompanied by their Spanish labels. The grammatical gender consistency effect also showed up when subjects were asked to attribute a man's or a woman's voice to each picture. Finally, Sera et al. found that by about second grade, Spanish speaking children assigned voices to objects in accordance with the grammatical gender of their labels.

Although results of these studies are suggestive, there are serious limitations common to these and most other studies of linguistic determinism. First, speakers of different languages are usually tested only in their native language. Any differences in these comparisons can only show the effect of a language on thinking for that particular language. These studies cannot tell us whether experience with a language affects language-independent thought such as thought for other languages, or thought in non-linguistic tasks.

Second, comparing studies conducted in different languages poses a deeper problem: there is simply no way to be certain that the stimuli and instructions are truly the same in both languages. This problem remains even if the verbal instructions are minimal. For example, even if the task is non-linguistic, and the instructions are simply "which one is the same?", one cannot be sure that the words used for "same" mean the same thing in both languages. If in one language the word for "same" is closer in meaning to "identical," while in the other language it's closer to "relationally similar", speakers of different languages may behave differently, but due only to the difference in instructions, not because of any interesting differences in thought. There is no sure way to guard against this possibility when tasks are translated into different languages. Since there is no way to know that participants in different languages are performing the same task, it is difficult to deem the comparisons meaningful.

Finally, in all of the tasks so far, participants were asked to provide some subjective judgment (there were no right or wrong answers). Providing such a judgment requires participants to decide on a strategy for completing the task. When figuring out how to perform the task, participants may simply make a conscious decision to follow the grammatical gender divisions in their language. Evidence collected from such subjective judgments cannot tell us whether gender is actually part of a person's conceptual representation of an object, or if (left with no other criterion for making the subjective judgment) the person just explicitly decided to use grammatical gender in answering the experimenter's questions.

The present study improves on the previous studies in two important ways. First, both Spanish and German speakers were tested in English. This allows us to test whether experience with a language affects language-independent thought (here, thinking for other languages). Second, participants were tested in a memory task and at test were asked to provide the right answer (not a subjective judgment). The present study examined the ways in which previous knowledge (experience with Spanish or German) interfered with participants' ability to correctly perform the task.

In this study, participants were taught proper names for objects (e.g., an apple may have been called "Patrick") and were tested on their memory for these object—name pairs later in the experiment. First, we were interested in whether English speakers would be better at remembering female names for objects that another group of English speakers had rated as more feminine (and male names for objects rated more masculine). Second, we were interested in whether Spanish and German speakers would be better able to remember a proper name for an object if the proper name was consistent with the grammatical gender of the object name in their native language. All objects were chosen to have opposite grammatical genders in Spanish and German (e.g., the word for "apple" is feminine in Spanish, but masculine in German). So, we predicted that German speakers would be better at remembering a proper name for "apple" if the name was "Patrick" than if it was "Patricia". The opposite should be true for Spanish speakers. Since the experiment was conducted entirely in English, this is a particularly conservative test of whether grammatical gender influences the way people think about objects.

## Experiment 2

### Methods

#### Participants

Twenty-five native Spanish speakers, sixteen native German speakers, and twenty English speakers participated in the study in exchange for payment.

#### Materials and Design

A set of 24 object names (e.g., apple, arrow) and 24 proper names (e.g., Patricia, Patrick) was constructed (see Appendix A). The object names were chosen such that half

were grammatically masculine and half were grammatically feminine and the grammatical gender in Spanish and German was opposite for each object name (if an object name was grammatically masculine in Spanish, it was grammatically feminine in German and vice versa). A separate group of 30 English speakers rated the 24 objects chosen for this experiment as masculine or feminine.

Half of the proper names were male and half were female; male and female proper names were chosen to be similar to one another (e.g., Alexander, Alexandra). This was done to increase the difficulty of the memory task. All of the materials used including the instructions were in English. For each participant, the computer randomly arranged the object names and proper names into object—name pairs, and presented them in a random order.

Spanish, German, and English speakers completed the same experimental task. Participants read the following instructions "For this experiment, we have given names to a bunch of objects. For example, we may have decided to call a chair 'Mary'. You will see objects and their names appear on the screen (e.g., chair—Mary), and your task is to try to memorize the name we have given to each object as well as you can. Your memory for these names will be tested later in the experiment."

#### Procedure

Participants were tested individually. A computer presented the experimental materials and recorded the participants' responses.

**Learning:** Participants learned 24 object—name pairs presented to them on a computer screen in a random order. Each object—name pair was presented on the screen for five seconds, and was automatically followed by the next pair. Each pair was presented only once.

After the learning, participants completed a five-minute distraction task unrelated to this study which was inserted to promote forgetting.

**Test:** Object names from the learning set were presented on the computer screen one at a time and participants were instructed to indicate the gender of the proper name that had been associated with that object name in the learning set by pressing one of two keys on the keyboard.

### Results

As predicted, English speakers remembered object—name pairs better when the gender of the proper name was consistent with the object's rated gender (86% correct) than when the two genders were inconsistent (78% correct),  $t=2.17$ ,  $p<.05$ . The results suggest that people do include gender in their conceptual representations of inanimate objects. Further, Spanish and German speakers showed language-specific biases in memory. Both groups remembered object—name pairs better when the gender of the proper name given to an object was consistent with the grammatical gender of the object name in their native language (82% correct) than when the two genders were inconsistent (74% correct),  $t=2.55$ ,  $p<.01$ . Since the object names used in this study had opposite grammatical genders in Spanish and German, Spanish and German speakers showed opposite

memory biases—for those objects that Spanish speakers were most likely to remember female names, German speakers were most likely to remember male names (and vice versa),  $F(1, 39)=6.21$ ,  $p<.05$ . These findings suggest that people's ideas about the genders of objects are strongly influenced by the grammatical genders assigned to those objects in their native language.

### Summary

Two studies examined whether (1) the assignment of genders to nouns is truly arbitrary (as has been claimed), and (2) whether the grammatical genders assigned to nouns have semantic consequences. In the first study, English speakers' intuitions about the genders of animals (but not artifacts) were found to correlate with the grammatical genders assigned to the names of these objects in Spanish and German. These findings suggest that the assignment of genders to nouns is not entirely arbitrary but may to some extent reflect the perceived masculine or feminine properties of the nouns' referents. Results of the second study suggested that (1) people do include gender in their conceptual representations of inanimate objects, and (2) people's ideas about the genders of objects are strongly influenced by the grammatical genders assigned to these objects in their native language. Spanish and German speakers' memory for object-name pairs (e.g., apple-Patricia) was better for pairs where the gender of the proper name was congruent with the grammatical gender of the object name (in their native language), than when the two genders were incongruent. Since both groups performed the task in English, it appears that the semantic representation of gender (once it has been established) is not language-specific. These results suggest that grammatical gender may not be as arbitrary or as purely grammatical as was previously thought.

### Acknowledgments

This research was funded by an NSF Graduate Research Fellowship to the first author. Partial support was also provided by NIMH research grant MH-47575 to Gordon Bower. The authors would like to thank Michael Ramscar, Herbert H. Clark, Barbara Tversky, and Gordon Bower for many insightful discussions of this research, and Jill M. Schmidt who was indispensable in assembling the stimuli.

### References

- Boroditsky, L. (1999). First-language thinking for second-language understanding: Mandarin and English speakers' conceptions of time. *Proceedings of the 21st Annual Meeting of the Cognitive Science Society*, Vancouver, BC
- Braine, M. (1987). What is learned in acquiring word classes—a step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 65-87). Hillsdale, NJ: Erlbaum.
- Bowerman, M. (1996). The origins of children's spatial semantic categories: cognitive versus linguistic determinants. In J. Gumperz & S. Levinson (Eds.), *Rethinking linguistic relativity*. Cambridge, MA: Cambridge University Press, 145-176.
- Bowers, J., Vigliocco, G., Stadthagen-Gonzalez, H. & Vinson, D. (1999). Distinguishing language from thought: Experimental evidence that syntax is lexically rather than conceptually represented. *Psychological Science*, *10*(4), 310-315.
- Choi, S., & Bowerman, M. (1991). Learning to express motion events in English and Korean: The influence of language-specific lexicalization patterns. Special Issue: Lexical and conceptual semantics. *Cognition*, *41*, 1-3, 83-121.
- Gentner, D. & Imai, M. (1997). A cross-linguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition* *62*, 2, 169-200.
- Heider, E. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, *93*, 10-20.
- Jakobson, R. (1966). On linguistic aspects of translation. In R.A. Brower (Ed.), *On translation*. New York: Oxford University Press, 232-239.
- Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist*, *86*, 65-79.
- Konishi, T. (1993). The semantics of grammatical gender: A cross-cultural study. *Journal of Psycholinguistic Research*, *22* (5), 519-534.
- Levinson, S. (1996). Frames of reference and Molyneux's question: Crosslinguistic evidence. In P. Bloom & M. Peterson (Eds.), *Language and Space*. Cambridge, MA: MIT Press, 109-169.
- Lucy, J. (1992). *Grammatical categories and cognition: a case study of the linguistic relativity hypothesis*. Cambridge, England: Cambridge University Press.
- Lucy, J., & Shweder, R. (1979). Whorf and his critics: Linguistic and nonlinguistic influences on color memory. *American Anthropologist*, *81*, 581-618.
- Mullen, M. K. (1990). Children's Classification of Nature and Artifact Pictures into Female and Male Categories. *Sex Roles*, *23* (9/10), 577-587.
- Sera, M., Berge, C., & del Castillo, J. (1994) Grammatical and conceptual forces in the attribution of gender by English and Spanish speakers. *Cognitive Development*, *9*, 3, 261-292.
- Slobin, D. (1996). From "thought and language" to "thinking for speaking." In J. Gumperz & S. Levinson (Eds.), *Rethinking linguistic relativity*. Cambridge, MA: Cambridge University Press, 70-96.
- Twain, M. (1880). *A Tramp Abroad*. Leipzig : Bernhard Tauchnitz.
- Whorf, B. (1956). *Language, Thought, and Reality: selected writings of Benjamin Lee Whorf*, ed. J.B. Carroll. Cambridge, MA: MIT Press.

## Appendix A

Materials used in the study:

### Proper names

Christopher	Christina
Daniel	Danielle
Paul	Paula
Brandon	Brenda
Eric	Erica
Karl	Karla
Claude	Claudia
Phillip	Phyllis
Harry	Harriet
Donald	Donna
Alexander	Alexandra
Patrick	Patricia

Object-names	<u>Grammatical Gender</u>	
	Spanish	German
apple	(f)	(m)
arrow	(f)	(m)
boot	(f)	(m)
broom	(f)	(m)
fox	(f)	(m)
frog	(f)	(m)
moon	(f)	(m)
spoon	(f)	(m)
star	(f)	(m)
toaster	(f)	(m)
whale	(f)	(m)
pumpkin	(f)	(m)
bench	(m)	(f)
cat	(m)	(f)
clock	(m)	(f)
disk	(m)	(f)
drum	(m)	(f)
fork	(m)	(f)
mouse	(m)	(f)
snail	(m)	(f)
sun	(m)	(f)
toilet	(m)	(f)
toothbrush	(m)	(f)
violin	(m)	(f)

# In search of the minority default: the case of Arabic plurals

**Sami Boudelaa (sami.boudelaa@mrc-cbu.cam.ac.uk)**

MRC Cognition and Brain Sciences Unit, 15 Chaucer Road,  
Cambridge, CB2 2EF, UK

**M. Gareth Gaskell (g.gaskell@psych.york.ac.uk)**

Department of Psychology, University of York,  
Heslington, York YO10 5DD, UK

## Abstract

A Minority-default inflectional system is one in which a regular affixational process (e.g., the plural morpheme ~s in English) applies to fewer forms in the language than the irregular stem modifying process (e.g. the umlauting in “foot-feet”-like pairs). Following the work of McCarthy & Prince (1990), the plural system of Modern Standard Arabic has been cited as an archetype of a minority-default system with the regular sound plural involving fewer nominal forms than the irregular broken plural. On the basis of linguistic, statistical and distributional evidence, we argue that this assertion is wrong. We point out that while both broken and sound plural have qualitatively limited productivity, the latter is quantitatively the more productive process. Furthermore, the diversity of regularly inflected phonological forms ensures that they will be treated as the default by a connectionist model. In the light of these findings we argue that a good model of morphological processing should motivate the observation that so few of the world’s languages use minority defaults.

## Introduction

A major debate in psycholinguistics revolves around the question of how human language users employ finite means to produce large numbers of words and utterances. In order to deal with this generic question, several more specific questions need to be spelt out. One such specific question is whether or not the structural properties of regularly and irregularly inflected words correspond to their representational and processing properties. Focusing on the representational format would lead one to tackle the question of whether morphologically complex words are represented as full forms or as decomposed morphemes (Marslen-Wilson et al., 1994). Focussing on the processing aspect of the equation would lead one to raise the same question from a different standpoint, namely whether morphologically complex words are formed via a symbolic rule-based mechanism operating on grammatical categories or via a memory-based associative network that extracts probabilistic contingencies between them (Rumelhart & McClelland, 1986; Pinker & Prince 1988; Marslen-Wilson & Tyler, 1998).

The acquisition of the English past tense has been extensively studied in an attempt to decide between the different approaches to this problem. The literature on the subject provides at least three different models. The first, and most traditional assumes that the regular past forms in English like “walk-walked” are formed by a rule, whereas irregular past tenses like “eat-ate, give-gave” are learned individually

by rote (Berko, 1958). Because it fails to explain the sub-regularities among the irregular verbs and the expansion of irregular inflection to phonologically similar nonce forms, this view has largely been superseded by a second model which claims that a rule-governed process inflects all the regular forms while an associative memory takes care of all the irregular forms. The associative memory identifies the irregular forms and blocks the default process from applying to them (Pinker, 1991; Pinker & Prince, 1988). The third model is a connectionist one, which dispenses with explicit rules and assumes that language learning is better accounted for using a single mechanism, namely a network of interconnected units (Rumelhart & McClelland, 1986). Both regular and irregular forms are inflected by this network, with responses to novel forms depending on their phonological similarity to familiar patterns (Plunkett & Marchman, 1991).

Both dual-route models and connectionist networks are able to handle an inflectional system like English because of its distributional characteristics. The English system is one in which about 95% of the forms are regularly inflected. This is an unproblematic situation for a dual-route model, which deals with the small number of irregulars via associative memory and the rest via a default rule. A connectionist network would also exhibit relative ease handling such cases. The network would store information about all forms and the preponderance of regular forms would trigger a regularisation process, by virtue of the fact that any novel form is more likely to resemble a regular form than an irregular one. Proponents of the dual-route model have argued that a dual mechanism can also deal satisfactorily with linguistic systems where the default is a minority, such as the inflection system in German (Clahsen, 1999). This is because rule-like behavior does not need to be contingent on the default pattern applying to a majority of the forms in the language. Rather, a default can be defined, the argument goes, even in terms of the least frequent patterns, because this process depends on applying the same procedure to different items bearing the same symbol (Clahsen, 1999). Conversely, a connectionist network was predicted to be unable to simulate people’s regularisation of novel forms in a minority-default system like German.

The Arabic plural is perhaps the most widely cited example of a minority default system (McCarthy & Prince, 1990; Hare, Daugherty & Seidenberg, 1992; Ravid & Farah, 1999). For this reason it was used as a litmus test by Plun-

kett and Nakisa (1997) who found that a connectionist network can model generalisation behaviour to both regular and irregular patterns, despite the absence of a default rule. One of our aims here is to take issue with the position that Arabic has a minority default plural system, and show that it hinges on an inaccurate description of the language. In order to come to grips with this claim we will begin by laying out the morphological system of Modern Standard Arabic and argue that this language does not exhibit a minority-default, using linguistic and corpus analyses. Second, we will examine the phonological distribution of Arabic nominal forms using a more representative sample than the one used by Plunkett and Nakisa (1997). All these sources of evidence converge on the idea that the Arabic plural system has a majority default of the type learnable by a connectionist model. We conclude by considering why minority default systems seem scarce across world languages.

## The Morphological System of Arabic

Traditionally Arabic surface forms are analysed as consisting of two abstract morphemes a root and a word pattern<sup>1</sup>. The root usually comprises three consonants and carries semantic meaning, while the word pattern contains vowels and conveys syntactic information. According to this approach, the representation of a surface form such as [nuqil] “*be moved*” will consist of the root {nql}, and the word pattern {fuil} where the letters “f, l, i” indicate the slots into which the root consonants map.

The morphology of Arabic falls into two relatively distinct parts (Bohas & Guillaume, 1984). The first consists of primitive nouns that are thought to be unrelated to verbs, although verbs can be derived from them. For example, from the primitive noun [kalbun] “*dog*” the verb [kaliba] “*get infected with rabies*” can be formed. The second part relates to verb morphology and subsumes verbs proper and deverbal nouns. Verb morphology can be further divided into unaugmented and augmented verb forms. There are three unaugmented forms and 14 augmented forms of which only 9 are frequently used in Modern Arabic. As for deverbal nouns, there are about 10 types such as the active participle, passive participle, instance noun, manner noun, “*assimilated noun*”, and the “*masdar*” (Holes, 1995).

## Verb Morphology

Verb morphology with its two components is the most productive part of the language in the sense of being the main source of most of the *transparent derivatives*<sup>2</sup>. For example, combining the root {xrʒ} “*go out*” with the pattern {faʕal}, produces the form [xaraʒ] “*go out*”. The same root can be further combined with as many as 5 augmented patterns yielding the following surface forms: [xarraʒ] “*move out*”

<sup>1</sup> Within the framework of multilinear phonology the word pattern is further broken down into a vocalic morpheme and a skeletal morpheme (McCarthy, 1981).

<sup>2</sup> Some authors claim that “as many as 400 different surface word forms can be derived from some trilateral verbs” (Xasaara, 1994, p. 134)

[ʔaxraʒa] “*take out*”, [taxarraʒ] “*graduate*” [taxaaraʒ] “*disengage*”, [ʔistaxraʒ] “*extract*”. From each of these forms a host of deverbal nouns can be derived. For example, the masculine active participle [xaariʒun] can be derived from the unaugmented surface form [xaraʒ]. Also, the following active participles can be derived respectively from each of the augmented verb forms above: [muxarriʒ], [muxriʒ], [mutaxarriʒ], [mutaxaariʒ], [mustaxriʒ]. Passive participles can also be formed from these verb forms. In addition to this, an “*instance noun*”, a noun denoting that the action takes place only once, [xarʒatun] “*one departure*” can be obtained from the verb [xaraʒ], the noun [taxaaruʒ] can be derived from the verb [taxaaraʒ], the noun [ʔistixraaʒ] can be derived from the verb [ʔistaxraʒ] and so on. This pattern of productivity holds even for verbs that are originally derived from primitives. Thus from the primitive noun [kalb] “*dog*” the verb [takaalab] “*to rave*” is derived and from the latter an active participle [mutakaalib] “*someone who raves*” and a noun [takaalub] “*raving*” are formed. Similarly, loan words like [talifuun] “*telephone*” can be used to derive verbs such as [talfan] “*to telephone*”, and an active participle like [mutalfin] “*phone-caller*”.

## Nominal Morphology

Arabic nouns undergo various morphological alterations of which the most frequent is perhaps pluralization. This is achieved either via suffixation or via pattern modification. In the first case, known as sound pluralization, the suffix ~uun is added to masculine nouns (e.g. [naazihun-naazihuun] “*successful*” male) while ~aat is appended to feminine nouns (e.g. [naazihatun-naazihaat] “*successful*” female). In the second, often referred to as broken pluralization, the pattern of the singular noun is dramatically altered and in some cases some of its consonants are lost (e.g. [ʃunquud-ʃanaqiid] “*cluster*” [ʃandaliib-ʃanaadil] “*nightingale*” (Murtonen, 1964; Xasaara, 1994; Holes, 1995). Sound pluralization is considered as regular inflection because it involves little or no allomorphy while broken pluralization is irregular because it is rich with allomorphic variations.

McCarthy & Prince’s (1990) work on the broken plural in Modern Standard Arabic has promulgated the idea of Arabic having a minority default system of pluralization. According to them the sound plural is “*systematically found only with the following short list: proper names; transparently derived nouns or adjectives such as participles, deverbals and diminutives; non-canonical or unassimilated loans and the names of the letters of the alphabet*” (McCarthy & Prince, 1990: p. 212). Phrased as such, the above claim is misleading because it fails to distinguish between qualitative and quantitative productivity. The distinction between these two aspects of productivity rests on the difference between the number and/or the strength of the constraints weighing on a particular morphological process (Aronoff & Anshen, 1998). Perhaps an English example will help to bring our point home. The suffix ~ity is qualitatively productive but quantitatively unproductive. This is because it tends to be appended preferentially to adjectives ending in suffixes like ~ible, ~able, ~ic, ~id etc. Conversely the suffix ~ness, is quantitatively productive because it is subject to fewer constraints and is not restricted to follow a



limited set of suffixes (Aronoff & Anshen, 1998). Arabic sound and broken pluralization processes lend themselves readily to a description in terms of a distinction between qualitative and quantitative productivity. Both are subject to few constraints. Sound pluralization is restricted to a set of nominal forms that must meet formal (e.g., length in syllables) and syntactic criteria (e.g. being preferably adjectives). But broken pluralization is subject to even more rigid and more numerous formal (e.g. length and syllabic structure) and syntactic criteria (e.g. being preferably a substantive). Quantitatively, however sound pluralization would not be a minority case even if it were found *only* with transparent derivatives. Transparent derivatives, as we will shortly show, correspond to the most productive part of the language. Additionally, sound pluralization affects systematically all recent loan words comprising more than three letters like [dimuqraat<sup>5</sup>iyun] “*democracy*”, [tafazatun] “*television*”.

### Type Frequency of Broken and Sound Plurals

A given trilateral root in Arabic can be productively mounted on some combination of the 9 frequent augmented word patterns to create new words. For instance, the trilateral unaugmented surface form [katab] “*write*” can be combined with as many as 7 augmented forms. Conversely, the unaugmented trilateral [ʔabaθ] “*fool around*” gives rise only to one augmented form [ʔaabaθ] “*banter*”. Although no systematic statistical work on the number of augmented and unaugmented verb forms is available in Arabic, one may safely hypothesise that trilateral roots can yield on average at least three surface forms. Confining our analysis to active and passive participles in the masculine and feminine forms, we can plausibly say that each of the augmented forms gives rise to at least 4 deverbal forms. There are 11978 roots of which 7597 are trilaterals, 4081 are quadrilaterals and 300 are quinquilaterals (Moussa, 1996). Assuming that the derivation of four masculine and four feminine deverbal surface forms from each root is not an overestimate, the trilateral roots alone will yield as many as 91164 surface forms that take a sound plural. If we consider the derivatives from quadrilateral and quinquilaterals, this estimate will increase greatly.

It is true that some transparent derivatives like “assimilated nouns” and lexicalized active participles often pluralize in the broken way. This does not mean that nouns taking a broken plural will outnumber those pluralizing regularly because for almost every assimilated noun or indeed for any other noun that has a broken plural, there is either a diminutive form, a feminine form or both, and these take a sound plural. Thus the assimilated noun [ʔaaqir] “*barren*” has the broken plural [ʔawaaqir], whereas its diminutive [ʔuwaiqir] has the sound feminine plural [ʔuwaiqiraat]. Likewise, the primitive noun [qird] “*monkey*” has a broken plural [quruud] but its feminine form [qirdatun] “*female monkey*” has a sound plural form [qirdaatun].

The type of pluralization taken by a particular nominal form may be driven by semantic considerations as well. Many active participles, derived from roots mounted on the unaugmented pattern, like [kaatib] may pluralize regularly

or irregularly depending on whether they function as a substantive or as an adjective. Used as a substantive to denote a permanent activity or quality, they form a broken plural. Thus when the token [kaatib] is used in the sense of “*author*”, it has the broken plural [kuttaab]. By contrast, when it is used in the sense of “*someone who writes*”, it pluralizes regularly as [kaatibuuna].

In order to support our claim statistically, we analysed all nouns listed in the “Basic Lexicon of Modern Standard Arabic” (henceforth BLMSA), which consists of the 3000 most frequent words in the language (Khoulooghli, 1992). The BLMSA is based on a statistical analysis of more than 200,000 words drawn from newspapers and literary work throughout the Arab world. The author reports a total of 1670 nominal forms (i.e. nouns and adjectives).<sup>4</sup> Of these, 666 tokens are explicitly listed as taking a broken plural and 610 as taking a sound plural (215 masculine and 395 feminine). For the remaining 394 words, the author lists either the plural form (sound or broken) with no mention of the singular or vice versa. The 394 words divide into 357 singular forms for which the corresponding sound plural is not listed, 11 sound plural forms without their relevant singular forms, 20 singular forms without their corresponding broken plurals, and 6 broken plurals for which the corresponding singulars are not listed. Possibly the author lists only the singular or the plural of these forms because the other is not one of the 3000 most frequent words of the language. However, this does not mean that they would be *hapax legomena* in a larger database if this were available. Indeed many of the unlisted words like [murabbaʔaat] “*squares*” and [ʔaaliha] “*gods*” the respective sound and broken plural forms of the listed singular forms [murabbaʔ] “*square*” and [ʔilaah] “*god*” are part of the familiar repertoire of words that can be encountered even in children’s books.

In sum, of the 1670 most frequent nominal forms of the language almost two thirds, 978 nouns, pluralize via suffix addition and the remaining forms take a broken plural. This is important for two reasons. First, testing a few random samples taken from the BLMSA shows that it has an average coverage of 75 to 95% of any Modern Arabic text. So if the BLMSA is representative, we can infer that about 56% of Arabic words are nouns (i.e. lexical nouns and adjectives) and most critically that about 59% of all nouns of the language take a sound plural while only 41% take a broken plural. Because BLMSA is a sample of the most frequent words, it is likely that lower frequency nouns are even more skewed towards the regular plural.

In view of this, it seems untenable to consider Modern Standard Arabic as an example of a minority-default system. Just why this stance has come to be held is an offshoot of Arabic lexicographers’ work that lists only the broken plural forms because they are unpredictable.

In this section, we have laid out linguistic and corpus-based evidence that the Arabic plural system is not a minority-default. The affixational process involves far more words than the templatic processes, although the proportion

<sup>4</sup> The remaining 1330 items listed in the BLMSA comprise verbs and the closed classes of particles, prepositions and conjunctions.

is still not as high as the English past tense system, with 95% regulars (Daugherty & Seidenberg, 1992).

### The Phonological Distribution of Sound and Broken Plurals

The supposed status of the Arabic plural as a minority default system has resulted in claims that it cannot be accommodated by a connectionist model. Plunkett and Nakisa (1997) examined this claim using statistical analyses and connectionist simulations. They noted that a minority default is not necessarily a problem for a connectionist account provided there is an even distribution of regulars and relatively tight clustering of irregulars in the phonological space spanned by the uninflected forms (cf. Hare, Elman & Daugherty, 1995). In cases where irregulars share strong phonological resemblances, but the minority of regulars vary widely in their phonological form, a multi-layered connectionist network can develop “distributional default” behaviour. Although the irregulars may be dominant in number, they are concentrated in relatively small pockets of the network’s input space, and so are unlikely to be similar to novel items. Instead, most novel inputs will be more similar to a regular item, and so will be inflected in the same way leading to default behaviour.

Plunkett and Nakisa (1997) examined the phonological distribution of Arabic singulars in this respect using a set of nouns drawn from the Wehr Arabic Dictionary (Wehr, 1976). On the basis of statistical analyses of the distribution of singulars in phonological space, they argued that the Arabic plural system does not provide a basis for developing a distributional default. Instead of evenly spanning the phonological space, the sound plurals appeared to be even more phonologically coherent than many of the broken plural sets. A connectionist network trained on the singular to plural mapping for these items would therefore be unlikely to develop behaviour resembling a default rule.

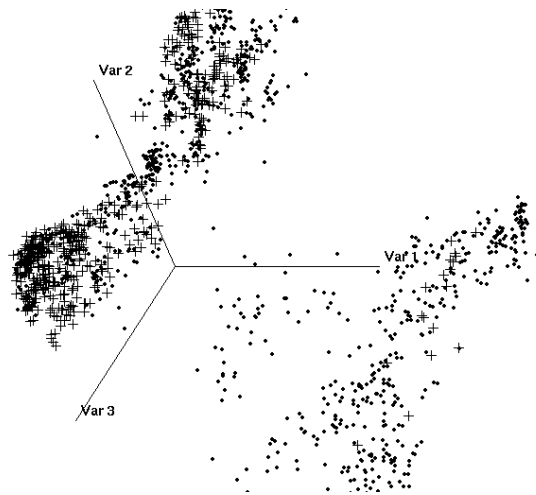
Plunkett and Nakisa (1997) also showed that despite the absence of the conditions necessary for developing default behaviour, a connectionist model was able to learn and generalise the pluralization task rather well. In fact generalisation (i.e., performance on untrained patterns) in the network was superior to a dual-route model irrespective of the division of labour between the two routes. In effect, the network was performing adequately with neither a majority nor a minority default.

The work of Plunkett and Nakisa (1997) is important because it marks out the conditions necessary for default-like behaviour in a connectionist model of morphological processing. The behaviour of a connectionist system does not just depend on the numbers of regular and irregular items. It also depends on the distribution of these items in phonological space. However, with respect to the specific case of Arabic, there are still many unanswered questions. Since the data-source used by Plunkett and Nakisa (1997) has, as we have argued, a bias in the proportions of sound and broken plurals, the detailed predictions made in their paper may be unfounded. We have already argued that sound plurals are in the majority in Arabic, but this is not enough to demonstrate that a connectionist system will learn

to treat them in a default-like way. The phonological properties of a representative sample of the language must also be examined in order to assess the basis for a distributional default. If it turns out that both sound and broken plural classes are phonologically well defined and compact, then a “no default” system would be predicted on the basis of Plunkett and Nakisa (1997).

The 1670 nominal forms were classified by plural type, and the 16 categories that contained 10 or more members were used in the analyses and these amounted to 1491 items. Of these, 972 took the sound plural (273 masculine forms and 699 feminine forms). The remaining 519 items were members of 14 broken plural subtypes, containing between 13 and 121 nouns). In order to examine the phonological similarities between the members of these groups, each singular form was translated into a featural code based on a slight modification of the template system of Plunkett and Nakisa (1997). First, the phonemic transcriptions for the singular forms were aligned to an 18-slot template consisting of alternating consonants and vowels. The slots were filled from left to right, with consonants placed in consonant slots and vowels in vowel slots. When a word contained two consonants or vowels in a row, this procedure led to an empty slot between them, but it also ensured that as far as possible the representations reflected similarities between words by comparing like with like. For example, the representation of /jurʔhun/ “scar” in the template was jur-HUn----- . The slot-based phoneme representations were then translated into featural representations in order to capture similarities between different phonemes. The outcome of this transformation was an 18 slot x 20 features (360 dimensional) vector for each singular form. Taking the dataset as a whole, the vectors span a 360 dimensional space, in which each word form is a point. The issue we address is how the different plural classes are distributed in this multidimensional space.

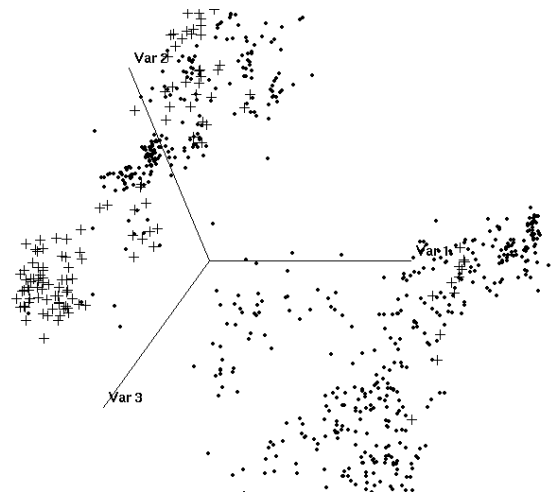
Principal components analysis takes a set of points in a high dimensional space and determines a smaller set of orthogonal vectors within this space that captures the greatest variation between the points. The original points can be projected on to these principal components to extract a low dimensional plot preserving the most important information from the high dimensional space and eliminating redundant dimensions. Figure 1 plots the positions of the different plural subtypes in the space defined by the first three principal components. For the sample used by Plunkett and Nakisa (1997), the sound plurals occupied relatively restricted positions in the space. For our sample, the sound plurals are fairly ubiquitous. There are many completely empty regions of the space, corresponding to phoneme combinations that are in some way badly formed, but most of the occupied regions are occupied by sound plurals, whereas the broken plurals sets are generally more coherent. Plunkett and Nakisa (1997) quantified their observations by calculating a coherence measure for each plural subtype. However, this measure is less valuable for our dataset (containing plural types of greatly varying size) because it is confounded with set size, such that larger sets will be rated as more coherent purely because of their size.



**Figure 1:** Phonological distribution of Arabic singulars across a plane through the first three principal component. Pluses mark broken plurals, dots mark sound plurals.

Instead, we looked at the relative isolation of the regular and irregular groups as a whole. Put simply, for the regulars to act as the distributional default in a connectionist model there should be a high chance that a randomly chosen non-word will be most similar to one of the existing regulars, and therefore will be processed in the same way. Each word in the language will have its own “sphere” of influence in the phonological space—if any novel form falls in this area, it will be closest to that point and will tend to be inflected in the same way.

The most influential items in the language will be the ones with the largest area of influence. We can analyse these areas by calculating, for each word in the language, the distance from the nearest neighbour (both of the same class and of any class). The class that exerts the most influence will be the one that has the most isolated members, because these words will have the greatest influence in terms of generalisation to novel forms. This analysis shows that not only are there more sound plurals in Arabic, but they are more spread out in the phonological space, and so have a greater sphere of influence. Sound plurals differ from their nearest neighbour by 4.9 features on average, whereas broken plurals differ by 3.7. This advantage is independent of the number of items in each plural class. When nearest neighbour distances are broken down by overall class, the combined effect of numerical dominance and greater area of influence becomes clear, sound plurals differ from their nearest broken plural by 12.2 features on average, whereas broken plurals differ from their nearest sound plural by 6.0 features on average. This statistic implies that it is easy to find sound plurals that are unlike any broken plural but difficult to find broken plurals that are unlike any sound plural. This finding is confirmed in Figure 2, which plots only the singular forms that are 8 or more features different from their nearest neighbour of the opposite class (68% of the sound plurals, and 25% of the broken plurals). The broken plurals are quite closely packed in tight pockets of the space, whereas the sound plurals are more spread out. This is exactly the state of affairs required for distributional default behaviour to develop in a connectionist model.



**Figure 2:** Distribution of “isolated” Arabic singulars. Pluses mark broken plurals, dots mark sound plurals.

## General Discussion

Much of the evidence relating to the debate between symbolic and connectionist accounts has stemmed from the study of the English past tense, in which regulars are numerically dominant. Proponents of the symbolic account have challenged the ability of connectionist models to deal with inflectional systems in which the default inflection is a minority. Modern Standard Arabic and German were taken as instances of languages that do not depend on the regular pattern involving the majority of forms. Connectionist simulations of minority default behavior (Hare et al., 1995; Plunkett & Nakisa, 1997) have refined the debate, by showing that minority default systems are not necessarily problematic for a connectionist model. If the distribution of regulars is sufficiently broad, then a connectionist model can develop default-like behavior (Hare et al., 1995). Even in the case where regulars are more tightly clustered, a connectionist model can learn the mapping, and perform generalization, although the regular will not become a true default (Plunkett & Nakisa, 1997). These studies emphasize the importance of phonological distribution in the analysis of linguistic systems, alongside the numerical information.

Our main point in this paper was to argue that the Arabic plural system is not a minority default, with regular sound plural applying to fewer forms than the idiosyncratic broken plural. Three sets of arguments were brought to bear on our claim. First, we have shown that while both broken and sound plural are qualitatively productive, only the latter reflects quantitative productivity. Second, the empirical investigation of the most frequent nominal forms collected from BLMSA demonstrates that sound pluralization involves almost twice as many word forms as broken pluralization. The sound plural does not have a low type frequency. Third, analyses of similarities in phonological space showed that the distribution of Arabic nominal forms follow much the same pattern as that of English verbs.

Our analysis raises a set of problems relative to current models of human language productivity. Symbolic models are perfectly compatible with languages exhibiting a minority default inflectional system, but do not provide a princi-

pled explanation for the scarcity of these cases. This follows from the assumption that the human cognitive processor manipulates symbols and does not need a majority of forms to show a rule-based behavior. So far as we know only German and Arabic are cited as current examples of such systems. As it is demonstrated above Arabic is not and Bybee (1995) offered an account that questioned the claim that German is a minority default. Note however, that from the perspective of language change we do not exclude the possibility of a linguistic system passing through a minority default inflectional system. Rather, our point is: if minority default systems are as natural and as easy to handle as symbolic models would have it, then why do they seem to be scarce?

Connectionist models, meanwhile, have responded to the challenge of the minority default. These systems are less at ease with a minority default system, since they require the regulars to have sufficient variety in their phonological form if they are to be treated as the default case. But more critically, they also offer an explanation for the lack of minority defaults in most modern languages. Hare and Elman (1995) used connectionist networks to model the diachronic changes in the verb system of Old English, which at some stage is likely to have been a minority default system. Developments in the structure of language were assumed to be the product of imperfect learning from generation to generation, modeled by generations of connectionist networks. In essence, the development of the language was one of regularization, with regulars becoming more and more dominant in each successive generation. Thus, minority defaults can be learned by a connectionist network as long as certain distributional conditions are met. Even when those conditions are met, however, the state of the language is somewhat unstable, with a diachronic movement towards majority default likely in the long term. This fits in with the observation that the vast majority of linguistic systems—including the Arabic plural—do not employ a minority default.

### Acknowledgments

We thank William Marslen-Wilson and Matt Davis for valuable comments on an earlier version of this article.

### References

- Aronoff, M. & Anshen, F. (1998). Morphology and the lexicon. In A. Spencer & A. Zwicky (eds.), *The Handbook of Morphology*, Blackwell Publishers.
- Berko, J. (1958). The child's learning of English morphology. *Word*, 14, 150-177.
- Bohas, B. & Guillaume, J. P. (1984) *Etudes des Théories des grammairiens Arabes: I Morphologie et Phonologie*: Damas, Syria.
- Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, 10, 425-455.
- Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences*, 22, 991-1060.
- Daugherty, K. & Seidenberg, M. (1992). Rules or connections? The past tense revisited. *Proceedings of the 14<sup>th</sup> annual Conference of the Cognitive Sciences Society*, 259-264.
- Hare, M. & Elman, J. (1995). Learning and morphological change. *Cognition*, 56, 61-98.
- Hare, M. Elman, J. & Daugherty, K. (1995). Default generalization in connectionist networks. *Language and Cognitive Processes*, 10, 601-630.
- Holes, C. (1995). *Modern Arabic*. Longman, London-NY.
- Khouloughli, D.-E. (1992). *Basic Lexicon of Modern Standard Arabic*. L'Harmattan, Paris.
- Marslen-Wilson, W. D., Tyler, L.K., Waksler, R. and Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, 101, 3-33.
- Marslen-Wilson, W. & Tyler, L. K. (1998). Rules, representations and the English past tense. *Trends in Cognitive Sciences*, 2, 428-435.
- McCarthy, J. J. (1981). A prosodic theory of non-concatenative morphology. *Linguistic Inquiry*, 12, 373-418.
- McCarthy, J. & Prince, A. (1990). Foot and word in prosodic morphology: The Arabic broken plural. *Natural Language and Linguistic Theory*, 8, 209-283.
- Moussa, A. H. (1996) Database for major Arabic dictionaries. *Proceedings of the 5<sup>th</sup> International Conference and Exhibition on Multilingual Computing*, Cambridge.
- Pinker, S. & Prince, A. (1988) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28, 73-193.
- Plunkett, K. & Marchman, V. (1991). U-shaped learning and frequency effects in a multilayered perceptron: implications for child language acquisition. *Cognition*, 38, 3-102.
- Plunkett, K. & Nakisa, C. (1997). A connectionist system of the Arabic plural system. *Language and Cognitive Processes*, 12, 807-836.
- Ravid, D., Farah, R. (1999). Learning about noun plurals in early Palestinian Arabic. *First Language*, 19, 187-206.
- Rumelhart, D. & McClelland, J. (1986). On learning the past tense of English verbs. In D. Rumelhart & J. McClelland (eds) *Parallel distributed Processing*, Vol. 2. Cambridge MA: MIT Press.
- Xasaara, M. (1994). *Translation and Linguistic Development*. Damascus. (in Arabic).

# Representing Categories in Artificial Neural Networks Using Perceptually Derived Feature Networks

**Robert B. Branstrom (branstrm@socrates.berkeley.edu)**

Department of Psychology; 3210 Tolman Hall  
University of California  
Berkeley, CA 94720

## Abstract

How might categories be represented in artificial neural networks while satisfying biological constraints? This article proposes using feature networks, an architecture based on two types of neural organization in perceptual systems, receptive fields and topographic representation. Using these two organizing principles, category features are represented in distributed networks that allow precise, graded or probabilistic interpretations. Simulations are illustrated that show these networks have characteristics consistent with human behaviors of assimilation, contrast, and chunking. A brief discussion and simulation show how these feature networks can be combined associatively to form complex multiple-feature categories. Implications of the architecture for representation and the nature of symbol processing are discussed.

## Introduction

Regardless of the nature of the representation (i.e., visual image, verbal, etc.), categories are a foundational aspect of higher level cognition. The nature of categories remains a topic of considerable debate. The classical, or Aristotelian, view is that characteristics or traits define categories: things which have those characteristics are in the category and those which do not are not. This is simplistic, because sometimes something is in a category but does not have all necessary characteristics. For example, a three-legged animal that chases cats and cars would still be classified as a dog, even if it doesn't have the requisite four legs. Two approaches, both dealing with uncertain information, have evolved to address this problem. The first approach is probabilistic, asserting that something may be in a category if its characteristics are likely, rather than necessary. Thus the three-legged dog is still a dog because dogs usually, but not always, have four legs. The second approach applies the concept of graded structure (Rosch, 1973), asserting that membership in the category is a matter of degree, not an all-or-nothing feature. Thus a three-legged dog would still be a dog, albeit not as good an example as a four-legged dog. The condition "has four legs" is only partly satisfied, so the animal is not as good an example of a dog.

What approach might be taken to model categories? The classical view can be represented by formal set theory. Modifications to this view have been made to accommodate the probabilistic view and the graded structure view. In the probabilistic view, something is in the category if it has, say, eight of the necessary 10 conditions (Medin & Smith, 1984). The graded structure view has been approximated by fuzzy set theory (Zadeh, 1965). However, these views were developed for their formal properties, not their biological realism, so they don't offer plausible mechanisms that might underlie categorization processes.

An approach that steps closer to the biological structures of the brain is connectionism. Loosely, connectionist (or artificial neural network) models, assert that the brain is composed of many highly interconnected neurons, and that the processing power of the brain comes from these many connections. Network models of categorization typically represent categorical structure as a set of nodes representing characteristics (cf., Anderson, 1995). The characteristic may be absent or present (valued at 0 and 1, respectively). This vector of characteristics can also have graded values between 0 and 1. These values could represent either the probability or degree of the characteristic being present.

While these network models of categorization have useful functional characteristics, it's generally accepted that they still do not represent an approach that is close to the brain's actual organization. Among other things, real brains are expected to have more distributed representations for high level concepts. Anderson (1995, p. 345-6) proposed a number of principles to guide the development of "natural data representations," based on what is known about vertebrate nervous systems. These are worth summarizing here:

1. Similar events should give rise to similar representations.
2. Things should have separate representations if they need to be separated, thus categories could be separated by their features.
3. If something is important it should be represented by multiple elements.
4. Preprocess information as much as possible in the hardware.

5. Make the representation flexible so it is not problem specific.

Anderson also asserts (p. 346) that it would be easy to use "rather crude spatial means--say, spatially organized excitation and inhibition--to emphasize or deemphasize one or another aspect of the computation." Following Anderson's guidelines, this article proposes a network model of categorical and conceptual representation in which each feature is represented by a set of spatially organized nodes. The model accommodates both probabilistic and graded structure theories. The paper is organized as follows. First, two key structures of brain organization in perceptual systems are introduced and adapted for representation of category features. Then a number of simulations are provided to illustrate key behavioral characteristics of the features model. A proposal is then made for how these feature networks could be interconnected to provide an aggregate model of a category or concept. Finally, some implications of the model for cognitive science are discussed.

### Representing single attributes

There are two common characteristics of perceptual systems that are spatially based. The first is the organization of sensory inputs using receptive fields. Receptive fields are sets of input cells that are interconnected such that closer cells have a common effect (excitatory or inhibitory) on the next level of processing. More distant cells have the opposite effect. In two dimensions, these are described as center-on, surround-off if the closer cells are excitatory, or center-off, surround-on if the closer cells are inhibitory. The second characteristic of perceptual systems is analogical representation of the physical world in neural structure. In visual and haptic systems this is spatially based topographic representation, and in the auditory system it is frequency based tonographic representation. In both cases, the principle is the same: values close to each other in the physical world are close to each other in the neural structure.

Sometimes these structures are combined, with rows of interconnected receptive fields. In the visual system, this architecture is responsible for the well-known effect of Mach bands, in which differences in contrast in input data are enhanced at edges to increase contrast sensitivity. This effect is illustrated in Figure 1. The lower graph of the figure shows the specific inputs to each cell. The upper graph shows the output pattern across many cells, including the enhanced contrast where the input pattern changes.

This model applies the same architecture (rows of interconnected receptive fields) to features of categories and concepts. Two important observations are important here. First, features are usually scalar in nature, i.e., they carry ordinal (and sometimes higher level) information.

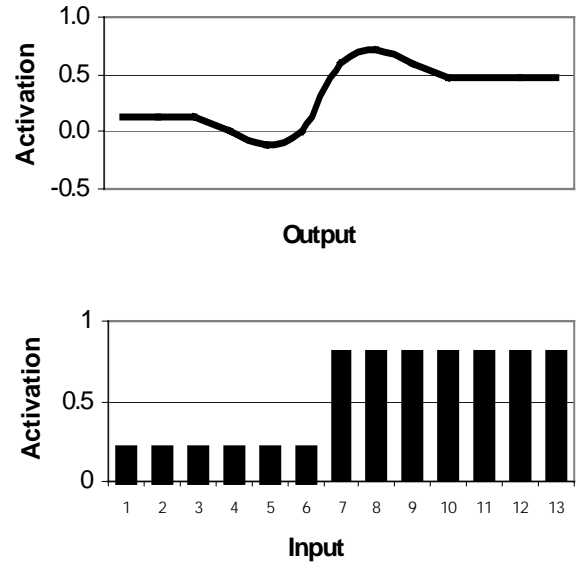


Figure 1: Edge contrast--Interconnected receptive fields enhance differences in input values at the edge where the difference occurs.

For example, dogs typically have fur. This can be represented on a scale from no-fur (Mexican hairless) to heavily furred (St. Bernard). Second, characteristics may be precise (24 inches tall) or vague (about 24 inches tall). This model allows for both of these characteristics. The ordered nature of a feature (i.e., the degree to which it holds) is mapped topographically onto the ordered organization of the nodes in the network. Precise values are represented as single nodes and vague values are represented as a cluster of adjacent nodes.

Several comments are in order before describing the model more specifically. First, the use of conceptual topographic mappings (as compared to physical or spatial topographic mappings) shouldn't be surprising if we take seriously the claim of evolutionary biologists, who argue that the easiest way to create a new structure is to borrow an old one. Second, representations of number are assumed to be at the level of an interval scale, so that both the order and distance between nodes is relevant to the representation. Third, nodes in the model's feature network are not suggested to be at the level of neurons, nor are they intended to be physically adjacent to each other. The *organization* of the nodes is the important factor; if this architecture holds in real brains it is expected that each node would be made up of many neurons and that connections would be distributed over wide areas. Finally, it should be noted that the idea of distributing features over multiple nodes was used by Shultz and Lepper (1996) to model cognitive dissonance. They distributed features across two-node polarized pairs.

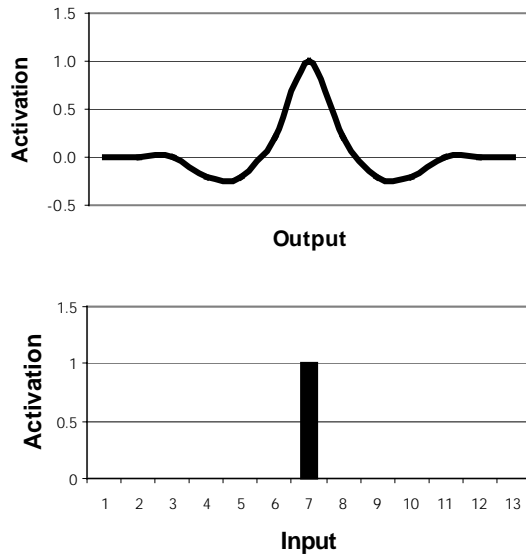


Figure 2: Point-valued representation--Input to a single node results in a characteristic "Mexican hat" output pattern.

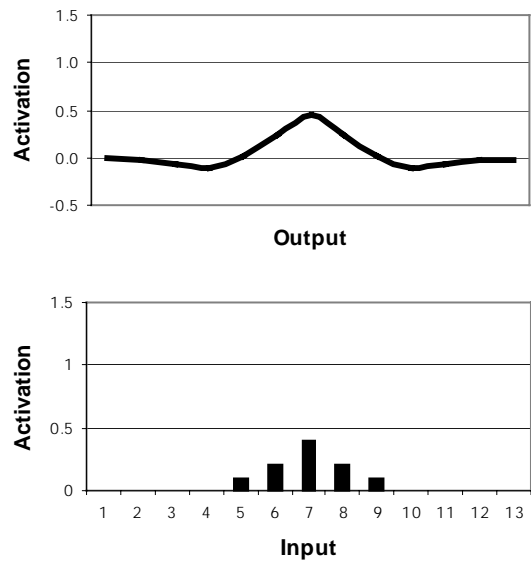


Figure 3: Vague-valued representation--Input to several adjacent nodes results in the same output pattern, but one that is more dispersed.

## The model

The model was created on a spreadsheet. Specifically, a one-dimensional row of nodes (cells) was used to represent a feature. Each node's activation was calculated as the sum of its input and the weighted-sum of the inputs of the six nearest nodes. Neighboring node inputs were all weighted at 0.2 of their actual value, and were positive for adjacent nodes and negative otherwise. In other words, the neural representation was a set of one-dimensional, overlapping, center-on/surround-off receptive fields. Inputs are modeled as values from 0 to 1, and outputs can be either positive or negative. (Although this latter effect is neurally unrealistic--neurons don't have negative activations--it is assumed this is reasonable given the usual positive base activation rate, which may be reduced. The zero base rate is used for simplicity of exposition.)

## Point-valued vs. vague-valued representations

Representations may be either point-valued or vague. This is modeled as either a single input or input spread across several nodes. Figure 2 shows a point-valued representation and Figure 3 shows a vague-valued representation. In both cases, the effects are similar: from the center of input the activation spreads slightly to neighboring cells, with closer cells being less activated than the central point and further cells being inhibited to negative values.

Vague representations may be interpreted as either probabilistic or graded. Thus, in Figure 3, the input value for 7 may be interpreted as a 40% probability of 7 occurring or as 7 to degree 0.4. When interpreted as probabilities, it isn't required that these values sum to 1.

This is consistent with empirical findings on subjective estimates of probabilities (Edwards, 1961).

## Assimilation and contrast effects

In addition to probabilities, judgments of similarity are also subjective. Sherif, Taub, and Hovland (1958) found that, when comparing two weights, subjects' estimates of the weight of one item depended on the similarity of the comparison weight. When the two weights were very similar, subjects shifted their weight judgments of the test weight (relative to when there was no comparison weight) towards the value of the comparison weight. This effect (or bias) they labeled assimilation. As the difference between weights increased, subjects shifted their estimates of the test weight more than the actual changes. This effect (or bias) was labeled contrast. In short, when two items were compared, the subjective judgment of difference depended upon the amount of the actual difference. Small initial differences were reduced so the two items appeared more similar than they actually were, while larger initial differences were enhanced so the two items appeared more different than they actually were.

The feature model yields the same effects. Figure 4 illustrates two point-valued inputs that are close to each other, yet still separated by another node. Their output, however, is merged into a single lump. (In this case, the output is two-peaked. The actual shape depends upon several factors, including the number of cells between inputs, the size of the receptive fields and the value used to weight neighboring cell inputs.)

A contrast effect, which occurs when the distance between the initial inputs is increased, is illustrated in Figure 5. The contrast occurs in two ways. First, the

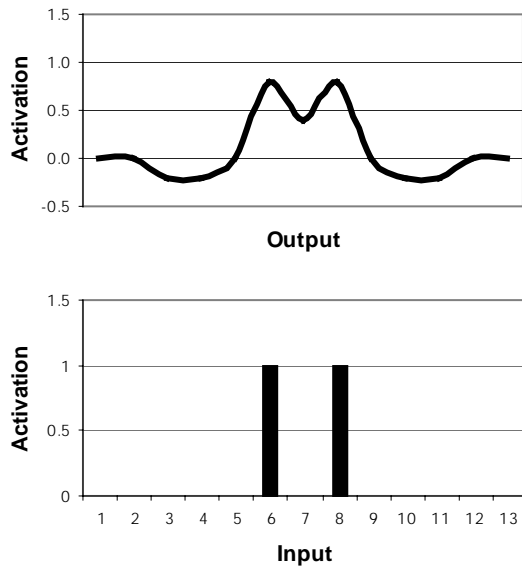


Figure 4: Assimilation effect--When two inputs are close to each other, the outputs from the feature network are merged into a single output.

activation value of the most intermediate node between the input values is inhibited below its normal base rate of zero, heightening the vertical contrast with the activation values of the nodes where the input actually occurs. Second, the "center of mass" of the output activations is shifted horizontally, slightly away from the actual value where the input occurs. This is seen in the actual output activation values. For example, input occurs at node 5, which has the highest output (activation value equal to 1). But node 4's activation is .2 and node 6's activation is 0. This asymmetry, in effect, shifts the mean activation of the representation for that input slightly away from its actual value.

Several observations are in order here. First, the effects are a result of the size of the receptive field. The assimilation effect occurs when the center (excitatory) parts of the receptive fields overlap and the contrast effect occurs when the surround (inhibitory) parts of the receptive fields overlap. Second, the assimilation effect could put a lower bound on what differences can be perceived; in effect they represent a just noticeable difference (Gregory, 1987, p. 405) for whatever is represented in the network. Third, if learning features from environmental inputs has created appropriately sized receptive fields, these effects are functionally adaptive. Essentially, assimilation allows for very small (and likely irrelevant) differences to be ignored, because they are merged and treated as one. Slightly larger (and likely more important) differences, which might not otherwise be noticeable, have their differences enhanced. (Even larger differences, which presumably would be easier to notice, aren't enhanced at all because the receptive fields of nodes receiving inputs don't overlap at all.) Fourth, these effects occur with vague representations as well as

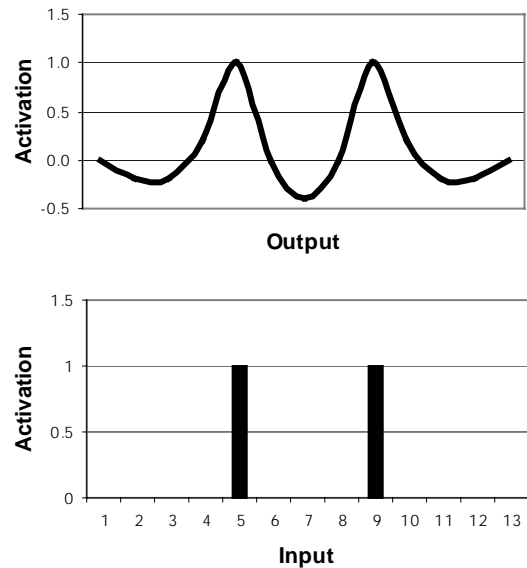


Figure 5: Contrast effects--When two inputs are slightly further apart, the outputs enhance the difference, both horizontally and vertically

the illustrated point-valued representations. Finally, this contrast effect is similar to the peak shift found in stimulus learning (Hanson, 1959). Peak shifts occur when a correctly learned stimulus (which generalizes over a symmetric gradient) must be discriminated from a new, closely related stimulus. The original stimulus gradient shifts slightly, creating an asymmetric gradient, but one that enhances discrimination. Because peak shifts are learned, they occur over time, whereas contrast effects occur immediately in real time. But both are adaptive mechanisms that enhance contrast.

### Chunking

One of the best known effects in cognitive science is chunking, the combination of several smaller bits of information into a single larger piece (Miller, 1956). When multiple pieces of information are represented as inputs in the feature network, assimilation and contrast effects provide a type of chunking. Figure 6 illustrates this, with seven inputs in two clusters of five and two, separated by one node with no input. The resulting output is two distinct "chunks," which could be called "low" and "high" on the particular feature in question.

### Multiple features

Typically, categories are made up of items with complex combinations of multiple features. This section begins an exploration of this issue by considering how feature networks might be combined to represent more complex concepts and categories. Due to the dynamic complexities of interconnected features, this section provides only a sketch of how multiple attributes might be represented.

Because each feature is represented as a network of



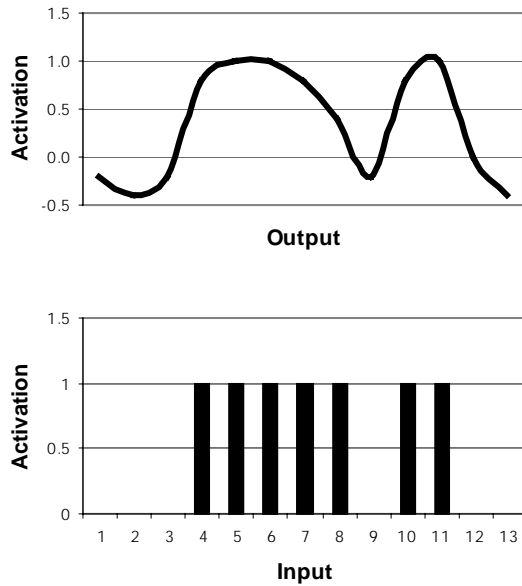


Figure 6: Chunking--When many input nodes are activated, their outputs are clustered into related groupings, such as "high" and "low."

ordered values, these networks can be related based on correlations between features. For example, the ability to fly is correlated with the presence of wings. Both are characteristics of birds and other flying species. Thus positive connections can be made between corresponding nodes in the two different attributes, such as good flying ability and good wings. Similarly, negative correlations can be made between opposite ends of the network. Figure 7 illustrates how these connections would be made from two nodes in a "flying ability" feature network to two nodes in a "wings" feature network. The straight-across connections are positive (shown as solid lines) representing positive correlations, and the diagonal connections are negative (shown as dashed lines), representing negative correlations. The double arrows on all connections represent that the connections are bilateral, that is they are mutually excitatory or inhibitory. This allows a dynamic interplay between the features, such that each node includes among its inputs the activations of the other feature's nodes from the previous iteration. These recurrent connections require a more complex formulation of the node activation functions, particularly the use of decay to dampen each node's activations over time. In the simulations presented here, correlative connections were weighted  $\pm 0.2$  and each node's activation value was decayed 80% from the prior period before computing the net input values.

When interconnected in this way, activation spreads from one feature to another. Figure 8 shows the spread of activation from an activated feature (flies well) in one period to a secondary feature (has good wings) in the following period. Two interesting characteristic of the

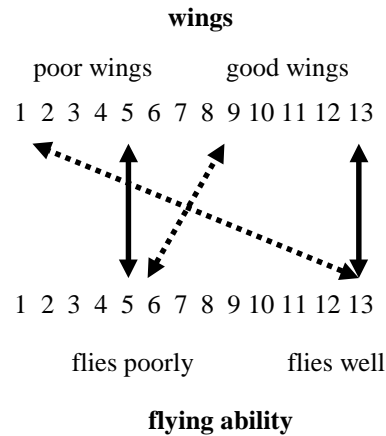


Figure 7: Multiple feature networks can be connected based on the correlations of features. Solid lines represent positive correlations and dashed lines represent negative correlations. Here, good flying ability (node 13) is positively correlated with good wings (node 13) and negatively correlated with poor wings (node 1). Double arrowheads indicate that the connections are bilateral.

secondary feature's output are the weaker level of activation relative to the activated feature and the drop in activation on the poor side of the scale, creating a contrast with the activated end of the feature network. The first characteristic is due to the weight of the correlation connections being less than one. The second characteristic results from the inhibitory connections that cross over to the opposite end of the secondary feature. The net effect of these two characteristics is that the activation level is lowered, but this is offset by an induced contrast effect.

### Categories

Treating categories as features can extend the use of interconnected feature networks to categories. For example, "birdness" is descriptive of a category, but can also be treated as a feature that is correlated with features like flight, wings, feathers, and egg laying. Because they are correlated, all the features of the category would be connected to the category network. Thus, networks for features like flying ability, wings, feathers, lays eggs, etc. would all connect to a bird feature network. When some of the features of being a bird are activated, the activation spreads to other features, including the bird feature.

Levels of categories (superordinate, basic, and subordinate) also appear to be easily computed in this structure, because the assimilation and contrast effects of the feature networks allow for generalization to higher category levels via chunking, and discrimination between lower level categories via contrast effects. Further simulations are needed to explore these dynamics.

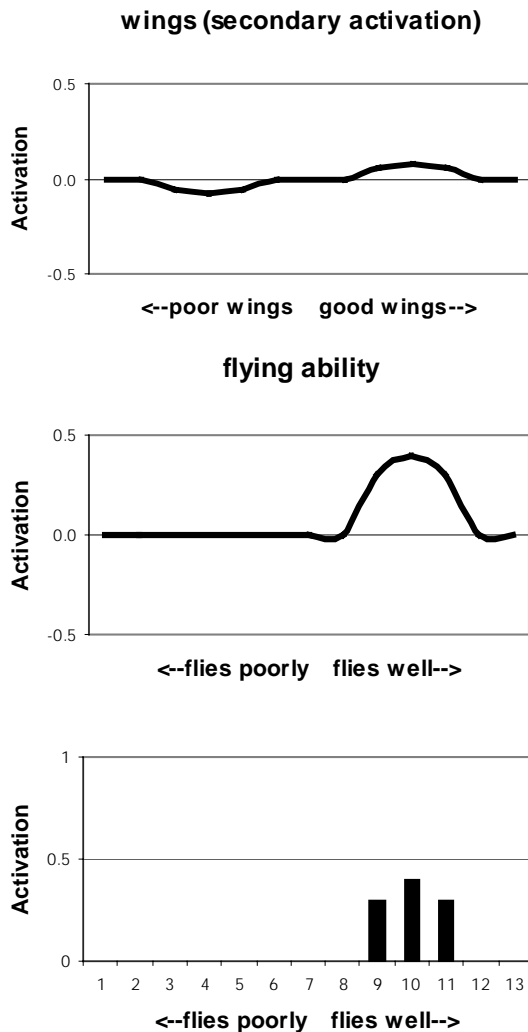


Figure 8: Activation of a correlated feature--Input (bottom graph) characterized vaguely as "good flyer" causes a similarly vague output of the "flying ability" feature network (middle graph). Both positive and negative correlations with the "wings" feature network (top graph) result in a slight contrast effect.

### Conclusions

If categories provide the foundation of higher level thought, then their representation in neural structures is an important nut to crack. The proposed method of representing categories via their features in ordered feature networks is promising because it is simple and based on known patterns of neural organization. These networks allow for crisp, vague, and probabilistic representations. Perhaps most unusual, they provide a natural way to dynamically generalize and bifurcate concepts because of their assimilation and contrast effects. While further research about the characteristics of these networks (especially more complex interconnected feature networks) is needed, their ability to perform these

basic tasks is intriguing.

Because these networks provide a means of representing symbolic information, they may shed light on the nature of symbolic thought. Those who view the mind as a symbolic processor and those who view the mind as a vast connectionist network have reached an uneasy truce. While not held universally, the view promoted by Smolensky (1988) is common: The mind is a symbol processor that runs on top of a neural network computing platform. The feature network model presented here suggests that this simple dichotomy may be unrealistic because the nature of the symbol processing itself may be important. In particular, dynamic grouping and splitting of fuzzy neural representations (i.e., generalizing and discriminating) and associations between correlated features may characterize thought more than logical operations.

### Acknowledgements

I appreciate the comments provided by Christine Diehl, Janek Nelson, and Michael Ranney on an earlier version of this paper.

### References

- Anderson, J. A. (1995). *An introduction to neural networks*. Cambridge, MA: MIT Press.
- Edwards, W. (1961). Behavioral decision theory. *Annual Review of Psychology*, 12, 473-498.
- Gregory, R. L. (Ed.). (1987). *The Oxford companion to the mind*. New York: Oxford Press.
- Hanson, H. M. (1959). Effects of discrimination training on stimulus generalization. *Journal of Experimental Psychology*, 58, 51-65.
- Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology*, 35, 113-138.
- Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for information processing. *Psychological Review*, 63, 81-97.
- Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. In T. Moore (Ed.), *Cognitive Development and the Acquisition of Language*: Academic Press.
- Sherif, M., Taub, D., & Hovland, C. I. (1958). Assimilation and contrast effects of anchoring stimuli on judgments. *Journal of Experimental Psychology*, 55, 150-155.
- Shultz, T. R., & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review*, 103(2), 219-240.
- Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11, 1-74.
- Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8(3), 338-353.

# A Sensorimotor Map of Visual Space

**Bruce Bridgeman (bruceb@cats.ucsc.edu)**

Department of Psychology; U.of Ca. Santa Cruz  
Santa Cruz, Ca. 95064 USA

## Abstract

The brain holds two representations of visual space: a cognitive representation that drives perception, and a sensorimotor representation that controls visually guided behavior. We separate spatial values in the two with the Roelofs effect: a target within an off-center frame appears biased in a direction opposite the offset of the frame. The effect appears for a verbal measure (cognitive) but not for a jab at the target (sensorimotor). Subjects might perform the jab by fixating the target during an exposure period, and jabbing where their eyes are aimed after the offset of target and frame. We show that normal humans use a context-free sensorimotor map even when they do not fixate the target: the motor map is a true 2-dimensional representation, not a 0-dimensional matching process.

## Two Visual Systems

Common sense tells us that one must accurately perceive an object's location and properties to interact effectively with it. This intuition is in error, however: several experimental designs now show that humans can engage in accurate motor behavior despite inadequate or erroneous perceptual information. Accurate perception is not required to visually guide an action.

Early experiments on separation of cognitive and sensorimotor systems showed that normal subjects could not perceive jumps of targets that take place during saccadic eye movements (a cognitive-system function). But they could still point accurately to the new locations of the same targets (a sensorimotor-system function), even if their pointing movements were controlled open-loop (Bridgeman, Lewis, Heit & Nagle, 1979). This showed that information about the new location of the target was accurate. But it was not available to perception, defined here as sensory information that is experienced, or more operationally information that can be described and remembered. If a visual stimulus is masked so that an observer denies seeing it, according to this definition the stimulus is not perceived even if it can affect later perceptual judgments or actions.

If each pathway can be probed without affecting the representation in the other, then they must be coding spatial information independently. A more rigorous way to separate cognitive and sensorimotor systems, then, is by double dissociation, introducing a signal only into the sensorimotor pathway in one condition and only into the cognitive pathway in another (Bridgeman, Kirsh & Sperling, 1981). A fixed target was projected in front of a subject, with a frame surrounding it. When the frame was displaced left or right, subjects saw illusory induced motion -- the target appeared to jump in the opposite direction. After target and frame

were extinguished, the subjects pointed to the last target position. They pointed to the same location despite the stroboscopic induced motion. But the illusion did not affect pointing, showing that the displacement signal was present only in the cognitive system.

In another condition we inserted displacement information selectively into the sensorimotor system by nulling the cognitive signal. Each subject adjusted the real target jumps until the target appeared stationary, with a real displacement in phase with the background jump equaling the induced displacement out of phase with the background. Thus, the cognitive pathway specified a stable target. Nevertheless, subjects pointed in different directions when the target was extinguished in the left or the right positions, showing that the difference in real target positions was still represented in the sensorimotor pathway. This is a double dissociation because in the first condition the apparent target displacement affected only the cognitive measure, while in the second condition the real displacement affected only the sensorimotor measure.

## A position-motion confound?

If a moving stimulus is sampled at different times for different functions, apparent dissociations might appear even though a unified visual representation underlies each function. Recently, methods have been developed, using static illusions, that can test dissociations of cognitive and sensorimotor function without possible confounding effects of motion. One method is based on the Ebbinghaus illusion, also called the Titchner circles illusion. A circle appears to be larger if it is surrounded by smaller circles than if it is surrounded by larger circles.

Aglioti, DeSouza and Goodale (1995) exploited this illusion by making the center circle into a 3-dimensional poker chip-like object and asking subjects either to judge the size of the circle or to grasp it. The grasp was adjusted closer to the real size of the circle than to its illusory size. Subjects were able to see their hands, however, so it is possible that subjects adjusted their grasp not to the non-illusory true size of the circle, but to the visible error between the grasp and the edge of the circle. The adjustments did not occur until just before the movement was completed, nearly 2 sec after it started.

Recognizing this problem, Aglioti et al. (1995) noted that calibration of grip aperture is largely refractory to visual information available during a movement, relying instead on motor programming that occurs before the movement begins. The experimental support cited for this open-loop

property, however, concerns movements to targets without illusory size modifications, so that visual recognition of grasp error and subsequent correction would not occur. The movements can be controlled open-loop because no correction is necessary. In a subsequent experiment that avoids the feedback confound, Haffenden and Goodale (1998) measured the illusion either by asking subjects to indicate the apparent size of a circle or to pick it up, in both cases without vision of hand or target. The illusion appeared for both estimations but was much smaller for grasp, indicating that the sensorimotor system was relatively insensitive to the illusion.

Another experiment contrasting grasp and perception, using the Müller-Lyer illusion, showed that while the illusion is significantly smaller when measured with grasp than with perception, there is some illusion under both conditions (Daprati & Gentilucci, 1997). Again, relatively slow grasp movements may be responsible, and vision of both hand and stimulus was allowed.

In summary, in normal subjects there is behavioral evidence for a distinction between processing in two visual streams, but we still know very little about processing in the sensorimotor pathway. With the exception of saccadic suppression and induced motion methods, all of the methods address the properties of objects rather than their locations.

A new method has produced large and consistent contrasts between cognitive and sensorimotor systems, differentiated by response measure. The dissociation is based on another perceptual illusion, the Roelofs effect: if a rectangular frame is presented off-center, so that one of its edges is directly in front of the subject, that edge will appear to be offset in the direction opposite the rest of the frame. A rectangle presented on the left side of the visual field, for example, with its right edge in the center, will appear less eccentric than it is, and the right edge will appear to the right of the subject's center (Roelofs, 1935).

We have extended and generalized this phenomenon to apply it to the study of the two-visual-systems theory. First, the frame need not have one edge centered in front of the subject; illusions occur whenever the frame is presented asymmetrically in the visual field. Second, if a target is presented within the offset rectangle, its location tends to be misperceived in the direction opposite the offset of the frame. Misperception of frame position induces illusions of target position; this is an induced Roelofs effect, but will be called simply a Roelofs effect here.

Roelofs effects can be observed reliably if subjects describe the target's position verbally, a task that addresses the cognitive system. If their task is to point to the target as soon as it disappears from view, however, they are not affected by the frame's position. This task addresses the sensorimotor system. Motor behavior for many subjects remains accurate despite the perceptual mislocalization (Bridgeman, Peery & Anand, 1997).

Though the motor task in our case is isomorphic with stimulus position, it is a communicatory act, and might be closely linked to cognitive representations. An alternative is to require an instrumental act, in which a subject must do something to the world rather than simply indicate a position

to another person. Behavior with a purely instrumental goal might be different from behavior with a communicatory goal, even if both the stimuli and the motor movements themselves are identical. Thus in our first experiment subjects jabbed a 3-dimensional target object, pushing it backward and making a clicking noise. Their intention was not to communicate anything, but only to do something to the world. With this improvement in our technique we achieve a cleaner separation of cognitive and motor systems. For a quick jab at a 3-dimensional target, rather than a pointing motion, almost all subjects show independence from Roelofs effects in immediate action, along with the previously observed robust Roelofs effects in verbal estimation of position.

Because this series of experiments follows up on earlier studies (Bridgeman et al., 1997), we were able to take advantage of the results of those studies to improve our experimental design. In the earlier data nearly all of the variance in responses as a function of target position was accounted for by a linear regression, so in the current experiments we did not need to present 5 target positions: two target positions would give us the same information, and allow us to increase the number of trials per condition.

## Experiment 1

Using these improved techniques, we begin the job of characterizing the psychophysics of the sensorimotor system.

### Method

**Subjects** Nine University of California undergraduates participated in the experiment, all right-handed with normal or corrected-to-normal visual acuity. Four were male and 5 female.

**Apparatus** Subjects sat with heads stabilized before a white hemicylindrical screen that provided a homogeneous visual field 180° wide x 50° high. A lever box located in front of the screen presented 5 white levers, each 1.8° wide, spaced 2.5° apart center-to-center (Figure 1). The center lever, marked with a black stripe, functioned as the target. Each lever was hinged at its base and spring-loaded. It activated a microswitch when pushed backward by 5mm. A long black baffle hid the microswitch assembly without revealing the position of the lever array. In the motor condition, the task was to jab the black target stripe rapidly with the right forefinger. The remaining levers served to record the locations of inaccurate responses.

A rectangular frame 38° wide x 1° in line width was projected, via a galvanic mirror under computer control, either centered on the subject's midline, 6° left, or 6° right of center. Inside the frame, the lever box occupied one of two positions, 3.5° left of center or 3.5° right of center. On each trial the frame and target were positioned in darkness during the intertrial interval. Then a computer-controlled shutter opened for one second. Stray light from the projected frame made the screen and the levers visible as well. As soon as

the shutter closed, the subject could jab the target or verbally indicate its position in complete darkness. Responses were recorded by the computer on an absolute scale (lever 1, 2, 3, 4, or 5).

**Procedure** Cognitive Measure: For the cognitive system the subject verbally estimated the position of the target spot on the center lever. The choices were 'far left', 'left', 'center', 'right', or 'far right', so that the response was a 5-alternative forced choice. Choices were identified with the five lever positions, which were centered before the subject during the instruction period, when the screen was illuminated by general room lighting and the frame was not projected. The five levers, and nothing else, were visible when the five alternatives were defined. By equating the responses with the visible levers in the apparatus, we could assign estimations in degrees of angle to the qualitative verbal responses. Interpretation of the data depends upon presence or absence of Roelofs effects, however, not on absolute calibrations of the cognitive measure. In the present series of experiments the cognitive measure serves as a control to assure that a cognitive illusion is present, differentiating the cognitive and sensorimotor systems. All quantitative results are based on the motor measure.

Subject instructions in the verbal condition emphasized egocentric calibration. Quoting from the instructions that were read to each subject, "In this condition you will be telling the experimenter where you think the target is in relation to straight ahead." Further, "If the target looks like it's directly in front of you, you will indicate this by saying 'center'." Thus center was defined in terms of the subject's body rather than the apparatus or the frame.

Sensorimotor measure: the subject rested the right forefinger on a foam pad mounted on the centerline of the apparatus just in front of the chin rest, then jabbed the target with the forefinger as soon as the target disappeared. Thus both cognitive and sensorimotor measures were open-loop, without error feedback. Before the experimental trials began, subjects practiced jabbing the target -- some were reluctant to respond vigorously at first for fear of damaging the apparatus. Subjects then received at least 10 practice trials in the jab condition and 10 the verbal condition.

Trial Execution: A computer program randomly selected target and frame positions, with the exception that an identical set of positions could not occur on two successive trials. For verbal trials, the experimenter recorded the subject's response by typing a number (1-5) on the computer's keyboard corresponding to the subject's verbal estimate. The computer recorded motor responses automatically.

In each trial one of the two target positions and one of the three frame positions was presented, exposed for one second, and extinguished. Since the projected frame provided all of the illumination, target and frame exposure were simultaneous. A computer-generated tone told the subject to respond. For no-delay trials the tone sounded as the shutter extinguished the frame, while on other trials the tone began after a 1-sec or 2-sec delay. During the delay the subject sat in darkness.

Two target positions x three frame positions x two response modes x three delays resulted in 36 trial types. Each trial type was repeated 10 times for each subject, resulting in a data base of 360 trials/subject. There was a brief rest and a chance to light adapt after each block of 60 trials.

Data were collated on-line and analyzed statistically off-line. Two-way ANOVAs were run for each subject, each response mode, and each delay condition. Factors were frame position and target position. Summary statistics were analyzed between subjects.

## Results

**Cognitive** The Roelofs effect, measured as a main effect of frame position, was significant under all delay conditions. Subjects tended to judge the target to be further to the left than its actual position when the frame was on the right, and vice versa. Six of 7 individual subjects showed a significant Roelofs effect ( $F(2,5) > 8.43$ ,  $p < 0.05$ ), and the magnitude of the Roelofs effect averaged across subjects was 2.23 deg (s.e. 0.86 deg).

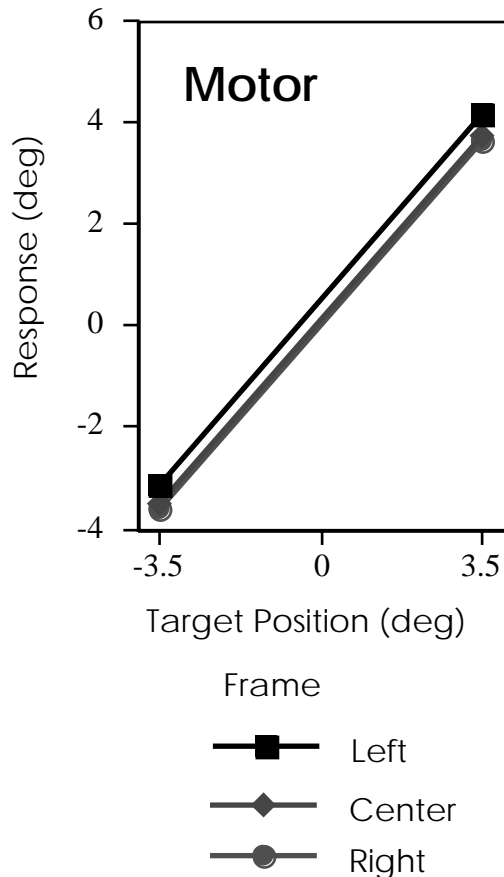


Figure 1. Motor responses, immediate action.

**Sensorimotor** The results can best be summarized with the generalization that subjects hardly ever missed the target, regardless of target position or frame position (Figure 1). Seven of 8 subjects showed no significant Roelofs effect (frame effect  $p > 0.094$ ). Averaged across subjects, the magnitude of the Roelofs effect was 20 min. arc (s. e. 22 min. arc).

**Comparing the two measures** Overall, ANOVA showed a significant difference between cognitive and motor measures ( $F_{1,43} = 12.45$ ,  $p = 0.001$ ), as expected from the robustness of Roelofs effects with the cognitive measure and the absence of Roelofs effects at short delays with the motor measure.

The sizes of the Roelofs effects under various conditions can be compared by measuring the difference between average response with the target on the right and with the target on the left. The cognitive measure shows a large and

consistent deviation, replicating Bridgeman et al. (1997), while the sensorimotor measure shows no deviation.

## Discussion

This experiment showed that the sensorimotor pathway can maintain veridical information about target position, unaffected by visual context, even when perception shows an illusion of position. The rules are different for the two systems. Cognition is conscious and must use context, even when that leads to errors of localization. The sensorimotor system does not use context, and its spatial values are held unconsciously. Conflicting spatial values can exist in the two systems simultaneously.

A possible mechanism of the sensorimotor store is that subjects might fixate the target visually when it is visible, then point where they are looking when the target is extinguished. This would mean a 0-dimensional storage of information of spatial information limited to the location of a single point, held in gaze position rather than in an internal register. If this interpretation is correct, subjects will be unable to perform the motor task if they are prevented from ever fixating the target. In the next experiment, extending the Roelofs effect paradigm, we seek to control for possible attention and fixation effects by preventing our subjects from fixating the target.

## Experiment 2

We hypothesize that if subjects cannot fixate the target, the motor system cannot use spatial information from gaze position and will be forced to call upon the cognitive system for spatial location information. Further, we prevent covert orienting to the target by requiring subjects to perform a continuous oculomotor task throughout the exposure period.

## Method

**Subjects** Seven University of California undergraduates participated in the cognitive condition, and 7 in the motor condition, all right-handed with normal or corrected-to-normal visual acuity. Each subject was run in only one condition, cognitive or motor.

For this experiment we need fixation points that define eye movements, but give the subject no information about target or frame positions. A pair of fixation points is added to the display, in positions statistically uncorrelated with target or frame positions, to elicit horizontal saccades.

**Apparatus** In order to present the target, frame and fixation points simultaneously, and also to improve the accuracy of our jab recordings, we move to an electronic apparatus with all stimuli displayed on a CRT screen. The screen is mounted with its face down and is viewed through a mirror mounted at 45 deg in front of the eyes, so that the display appears to be directly in front of the subject. A touch pad mounted vertically in the apparent plane of the display records jab responses made with a stylus. The frame's width is 24 deg, and the saccade targets are 23 deg apart, displayed above the frame. As before, targets are at 6 deg. left, center, and 6 deg. right.

## Results

Results were analyzed in the same manner as experiment 1. The cognitive subjects showed an effect of target position, frame position and fixation point position, all significant at  $p < 0.0001$ .

The motor subjects, in contrast, showed no Roelofs effect (no significant frame effect), but had a target significant at  $p < 0.0001$  and a fixation point effect significant at  $p < 0.0011$ . There were no significant interactions in either set of results.

## Discussion

Since the subjects in the motor condition showed no Roelofs effect, while those in the cognitive condition did, we can conclude that the sensorimotor representation was controlling the jab for the motor subjects. The representation is at least 2-dimensional, a true map and not a simple matching of gaze and jab positions. The single most important finding of the experiments reported here is that preventing direct fixation on the target, even when multiple targets must be discriminated, does not cause a Roelofs effect. These experiments show that oculomotor fixation and spatially selective attention are not responsible for accurate pointing behavior in an illusory visual context.

## Conclusions

Once again, the evidence can be interpreted in terms of two visual systems, one based on egocentric coordinates to govern motor behavior and another that uses information from visual context to represent spatial relationships in perception. Also, these experiments lend support to the claim that the price in performance the cognitive system must pay in order to take advantage of visual context information is a susceptibility to illusions of spatial context. While it has been shown that direct fixation driven by attentional selection is not the mechanism responsible for accurate pointing behavior in a visual context that creates illusory perceptions in the cognitive system, this shows only that fixation is not responsible. Other aspects of attention may be responsible for the continued accuracy of motor behavior in these experiments.

The visual mechanism by which motor behavior is governed has been shown to be extremely robust, both by these and previous studies. Indeed, the reappearance of a Roelofs effect for motor responses after a delay (Bridgeman et. al., 1997) shows that the cognitive system can provide information to the motor system when necessary, and this so far appears to be the only form of communication between the two systems. To date there is no evidence that the cognitive system can access spatial location information in the motor system, supporting the inference that spatial information can flow in only one direction, from cognitive to motor. In normal visual conditions, however (motor interactions with still-visible targets), spatial information in the two systems remains segregated.

## References

- Aglioti S., DeSouza J. F., & Goodale M. A. (1995) Size-contrast illusions deceive the eye but not the hand. *Current Biology*, 5, 679-685.
- Bridgeman, B., Kirch, M., & Sperling, A. (1981). Segregation of cognitive and motor aspects of visual function using induced motion. *Perception and Psychophysics*, 29, 336-342.
- Bridgeman, B., Lewis, S., Heit, G., & Nagle, M. (1979). Relation between cognitive and motor-oriented systems of visual position perception. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 692-700.
- Bridgeman, B., Peery, S., & Anand, S., (1997). Interaction of cognitive and sensorimotor maps of visual space. *Perception & Psychophysics*, 59, 456-469.
- Daprati, E. & Gentilucci, M. (1997). Grasping an illusion. *Neuropsychologia*, 35, 1577-82.
- Roelofs, C. (1935). Optische Localisation. *Archives für Augenheilkunde*, 109, 395-415.

# Unity of Consciousness: What It Is and Where It Is Found

Andrew Brook

Philosophy and Interdisciplinary Studies  
Carleton University  
Ottawa, ON, Canada K1S 5B6  
abrook@ccs.carleton.ca

## Abstract

The unity of consciousness is our capacity to be conscious of a number of items all at once, in what could be called a single conscious act. Such unity is found in at least three places: consciousness of the world in general, consciousness of self in general, and paying focal attention to aspects of either. In all three, unified consciousness has both a synchronic and a diachronic dimension. That is to say, consciousness is unified both at a given moment and over time. Unified consciousness can be breached in two ways: by splitting (into two unified centres of consciousness, as in brain bisection operations) and by shattering (as in some severe schizophrenias and dysexecutive disorder). Studying it in its breakdown conditions is a good way to throw light on it. In this paper, we will delineate the unity of consciousness, explore some situations in which it breaks down, and relate it to some other mental unities.

**1. Introduction** One of the most striking features of consciousness is that what is presented to us in it is usually highly unified. This unity takes the following general form. We are conscious not just of individual objects but of a multitude of objects related to other objects in a multitude of ways. I am aware not just of A and, separately, of B and, separately, of C, but of A-and-B-and-C, all at the same time – or better, as all parts of a single complex object of a single conscious state. Since at least the time of Kant (1781/7), this unity has been called the *unity of consciousness*.

There has been a huge resurgence of interest in consciousness in cognitive science in the past decade or two. Here is how the philosopher Daniel Dennett summarized the attitude of the nonphilosophical part of the cognitive community two decades ago:

Consciousness appears to be the last bastion of occult properties, epiphenomena, immeasurable subjective states – in short, the one area of mind best left to the philosophers. Let them make fools of themselves trying to corral the quicksilver of “phenomenology” into a respectable theory. [1978, p. 149]

He could have added that this was pretty much true of most philosophers, too.

This situation began to change in the mid to late 1980s, due to the work of psychologist, Bernard Baars (especially 1988) and many others. (Baars developed the methodology called contrastive analysis, in which we compare the difference made by performing a task consciously and without consciousness. This method gave researchers a method to study consciousness much better than the

traditional appeal to introspection.) Consciousness studies quickly became a major player in cognitive research. At least a hundred new books and thousands of articles written from both an experimental and a philosophical point of view have now appeared. Interestingly, even though one of the things that immediately strikes almost everybody about consciousness is its unity, relatively little attention has been paid to it in this burgeoning literature. Neither philosophers nor experimentalists have had much to say about it.

Here we need to make a distinction. Under the name, the binding problem, one phenomenon related to the unity of consciousness has received a lot of attention – our ability (better, the ability of our visual cortex) to ‘bind’ diverse features of objects sensed by diverse parts of the visual or other sensible cortices into representations of three-dimensional objects. Binding of this sort is not unity of consciousness, not as I am discussed the latter. First, the representations that result from binding need not even be conscious. Many perfectly good representations of three dimensional objects affect behaviour and even enter memory without us ever becoming conscious of them. Second, the unity that I am exploring in this paper concerns multiple objects, related to one another in such a way that one is aware of many of them together, not individual objects by themselves. Contrary to the situation with binding, unified consciousness of multiple objects has received little attention.

## 2. Breaches of Unified Consciousness

This lack of attention to the unity of consciousness notwithstanding, some clinical and experimental phenomena in which this unity in fact plays a central role have received a lot of attention, especially situations in which there is some drastic change in unified consciousness.

There are at least two ways in which the unity of consciousness can be breached without unity being destroyed altogether. First, there are the “brain bisection” operations (commissurotomies) much beloved by philosophers, in which it appears that one “centre of consciousness” becomes two under certain conditions (Nagel 1971; Marks 1981). Since the two centres coexist and are both active at the same time, this breach of unity occurs at a single time.

Much ink has been spilled on the question of what is



going on in the phenomenology of these patients. Some theorists have even claimed that there is no whole number of 'centres of consciousness' in these subjects: there is too much unity to say that they are two, yet too much splitting to say that they are one. Some reason work by Sergent (1990) might seem to support this conclusion. She found, for example, that when a sign '6' was sent to one lobe and a sign '7' was sent to the other in these subjects (in such a way that no crossover could occur), they could say that 6 is a smaller number than 7 but could not say whether the signs were the same or different. However, the interpretation of these data is controversial. In particular, there does seem to be a clear answer to any precise 'one or two?' question we could ask, so it is not clear that Nagel's no whole number view receives any support from them. ('Unified consciousness of the two signs with respect to numerical size?' Yes. 'Unified consciousness of the visible structure of the signs?' No).

At any rate, since there continues to be unified consciousness, whether in what are unambiguously two centres or in something less well delineated, we do not have the complete destruction of unity here, though it is a breach of some kind. Then there is the more controversial phenomenon that used to be called Multiple Personality Disorder, now called, more neutrally, Dissociative Identity Disorder. In the most common variety, the units (whatever we want to call them: persons, personalities, sides of a single personality) "take turns" and when one is active, the other(s) usually are not. This is another breach in unity without unity being destroyed, in this case across time.

Then there are phenomenon in which unity does seem to be destroyed. In both brain bisection and dissociative identity cases, we have at most one unified consciousness splitting into two or more – two or more at a time or two or more across time. It is, of course, a matter of debate whether we have even that, especially in the case of dissociative identity disorder, but we clearly do not have more than that. In particular, unity itself does not disappear. The unity may split but it does not shatter. There are at least two kinds of case in which unity does appear to shatter.

One is a certain particularly severe variety of schizophrenia in which the victim seems to lose the ability to form an integrated, interrelated representation of his or her world and his or her self at all. The person speaks in "word salads" that never get anywhere, indeed that sometimes never even reach the level of complete sentences. The person is unable to put together integrated plans of actions even at the level necessary to obtain sustenance or escape irritants. And so on. Here, unity of consciousness appears simply to have shattered.

In schizophrenia of this sort, the shattering of unified consciousness is part of a general breakdown or deformation of mental functioning: affect, desire, belief, even memory all suffer massive distortion. In another kind of case, the normal

unity of consciousness is just as absent but there does not seem to be a general disturbance of the mind. This kind of case has been called dysexecutive syndrome (Dawson 1998, p. 215). What characterizes the breakdown in the unity of consciousness here is that subjects are unable to consider two things together, even things that are clearly related to one another. For example, such people cannot figure out whether a piece of a puzzle fits into a certain place even when the piece obviously fits. They cannot crack an egg into a hot pan. And so on. The reason seems to be that they cannot focus on two items simultaneously and so cannot fit the two together.

The ability to unify the contents of consciousness, as these last examples show, is central to all cognitive functioning, certainly functioning of any complexity. Moreover, the phenomenon once received a lot of attention. For example, it is the centrepiece of Kant's model of the mind (Brook 1994). These facts notwithstanding, the phenomenon has received, as I said, relatively little attention in recent work on consciousness.

### 3. Two Kinds of Consciousness

Before we can draw out the morals for the nature of unified consciousness contained in breaches of unity of the kinds we have just sketched, we first need to say a bit about consciousness in general. In particular, we need to make a crucial distinction.

Current work on consciousness labours under a huge and confusing terminology. Different theorists talk about access consciousness, phenomenal consciousness, self-consciousness, simple consciousness, creature consciousness, state consciousness, monitoring consciousness, awareness taken to be coextensive with consciousness, awareness distinguished from consciousness, higher order thought, higher order experience, qualia, the felt qualities of representations, consciousness as displaced perception, memes, virtual captives .... and on and on and on. A terminology this florid, confused and overlapping is a good sign that consciousness research is still very immature science. For purposes of this article, we need to make just one distinction: between what we will call simple consciousness, on the one hand, and consciousness of self, on the other.

Simple consciousness is closely related to sentience and to being awake. It is (perhaps among other things) being in a certain informationally and behaviourally responsive state to one's immediate environment. It is the ability, for example, to process and act responsively to information about food, friends, foes, and other items of relevance. One finds simple consciousness a long way down the evolutionary ladder.

Consciousness of self is the ability to process and respond in a similar fashion to oneself, more specifically, to one's own psychological states and to oneself as oneself, as the

thing whose states they are. The latter form of consciousness of self, the ability to identify oneself as oneself, probably requires the use of indexicals and may therefore be restricted to human beings and perhaps a few other species of primate.

The importance of this distinction between simple consciousness and consciousness of self is that the literature tends not to distinguish them and even to run them together. Everyday English does so, too. We speak of someone regaining consciousness – where we mean simple consciousness of the world. Yet we also say things like, “She wasn’t conscious of what motivated her to say that” – where we do not mean that she lacked simple consciousness of the world but rather that she was not conscious of something about herself. Some theorists make this distinction but others treat consciousness as either synonymous or at least coextensive with consciousness of the second sort, what we are calling consciousness of self. A few even occupy a middle ground, those philosophers who talk about the felt qualities of things as central to consciousness, for example. They do not seem to hold that we must be *conscious of* these felt qualities for them to exist as conscious states – but they do not view them as objects of simple consciousness of the world either.<sup>1</sup> To understand the unity of consciousness, we need to make the distinction. We need to treat consciousness of self and simple consciousness of the world as distinct. Why? Because even though the distinctive unity associated with consciousness is found in both, it takes somewhat different forms.

#### 4. Unity of Consciousness

Indeed, we find unity of consciousness in at least three places. We might call them *unity of simple consciousness*, *unified consciousness of self*, and *unity of focus*.

Unity of consciousness in general starts from the intuitive idea laid out above that we are aware of a great many things at once. Here is a more informative definition:

The unity of consciousness =*df.* a consciousness of objects in which a number of representations of objects and sometimes also the representation themselves are combined in such a way that to be conscious of any of these objects and/or representations of them is also to be conscious of other objects and/or representations as connected to it/them and of the group together as a single complex whole of objects and/or representations.

*i. Unity of simple consciousness* Unity of simple consciousness is the consciousness that we have of the world around us (including, it should be noted, one’s own body and perhaps even psychological states) as a single world, of the various items in it as linked to other items in it. That is to say, it is simply unity of consciousness as found in the conscious representation of one’s environment.

*ii. Unified consciousness of self* Here one is aware of

oneself as not just the subject but, as Kant put it (A350), the “single common subject” of unified fields of representation (and the single common agent of unified activities of deliberation and action). Unified consciousness of self has been argued to have some very special properties, in particular that the reference to oneself as oneself that generates it is achieved without “identification” – that is to say, not via attribution of identifying properties or attributes to oneself (Castañeda 1966; Shoemaker 1968; Perry 1979) but we do not have room to go into that interesting issue here.

*iii. Unity of focus* Unity of focus refers to our ability to pay unified attention to objects and one’s own self. It may be part of unified consciousness in general. Whether it is or not, it is certainly not the same thing. In the two situations of unified consciousness just explored, consciousness ranges over many objects (or, in the case of unified consciousness of self, many occurrences of becoming aware of an object). Unity of focus is a matter of focussing on one such item. What I have in mind is Wundt’s old distinction between the field of consciousness (*Blickfeld*) and the focus of consciousness (*Blickpunkt*). The consciousness of an item on which one is focussing is just as unified as the consciousness of many such items at the same time. If so, we find an occurrence of unified consciousness *within* each of the two sites of unified consciousness laid out in (i) and (ii). We are talking, of course, about focal attention.

Note that, in addition to paying focal attention to individual objects, we can also unite a number of considerations in focal attention at the same time – desires, beliefs, alternatives, probabilities, and so on – and integrate them with, for example, available alternatives to reach decisions and choose courses of action. We can then go on to do the same with behaviour and resources, focussing on carrying out the choice in the face of obstacles, conflicting desires, and so forth. Moreover, there are costs attached to not having fully functioning focal attention, as the dysexecutive syndrome mentioned above makes painfully apparent. These remarks suggest that unified consciousness is not the only form of mental unity, a suggestion to which we will return briefly below.

Though this has often been overlooked, the unity found in unified consciousness comes in two very different forms, no matter which site we have in mind. The unity can consist entirely in phenomena occurring at the same time and it can consist in links of certain kinds among phenomena occurring at different times. In its synchronic form, it consists in such things as our ability to compare two items to one another, to see how two items fit or do not fit into one another, etc. Diachronically, it consists in the ability to retain a representation of an earlier object in the right way and for long enough to relate the earlier object to some currently represented object.

## 5. The Situations in which Unity is Breached

Let us now return to the four breaches of unified consciousness discussed earlier. We can see that in every case, at least one feature of unified consciousness as we defined it is absent.

In brain bisection cases, there are, notoriously, all sorts of situations in which a being in the body in question who is aware of some represented objects is not aware of others. Thus, for example, if the right hemisphere is asked to do arithmetic in a way that does not penetrate to the left hemisphere and the hands are shielded from the eyes, it is easy to set up a situation in which the left hand will be doing arithmetic while whatever controls the mouth insists that it is not doing arithmetic, indeed has not even thought of arithmetic today. And so on.

In DID cases, a central feature of the case is reciprocal amnesia (with all sorts of variations). Again, this is a situation in which a being aware of some represented objects is not aware of others.

The same pattern is even more clear in the cases of severe schizophrenia and dysexecutive disorder sketched. In both cases, awareness of some conscious states goes with lack of awareness of others. There is nothing aware of all the relevant conscious states together.

In short, our definition seems to illuminate the situations in which unity of consciousness is breached quite nicely.

## 6. Other Unities in Cognition

The unity of consciousness is far from being the only kind of mental unity as our remarks about what can be integrated in focal attention might indicate. There is unity in the early stages of cognition, unity that consists of integration of motivating factors, cognitive capacities, etc., and also unity in the outputs, unity that consists of integration of behaviour. First, the early stages of cognition.

One of the more striking things about human beings as cognitive systems is that we can bring an extremely wide range of factors to bear on a cognitive task, e.g., when we seek to characterize something or reach a decision about what to do about something. We can bring to bear: what we want; what we believe; our attitudes to self, situation, and context; input from each of our various senses; information about the situation, other people, others' beliefs, desires, attitudes, etc.; the resources of however many languages we have available to us; the various kinds of memory; bodily sensations; various problem-solving skills that we have acquired; and so on. Not only can we bring all these elements to bear, we can integrate them in a way that is highly structured and ingeniously appropriate to our goals and the situation(s) before us. This form of mental unity could appropriately be called *cognitive unity*.

At the other end of the cognitive process, we find an equally interesting form of unity, what we might call *unity of*

*behaviour*. To act, we need to coordinate our limbs, eyes, bodily attitude, etc., indeed in ways the precision and complexity of which would be difficult to exaggerate. Think of a concert pianist performing a complicated work.

And between the two is the unified consciousness laid out in the previous section.

## 7. The Unity of Consciousness as Evidence

It would seem that anything as central to human cognition as unified consciousness would have to play a role in any serious attempt to understand cognition. This, of course, has not been the case for a while. As has often been remarked, until about fifteen years ago, as cognition was modelled in cognitive science, it could just as well have been entirely nonconscious.

Historically, the unity of consciousness played a large role. Indeed, it is central to one of the most famous arguments in philosophy, Kant's "deduction" of the categories. In this argument, boiled down to its bare essentials, Kant argued that in order to tie various items together into a single unified conscious representation, we must be able to apply certain concepts to the items in question, in particular qualitative, quantitative, relational and what he called 'modal' concepts. (Modal concepts are the concepts we use when we decide whether something merely could exist, actually does exist, or [if this is ever the case] must exist.) By far the most important relational concept for Kant was the concept of cause and effect. Indeed, Kant thought that he could tease a complete defence of physics as a body of genuine knowledge out of the fact (as he saw it) that we have to be able to apply the concept of cause and effect to items in our experience if we are to have a unified consciousness of them.

It also played a role in arguments for dualism. Theorists otherwise as different as Descartes and Reid argued that unified consciousness could never be achieved by any system of components acting in concert. Give each of these components a part of a thought or perception divided up as finely as you please; the result will never be a unified thought or perception. As James famously put it,

Take a sentence of a dozen words, take twelve men, and to each one word. Then stand the men in a row or jam them in a bunch, and let each think of his word as intently as he will; nowhere will there be a consciousness of the whole sentence. [James, 1890, Vol. 1, p. 160]

The inference from this argument was that the human mind could not be any system of components. Now, anything material will be a system of components. If so, then the mind is not made out of matter.

Remarkably enough, some version of this argument impressed practically all theorists until well into the twentieth century, despite the complete absence of anything

like an alternative account and even though no less a figure than Kant poked a huge hole in it as early as 1781. (He noted that unified consciousness being achieved by a system of components acting together would be no more mysterious than it be achieved by something that has no parts or components.)

Nonetheless and whatever the merits of this argument for the simplicity and immateriality of the mind, the unity of consciousness did receive a lot of attention. And rightly so; cognition of any complexity must be unified in the way that consciousness is. Without the ability to retain representations of earlier objects and unite them with current represented objects, for example, the only language that we would be able to understand would be single words. The simplest of sentences is something spread over time. Now, unification *in consciousness* might not be the only way of achieving this unity but it is clearly a central way. If so, consciousness being unified is central to cognitive life as we know it.

In some circles, the idea that consciousness has a special kind of unity has fallen into disfavour lately. Davidson, Fodor, Dennett, Pylyshyn and the Churchlands come immediately to mind. The mind, they say, is modular (Fodor 1983) and most modules work out of the sight and control of consciousness. Moreover, we often do things that we don't intend, act for reasons of which we are not aware, and so on. Does any of this entail that consciousness is not unified? Not at all. The most these observations do is to shrink the range over which the unity extends. If something is out of the sight and/or the control of the conscious mind, we should ask: out of the sight or control of what? Unified consciousness. And we still need to understand the nature of this unity. Practically anything that could be said about the unity of consciousness when consciousness was conceived in the pre-twentieth century way as ranging over most everything mental can still be said about the unity of consciousness conceived in the twentieth-century way with a range that has shrunk dramatically.

Yet few recent philosophers and even fewer other cognitive researchers even raise the question of what the unity of consciousness is like. This is strange; it hardly seems controversial to say that we have unified consciousness, though how far this unity extends and over what can be debated. Indeed, without knowing what the unity of consciousness is, it is hard to see how we can even talk coherently about the situations so prominent at the moment where unity is absent or breached.

## 8. Background: Theories of Consciousness

We will close with a different question: Does the unity of consciousness have implications for the big debates about the general nature of consciousness currently raging? There are currently at least three camps. There are those who see consciousness as something quite unique, the “felt quality”

of representations or whatever. On this picture, representations could function much as they do even if, in Nagel's (1974) phrase, it was not like anything to have them. They would merely not be conscious. If such a split is possible, then the next question is whether consciousness plays any important cognitive role at all, its unity included. Maybe it is a free rider (Jackson 1986; Chalmers 1996).

Then there is a second camp. It holds, to the contrary, that consciousness is simply a special kind of representation: a representation of a representation, for example (Rosenthal 1991; Dretske 1995; Tye 1995).

Finally, there are those who hold that what we call consciousness is really something else. On this view, consciousness will in the end be “analysed away” – what we misleadingly label ‘consciousness’ is something very different from what we take consciousness to be like. Perhaps it is competing information-parcels in a Pandemonium architecture that have gained temporary dominance in the struggle for cognitive resources (Dennett 1991). Perhaps it is self-monitoring transformations of some sort in a multidimensional phase-space (Churchland 1995). Whatever, consciousness is not anything like the unified system of representations that both common sense and the Kantian model of the mind take it to be like.

No matter what one's view of the nature of consciousness, and the three views sketched above probably do not exhaust the possibilities, one will have to provide an account of the unity found in it. Indeed, even if one holds that this unity has been overrated and consciousness is much less unified than theorists have thought, one will still have to provide an account of this unity in those situations in which it does occur. The kind of integration of properties and objects into more complex objects of experience that we sketched above is too central to be ignored.

On the other hand, the unity of consciousness as we have defined it might not have much by way of implications for which of the three views is right. If it is as genuine and undeniable as I've urged, it may cut a bit against the third, eliminativist position. But adherents of this position have increasingly been treating consciousness as something real, i.e., nothing to be eliminated, in any case. The unity of consciousness seems neutral with respect to the other two positions. If so, curiously enough, which view of consciousness we start from may not matter much when we set out to understand the unity of consciousness.

## References

- Baars, B. 1988. *A Cognitive Theory of Consciousness* Cambridge University Press
- Brook, A. 1994. *Kant and the Mind*. Cambridge University Press.
- Kristjánsson, Hector-N. 1966. ‘He’: A study in the logic of self-consciousness. *Ratio* 8, pp.130-57.
- Chalmers, D. 1996. *The Conscious Mind* Oxford: Oxford

University Press

- Churchland, P. M. 1995. *The Engine of Reason, the Seat of the Soul* MIT Press
- Dawson, M. 1998. *Understanding Cognitive Science*. Blackwell's
- Dennett, D. 1978. Toward a cognitive theory of consciousness *Brainstorms* Bradford Books, pp. 149-73
- Dennett, D. 1991. *Consciousness Explained*. Little, Brown
- Dretske, F. 1995. *Naturalizing the Mind*. MIT Press
- Fodor, J. 1983. *Modularity of Mind*. MIT Press
- Jackson, F. 1986. What Mary didn't know *J. Phil.* 83:5, 291-5
- James, W. 1890. *Principles of Psychology*, two volumes. London: Macmillan
- Kant, I. 1781/87. *Critique of Pure Reason*. Trans N. Kemp Smith. Macmillan 1927
- Marks, C. 1981. *Commissurotomy, Consciousness and Unity of Mind*. MIT Press
- Nagel, T. 1965. Physicalism. *Phil. Rev.* 74, 339-56
- Nagel, T. 1971. Brain bisection and the unity of consciousness. *Synthese* 22, pp. 396-413.
- Nagel, T. 1974. What it is like to be a bat. *Phil. Rev.* 83, pp. 435-50
- Perry, J. . 1979. The problem of the essential indexical. *Noûs* 13, pp.3-21.
- Sergent, J. 1990. Furtive incursions into bicameral minds. *Brain* 113: 537-68.
- Rosenthal, D. 1991. *The Nature of Mind*. Oxford University Press
- Shoemaker, Sydney. 1968. Self-reference and self-awareness. *Journal of Philosophy* 65, pp. 555-67.
- Tye, M. 1995. *Ten Problems of Consciousness*. Cambridge, MA: MIT Press

---

<sup>1</sup> A full set of relevant distinctions here would distinguish among consciousness of self, consciousness of one's psychological states, one's conscious states themselves, and so on. We do not need to go into the differences among these things here. For purposes of understanding the unity of consciousness, it is enough to distinguish between consciousness of self, on the one hand, and simple consciousness of the "world", on the other.

# Are Retrievals from Long-Term Memory Interruptible?

Michael D. Byrne  
byrne@acm.org  
Department of Psychology  
Rice University  
Houston, TX 77251

## Abstract

Many simple performance parameters about human memory are not well-understood. One such parameter is how the cognitive system handles interruption at a relatively low level. This research is an attempt to determine if simple, well-practiced retrievals from long-term memory can be interrupted by a higher-priority task. An experimental paradigm referred to as a “reverse PRP” paradigm is introduced, and the results of one experiment in this paradigm reported. The results suggest that retrievals can indeed be interrupted, but that there is an interruption cost.

## Introduction

There are numerous situations in which people are interrupted in doing simple tasks by higher priority tasks and must drop what they are working on the new task. In most situations, this is merely an inconvenience. However, in high-performance tasks such as air traffic control, even a small delay in responding to the interrupting task can have more serious consequences. In many cases, the interruption may place demands on perceptual or motor performance, but in other cases it is a cognitive operation that is interrupted. Generally speaking, cognitive theories have little to say about what should happen in such situations. However, this does not mean that these phenomena cannot be understood in the context of, and do not have implications for, theories of cognition.

ACT-R/PM (Byrne & Anderson, 1998) provides a set of perceptual-motor extensions to the ACT-R cognitive architecture (Anderson & Lebiere, 1998). Communication between central cognition (the ACT-R production system) and the perceptual-motor modules takes two forms: [1] the left-hand, or THEN, side of productions can request activity from the perceptual-motor modules (e.g. shift visual attention, press a key), and [2] perceptual-motor modules deliver results (e.g. representations of percepts) to ACT-R’s declarative memory in the form of chunks.

Declarative memory chunks in ACT-R are accessed via retrieval, which is a time-consuming process. That is, retrievals take time, which is part of the process of matching the IF side of productions in ACT-R. Because perceptual-motor modules operate in parallel with the production system, it is possible for one or more of the perceptual-motor modules to change the contents of declarative memory while a retrieval is in progress. The fundamental question this research is attempting to address is what happens in this situation: Do retrievals always complete or can they be

interrupted? Rather than attempt to answer this question on theoretical or computational grounds, this research approaches this as an empirical question.

## Reverse PRP Paradigm

Consider this simple dual task: two digits appear on a display, and the product of those digit should be spoken aloud. On some trials, the digits are replaced a short time after they appear by a colored block. When the block appears, the task is to make a choice response based on the color of the block as rapidly as possible. Because the delay is short, the appearance of the color block may be interrupting the retrieval of the product of the two digits. Can the single, simple retrieval be interrupted?

This task shares a number of important properties with the psychological refractory period (PRP) paradigm, which is perhaps the simplest dual-task experimental paradigm. The PRP has a long history in psychology (see Pashler, 1994 for a review). In this paradigm, subjects are asked to do two tasks, usually referred to as T1 and T2, in rapid succession. The stimulus for T1 appears, then after some delay (called the stimulus onset asynchrony or SOA), the stimulus for T2 appears. Subjects are instructed to give T1 maximum priority and the typical results are that responses to T2 are slowed,

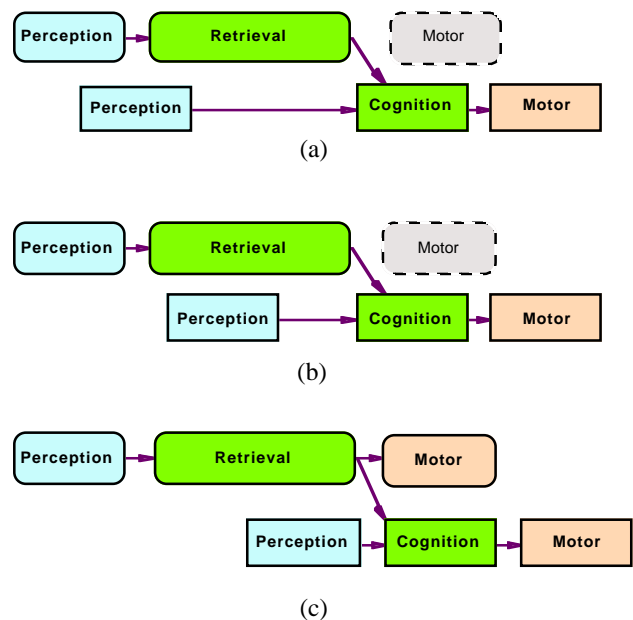


Figure 1. Predictions of the ballistic retrieval hypothesis

and more so at shorter SOAs. Results of such experiments have been taken as evidence that central cognition is effectively serial (again, see Pashler 1994 for a review).

The basic experimental paradigm used in this research inverts the priority instruction given to the subjects. That is, subjects are instructed to give T2 maximum priority; when the T2 stimulus onsets, subjects are to immediately give that stimulus highest priority. If T1 involves retrieval from declarative memory, the interruptibility of that retrieval will have a large impact on response time for T2. If the T1 retrieval is not interruptible (this will be termed “ballistic”), then, assuming serial cognition, cognitive processing of the T2 stimulus will be forced to wait for the completion of the retrieval and will thus be slowed. In particular, it should be slowed more at shorter SOAs. This situation is depicted in Figure 1. In Figure 1 and the following figures, time moves from left to right, and each stage of processing is represented by a box. Arrows represent dependencies. T1 stages are the upper set of boxes, T2 stages the lower set.

Panel (a) of Figure 1 shows the situation at short SOAs, which will result in a long T2 response time. Cognition for T2 must wait for the T1 retrieval to complete, which causes an elevated T2 response time. As SOA increases, T2 response time should decrease (Figure 1, panel b) until at long enough SOAs T2 should no longer be slowed at all (Figure 1, panel c). The slope of T2 response time as a function of SOA should thus be -1 until the “long enough” SOA is reached and the slope drops to zero. At this point, the response time for T2 should be the same as when T2 is not an interrupting task, that is, the single task time.

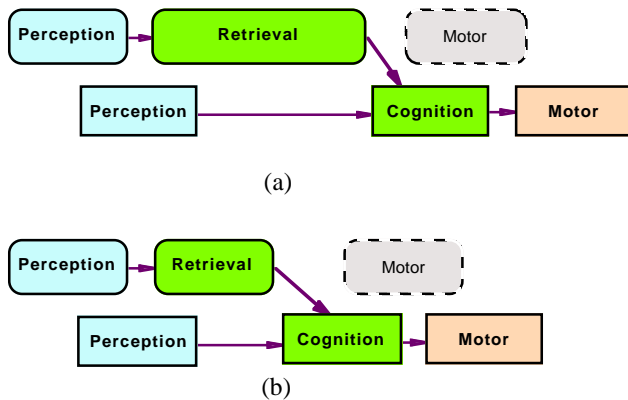


Figure 2. Difficulty effect under the ballistic retrieval hypothesis

A secondary prediction made by the ballistic retrieval hypothesis is that the duration of the T1 retrieval should directly impact the T2 response time. If processing for T2 must wait for the completion of T1 retrieval, extending the duration of that retrieval (e.g. by making the retrieval more difficult) should directly impact T2 response time. If processing for T2 must wait for the completion of the T1 retrieval, extending the duration of that retrieval should

result in a time cost for T2 identical in size to the increase in retrieval difficulty. This is depicted in Figure 2: panel (a) depicts a short T1 retrieval, panel (b) depicts a long T1 retrieval.

If, on the other hand, retrievals are interruptible, T2 response should be insensitive to the state of the T1 retrieval. That is, there should be no effect of either SOA or T1 difficulty. This situation is depicted in Figure 3.

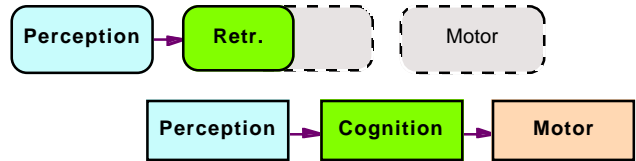


Figure 3. Interruptible retrieval hypothesis

There is a potential complication, which is interruption cost. The shift from T1 to T2 may have a cognitive cost. If such a cost exists, and it is fixed, then the T2 response time in the interruption situation should be elevated when compared to the T2 response time when T2 is performed in isolation (the single-task case). This should hold regardless of T1 difficulty or SOA. Figure 4 represents the situation in which retrievals are interruptible but with an interruption cost.

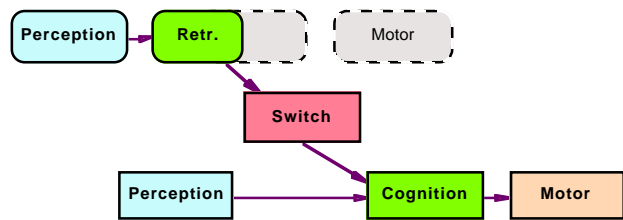


Figure 4. Interruptible retrieval with switch cost

To summarize, the ballistic retrieval hypothesis predicts that T2 response time should have a -1 slope as a function of SOA, and that adding difficulty to retrieval should generate a parallel function of SOA, with the distance between the RT functions equivalent to the single-task difficulty effect of T1. On the other hand, interruptible retrieval hypothesis predicts that T2 should be insensitive to either SOA or T1 retrieval difficulty.

## Methods

### Participants

The participants were 39 Rice University undergraduates who participated for credit in a psychology class.

## Stimuli and Design

There were three kinds of trials: multiplication only, color identification, and interruption. Single-digit multiplication was used as T1 in this paradigm. Participants saw two digits presented visually (e.g. “6 8”) and responded with the product of the two digits vocally (e.g. “forty-eight”). Retrieval difficulty was manipulated by varying the size of the digits used. This manipulation has been shown to be effective in previous work (Byrne & Anderson, 1999). “Easy” retrieval used the digits from 1 to 4, while “hard” retrieval used the digits 6 through 9. Squares (e.g. “7 7”) were not used.

A simple color identification task served as T2 in this paradigm. This was a choice reaction time task with two alternatives. A rectangular block of color appeared on the display. If the color block was blue, participants pressed one key on the keyboard; if the block was red, another key was pressed.

For interruption trials, the color block appeared and covered the digits on the screen. The SOA was the time between the onset of the digits and the onset of the color block, measured in milliseconds. SOAs of 200, 375, 550, and 725 ms were used. Participants were instructed that when an interruption occurred, they were to respond to T2 as rapidly as possible and that completion of T1 was not necessary. These instructions were given to maximize the priority given to T2; participants should have no reason to continue with T1 and thus should switch to T2 as rapidly as possible.

The design was also blocked, each block consisted of five sets of 40 trials. One set in each block consisted of only color identification trials, to provide an estimate of single-task response time. The remaining four sets were a mixture of multiplication-only trials and interruption trials, with interruptions occurring 20% of the time. Thus, for interruption trials, there were three factors, all within-subjects: block, from one to three, four levels of SOA, and two levels of difficulty. Which trials contained interruptions and the order of sets within a block were randomized.

## Procedures

Participants were first trained on the color identification task until they performed two consecutive sets of 40 trials with 95% or better accuracy. Participants were then given 40 trials of practice with multiplication-only trials, followed by a 40-trial set of multiplication trials, 20% of which contained interruptions.

## Apparatus

Stimulus presentation and data collection were done on Apple iMac personal computers. Vocal responses were timed with an Apple PlainTalk microphone by monitoring the microphone level and stopping the timer when a threshold level of input was exceeded. Keypress responses were timed by actively polling the state of the keyboard. Both measures should be accurate to approximately 5 ms.

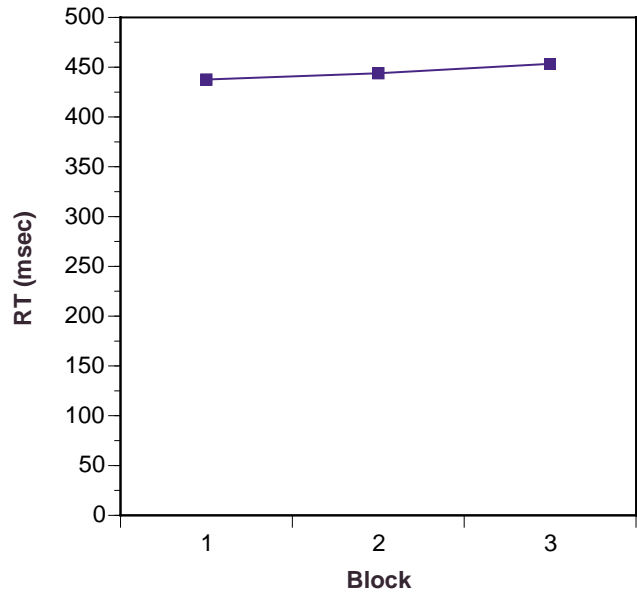


Figure 5. Color identification response time as a function of block

## Results

Due to the excellent power of the repeated-measures design and the large number of subjects and trials, an alpha level of 0.01 will be used for all statistical tests.

The color identification task is fairly simple and participants were forced to practice to a relatively stringent criterion, so performance was expected to be rapid but there was still the possibility that subjects may have been speeding up with practice. Figure 5 presents single-task color identification response time as a function of block. Clearly, there was no practice-related speedup in this case, in fact, the absolute response times actually went down slightly with practice, though this is probably coincidental. Overall, the effect of block was not reliable,  $F(2, 70) = 1.83, p = 0.17$ . The lack of learning on this task suggests that performance on this task is limited primarily by fixed architectural properties such as perceptual-motor limitations; the cognitive demands of this task are fairly minimal.

Multiplication-only trials demonstrated a much more complex pattern. Response time for multiplication-only trials is shown in Figure 6. As expected, there was an effect of difficulty,  $F(1, 35) = 81.74, p < 0.001$  with hard problems clearly slower than easy ones, on average, about 350 ms slower. There was also a main effect of block,  $F(2, 70) = 10.30, p < 0.001$ ,<sup>1</sup> and a block by difficulty interaction,  $F(2, 70) = 12.14, p < 0.001$ , both primarily a function of improvement on hard problems. If retrievals are ballistic, all of these effects should show up in T2 response time in the interruption trials, since T2 cognition should be forced to wait for the completion of the retrieval.

<sup>1</sup> To control for sphericity problems, repeated-measures factors with more than two levels were adjusted with either Huynh-Feldt epsilon or Greenhouse-Geisser epsilon where appropriate.



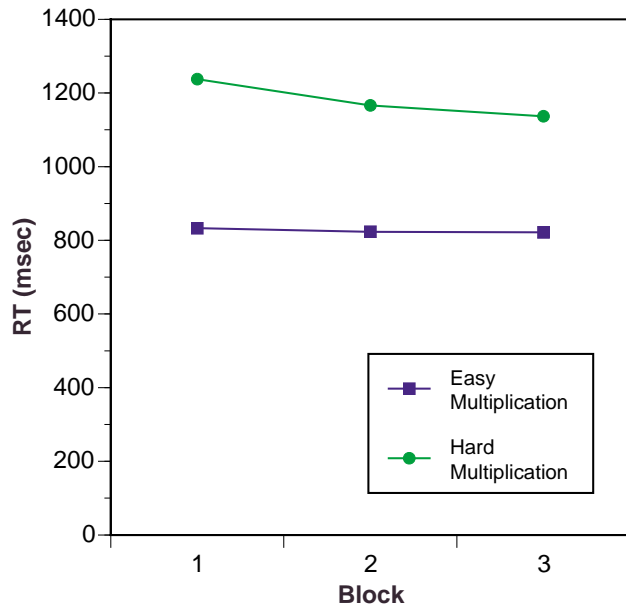


Figure 6. Multiplication-only response time for easy and hard problems as a function of block

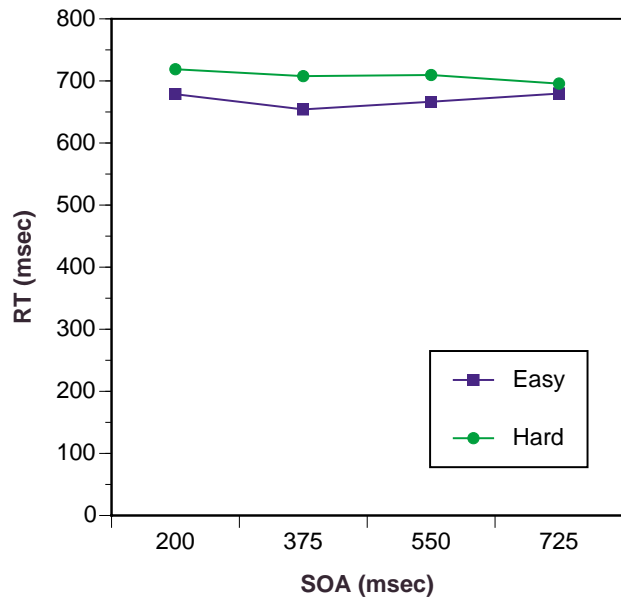


Figure 7. T2 interruption response time as a function of SOA for hard and easy T1 retrievals

The interruptible retrieval hypothesis, given the lack of practice effects on color identification in isolation, should predict no effect of difficulty or block on T2 interruption performance.

The data of primary interest, of course, are the data for the interruption trials. These data, as a function of SOA, are presented in Figure 7. The results are generally consistent with the interruptibility hypothesis. Most importantly, there was no effect of SOA,  $F(3, 105) = 0.90, p = 0.40$ . There is clearly a potential problem of accepting the null hypothesis here. However, the prediction made by the ballistic hypothesis is specific: there should be a -1 slope with SOA. This can be tested with a linear contrast on SOA, which was not reliable,  $t(35) = -0.51, p = 0.61$ . A -1 slope would be a large effect in this context, and power to detect a large effect in this situation was estimated to be 0.99 (see Cohen, 1988 for details on this procedure). Thus, accepting the null hypothesis in this case is statistically justifiable.

All other effects and interactions were also not reliable, save one: the effect of T1 difficulty (the difference between the easy and hard conditions) was reliable,  $F(1, 35) = 8.29, p = 0.007$ . The absolute magnitude of this difference is small, however, at just under 40 ms. The two difficulty effects, one in multiplication-only trials, and one in interruption trials, is presented for each block in Figure 8. These effects are obviously different, and indeed a repeated-measures ANOVA on the difficulty reveals a very reliable effect of multiplication-only vs. interruption,  $F(1, 35) = 65.46, p < 0.001$ . This suggests that while the difficulty effect did manifest itself in the T2 response time, this effect is probably not due to retrieval difficulty in T1, since that difficulty effect was roughly nine times larger.

There was also a reliable effect of block,  $F(2, 70) = 6.73, p = 0.004$ , and an interaction,  $F(2, 70) = 10.68, p < 0.001$  on the difficulty effect. This seems to be driven primarily by the previously-mentioned improvement in “hard” multiplication problems over time, which results in a reduction in difficulty effect for the multiplication-only trials; in contrast, the small

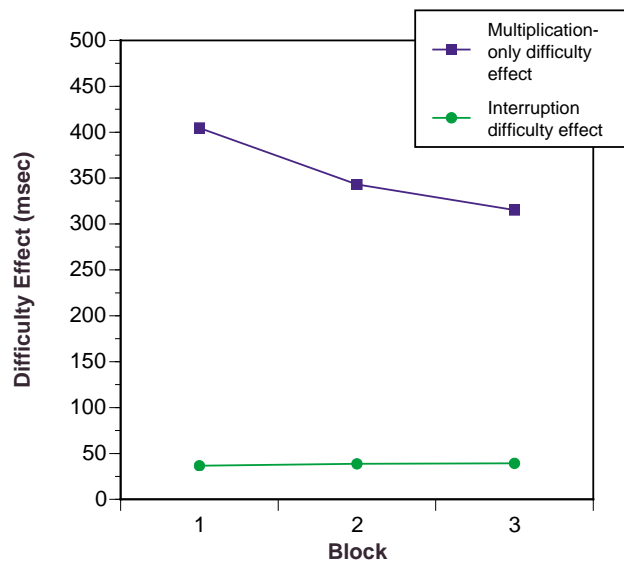


Figure 8. Difficulty effect as a function of block for multiplication-only trials and interruption trials

difficulty effect seen in T2 interruptions is fairly stable over blocks.

Of course, the lack of SOA effect may be due to subjects adopting a strategy of delaying response to T1 until they could be confident that an interruption would not occur. The multiplication-only response times are fairly rapid, suggesting this is unlikely, but there is a more direct test. Subjects often responded to T1 even when the interruption occurred, but they did so more often for long SOAs than for short SOAs and more often for easy problems than hard problems. This is shown in Figure 9. Effects of block, SOA, and their interaction were reliable, [for SOA,  $F(3, 105) = 108.70, p < 0.001$ ; for difficulty,  $F(1, 35) = 152.24, p < 0.001$ ; for the interaction,  $F(3, 105) = 11.42, p < 0.001$ ] but there were no reliable effects or interactions involving block. This sensitivity to SOA and difficulty suggests that participants did indeed attempt to respond as rapidly as possible to T1 and did not uniformly postpone T1 in anticipation of an interruption.

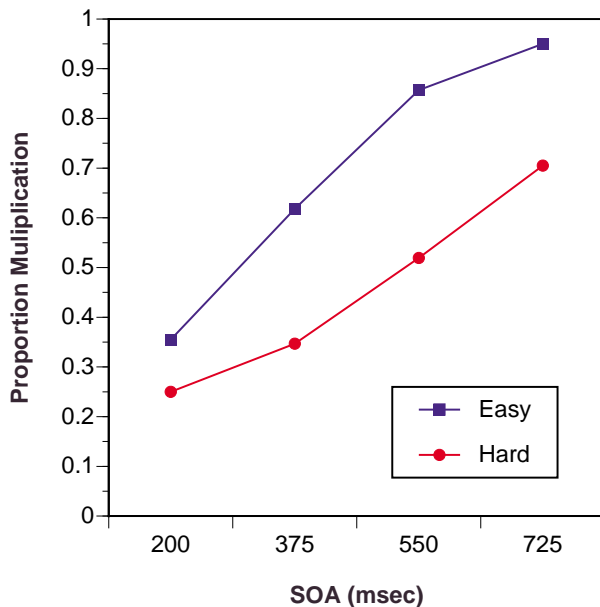


Figure 9. Proportion of interruption trials on which a T1 response was emitted

The final effect to consider is the interruption cost, that is, the difference between color identification response time when it was in isolation vs. when it was the interrupting task. Figure 10 presents the results. Clearly, there was an interruption cost,  $F(1, 35) = 235.58, p < 0.001$ . The absolute magnitude of this difference is large relative to the single-task color identification response time. Single-task response time for color identification averaged just under 450 ms, but with interruptions it was close to 700 ms, a 250 ms penalty.

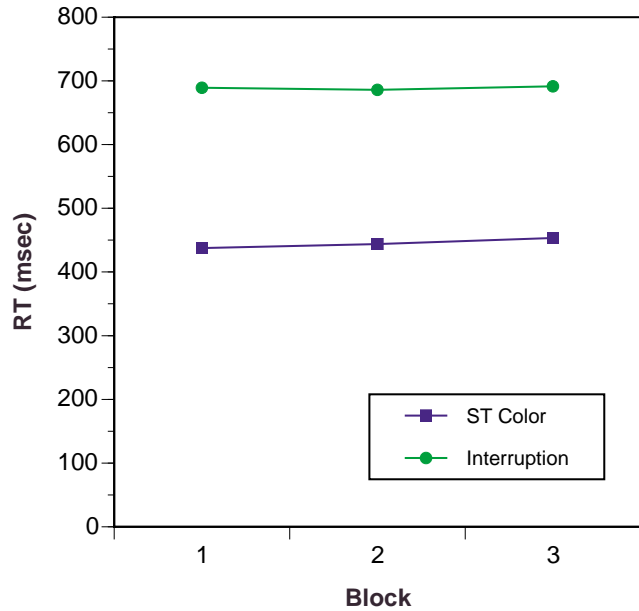


Figure 10. Color identification response time as a function of block when it was in isolation (single-task or ST) vs. as the interrupting T2

There is no real evidence that this cost was reduced with practice as there was no reliable effect of block,  $F(2, 70) = 0.80, p = 0.45$ , or an interaction of task condition and block,  $F(2, 70) = 0.81, p = 0.45$ .

## Discussion

These data are clearly more consistent with the interruptibility hypothesis. The lack of SOA effect on T2 response time is most telling. However, the data are not entirely equivocal. There was a reliable effect of T1 retrieval difficulty on T2 response time, though this effect was small and clearly of a different magnitude than the difficulty effect present in T1. The source of this effect is unclear. One possible explanation is that perceptual processing of T1 is more difficult for larger digits but this is purely speculative.

For the purposes of setting architectural policy in ACT-R/PM, these results certainly suggest that retrievals should be interruptible. However, whether retrievals should always be interrupted by any change in declarative memory or whether they should only be interrupted under certain conditions is unclear. In this experiment, the retrieval is interrupted by a higher-priority change that is both presented foveally and displaces the T1 stimulus in the visual array. These conditions at least appear to favor interruption. The frequency of interruption in this experiment, 20%, may also play a role.

At a more general level, the interruption cost itself is quite intriguing. The source of this cost is not clear, though something of its nature was revealed; it appears not to change with practice (blocks) and appears not to be affected

by SOA. Whether this cost is sensitive to factors such as interruption frequency, modality match with the T1 stimulus, and T2 difficulty, is unknown. Follow-up research certainly appears appropriate.

However, in some sense, the change from T2 to T1 processing can be thought of as a task-switch (e.g. Rogers & Monsell, 1995). While a great deal is known about task-switching (Altmann & Gray, 1999 provides an excellent account), it is not clear whether or not this is a special case of task-switching phenomenon. In traditional task-switching experiments, one type of task follows the completion of another, but the two tasks do not temporally overlap, that is, one does not interrupt the other. The ramifications of this difference in experimental paradigm are not entirely clear; the interruption cost may be related to the cost associated with task-switching or it may be an independent effect. Again, further research will be required to better understand the interruption cost.

### **Acknowledgements**

I would like to thank Michael Fleetwood and Bryan Blauvelt for their assistance in data collection, and to Erik Altmann for comments on an earlier draft.

### **References**

- Altmann, E. M., & Gray, W. D. (1999). The anatomy of serial attention: An integrated model of set shifting and maintenance. Manuscript submitted for publication.
- Anderson, J. R., & Lebiere, C. (Eds.). (1998). *The atomic components of thought*. Hillsdale, NJ: Erlbaum.
- Byrne, M. D. & Anderson, J. R. (1998). Perception and Action. In J. R. Anderson & C. Lebiere (Eds.) *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum.
- Byrne, M. D., & Anderson, J. R. (1999). Serial modules in parallel: The psychological refractory period and perfect time-sharing. Manuscript submitted for publication.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, 116, 220-244.
- Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, 124(2), 207-231.

# Hemispheric Specialization During Episodic Memory Encoding in the Human Hippocampus and MTL

Daniel J. Casasanto† ([dcasasan@mail.med.upenn.edu](mailto:dcasasan@mail.med.upenn.edu))  
William D. S. Killgore† ([killgore@mclean.harvard.edu](mailto:killgore@mclean.harvard.edu))  
Guila Glosser† ([glosser@mail.med.upenn.edu](mailto:glosser@mail.med.upenn.edu))  
Joseph A. Maldjian‡ ([maldjian@oasis.rad.upenn.edu](mailto:maldjian@oasis.rad.upenn.edu))  
John A. Detre†‡ ([detre@mail.med.upenn.edu](mailto:detre@mail.med.upenn.edu))

Departments of Neurology and Radiology, University of Pennsylvania Medical Center  
3400 Spruce Street, Philadelphia, PA 19104 USA

## Abstract

Hemispheric specialization during episodic memory encoding was examined using three functional magnetic resonance imaging (fMRI) tasks. Stimuli for the three tasks differed in the degree to which they elicited subjects' use of verbal and image-based encoding strategies. Intentional encoding of visually presented scenes, sentences, and faces was associated with neural activity in the hippocampus and surrounding mesial Temporal Lobe (mTL) structures. Across tasks, material-specific lateralization of neural activity was observed in the posterior mTL. In contrast, hippocampal activation did not lateralize according to material type for two of the three tasks. These results suggest a functional dissociation between the hippocampus and other mTL subcomponents, and indicate that material-specificity may not fully explain hemispheric specialization in the mTL memory system.

## Introduction

The human hippocampus and adjacent mesial temporal lobe (mTL) structures are believed to subservise encoding of new information into episodic memory: the form of long-term memory that supports conscious recollection of ongoing experiences (Squire and Zola-Morgan, 1991; Tulving, 1998). The role of the mTL in long-term memory processing has been investigated extensively ever since Scoville and Milner (1957) reported profound global anterograde amnesia in patient HM following bilateral resection of the hippocampus, uncus, and amygdala. Numerous studies of unilateral mTL resection have documented that in left-language dominant patients, resection of the left anterior temporal lobe consistently produces verbal memory impairment, and although the findings are less robust, that resection of the right, non-language-dominant anterior temporal lobe produces visuospatial memory impairment (Milner, 1958; Blakemore and Falconer, 1967; Milner, 1968; Jones-Gotman, 1986). Such findings gave rise to the ipsilateral deficit model, or *material-specific model*, which asserts that memory function lateralizes with cerebral function: in left-language dominant individuals, the left hemisphere mediates verbal memory, and the right hemisphere visuospatial memory (Saykin, et al., 1992).

Neuroimaging results have not been entirely consistent with lesion data regarding material-specificity during memory processing. Although several studies have demonstrated material-specific laterality in the frontal lobes (Wagner, et al., 1998; McDermott, et al., 1999) and in the mTL (Grady, et al., 1995; Stern, et al., 1996; Nyberg, et al., 1996a; Kelly, et al., 1998; Detre, et al., 1998), numerous studies suggest that hemispheric effects depend upon the memory process being instantiated (encoding vs. retrieval), rather than the type of stimulus material (Tulving, et al., 1994; Schacter, et al., 1995; Nyberg, et al., 1996b). Other studies suggest that the right and left medial temporal regions respond differentially to novel and familiar stimuli (Tulving, et al., 1996; Fujii, et al., 1997), or that laterality of activation varies with depth of encoding (Nyberg, et al., 1996a; Martin, et al., 1997), success of encoding (Casasanto, et al., 2000), or with task parameters such as the stimulus presentation rate (Kelly, et al., 1998).

The present study examined fMRI activation during intentional encoding of unfamiliar faces, complex visual scenes, and four-word declarative sentences. The goal of the study was to determine whether activation in the mTL lateralizes according to the type of stimulus material presented. Other variables that may affect hemispheric laterality, such as stimulus novelty, task instructions, and stimulus presentation parameters, were held constant across the three tasks. It was hypothesized that encoding of unfamiliar faces would be associated with preferential activation of right-hemisphere mTL structures, encoding of sentences with preferential activation of left-hemisphere mTL structures, and encoding of complex visual scenes, which are amenable to both verbal and visuospatial encoding, would be associated with bilateral mTL activation.

## Materials and Methods

### Subjects

Healthy, normal volunteers between the ages of 18 and 30 were consecutively recruited from the University of Pennsylvania community, and paid \$20 for their

participation (Scenes task: N=19, 6 male; Sentence Task: N=15, 6 male; Face task: N=6, 2 male). All subjects were right-handed by self-report, and all of the sentence task participants were native speakers of English.

### **Cognitive Task Design**

For each encoding task, subjects viewed a total of 60 stimuli, presented over six 40-second blocks (10 stimuli per block, 3500 ms presentation, 500 ms ISI) while lying supine in the bore of the MRI scanner. Stimulus blocks alternated with blocks of control images, matched with target stimuli for size, color, luminosity, and presentation rate. Scene stimuli were obtained from a commercial library of digitized images (PhotoDisc, Inc., 1995, Seattle, WA). (See figure 1a.) Face stimuli were constructed from University of Pennsylvania ID card photographs. Consent for use of the photographs was solicited via an e-mail advertisement to approximately 3000 members of the University community, and only photographs from those providing consent were used. The face photographs were equated for size and image quality, and were cropped so as to include the brow, eyes, nose, and mouth, but exclude ears, hair, and any extraneous objects such as eyeglasses or jewelry. (See figure 1b.) Sentence stimuli were four-word, active, declarative sentences culled from children's books estimated to be at the fifth-grade reading level, and presented in Chicago 24-point font. Simple sentences were chosen so that this task could be administered to neurologically impaired patients with cognitive deficits, although all data reported presently pertain to healthy subjects. (See figure 1c.) For the faces and scenes tasks, the control images were visual noise patterns, created by transforming a stimulus image with a random retiling algorithm iterated 10,000 times. For the sentence task, the control image was a set of four strings, composed of asterisks, of the same mean length as the stimulus words. Stimulus presentation routines were developed on a Macintosh Powerbook (Apple Computer, Cupertino, CA), using Psyscope software (Cohen, et al., 1993). Stimuli were back-projected using an Epson LCD projector (model ELP-5000) onto a viewing screen positioned approximately 7 feet from the subject's eyes, which was easily visible via a mirror mounted in the scanner head coil. Subjects were instructed to remember the stimuli for a recognition test immediately following each encoding task, and to attend to the control images, but not to memorize them. The sequence of cognitive tasks was pseudorandomly varied across subjects.

For each recognition test, subjects viewed all sixty of the stimuli presented during the preceding encoding task, randomly intermixed with an equal number of novel distractors. While still lying in the scanner bore, subjects were required to distinguish studied stimuli from unstudied distractors, and to respond using a two-button box interfaced with the Macintosh computer via fiberoptic cable. The forced-choice recognition test was self-paced, and subjects were informed that both the speed and accuracy of their responses was of interest. Functional

imaging data were collected during encoding, but not during recognition testing.

### **Image Acquisition**

Imaging data were collected on a 1.5 Tesla GE Signa MRI scanner equipped with a fast gradient system for echo-planar imaging, using a standard quadrature radiofrequency (RF) whole-head coil. Foam padding was used to comfortably restrict head motion. Sagittal and axial T1-weighted structural images were obtained for each subject. Prior to functional activation, data were acquired for correction of image distortion due to static susceptibility effects (Alsop, 1995). T2\*-sensitive, gradient echo, echoplanar functional images were then obtained with BOLD contrast (TR = 2000 ms, TE<sub>eff</sub> = 50 ms) in 18 to 20 contiguous 5-mm-thick axial slices, in a 24-cm field of view with a 64x64 acquisition matrix, resulting in a nominal pixel resolution of 3.75 x 3.75 x 5 mm. For each encoding task, functional activation was measured over a single 240-scan run consisting of six 80-second task/control cycles. Raw imaging data were extracted onto digital audiotape (DAT) for subsequent analysis.

### **Image Processing and Data Analysis**

Imaging data were reconstructed offline on SUN UltraSparc workstations, (SUN Microsystems, Mountain View, CA) using software developed in Interactive Data Language (Research Systems, Boulder, CO). Raw data were corrected for static susceptibility-induced distortions, and a motion-compensation algorithm targeting translational artifacts occurring along three orthogonal vectors was applied to each data set. Data were convolved in space using a three-dimensional nonisotropic gaussian kernel (full width half-maximum [FWHM] = 8 X 8 X 10 mm). Using SPM97 software (Wellcome Laboratories, London, UK), a linear model for temporally autocorrelated observations was applied voxelwise to each data set. FMRI signal at each voxel was correlated to a reference function obtained by convolving the square wave describing the task/control alternation with an estimate of the subject's hemodynamic response function (Friston, et al., 1994). Statistical parametric maps (SPMs) were generated for each subject's encoding runs. Multisubject SPMs were then constructed for each task using the random effects model, with SPMt maps as input. Normalized group maps were viewed in Talairach atlas space, with across-subject averaged functional images superimposed on a standard pseudosubject structural image. Cognitive subtraction (task condition – control condition) produced a difference image showing activation associated stimulus encoding for each task.

Anatomical regions were defined using the SPL anatomy browser (Kikinis, et al., 1996), interfaced with IDL and SPM98 software. Based on these anatomical regions, an mTL region of interest (ROI) was defined comprising the hippocampus, parahippocampus, and

fusiform gyrus. Although whole-brain data were collected, secondary analysis was restricted to this a-priori defined region of interest. Only activation exceeding a mapwise statistical threshold ( $\alpha = .05$ ) was considered. Suprathreshold activation was quantified for each lateralized anatomical structure within the mTL ROI, by counting the number of active suprathreshold voxels. The hemispheric asymmetry of activation correlating with each cognitive task was determined by calculating an asymmetry ratio for each search region ( $AR = \text{Voxels}_R - \text{Voxels}_L / \text{Voxels}_R + \text{Voxels}_L$ ). The significance of activation asymmetry was assessed by comparing the proportion of active suprathreshold voxels in each lateralized search region, using a standard test for the independence of two proportions (Hinkle, et al., 1988).

Recognition test performance was assessed by computing a discriminability index for each subject (Discriminability = (% hits) – (% false positives)).

## Results

### Behavioral Results

Performance on the post-scan recognition tests confirmed that subjects were able to encode target stimuli satisfactorily. Results show that all subjects performed significantly above chance on all tasks. Subjects' mean discriminability score for the face task was 0.50 (SD +/- .20,  $t = 6.13$ ,  $p = .0008$ ), for the scene task 0.80 (SD +/- .17,  $t = 19.18$ ,  $p = .0001$ ), and for the sentence task 0.70 (SD +/- .17,  $t = 13.97$ ,  $p = .0001$ ).

### Imaging Results

Suprathreshold activation associated with encoding was found in the mTL region of interest across all three tasks. Table 1 presents the Talairach locations of peak activation during encoding for each anatomical structure within the ROI. Figure 2 presents selected slices of the multisubject functional activation maps for each encoding task. It was observed, for the face and scene tasks, that active suprathreshold voxels in the parahippocampus were contiguous with those in the fusiform gyrus, constituting a "cluster" of active voxels. Hippocampal activations formed separate clusters. Therefore, for analysis of hemispheric effects, the region of interest was divided into two subregions: a hippocampal ROI comprising the hippocampus proper (horn of Ammon, subiculum, and dentate gyrus), and a posterior mTL ROI comprising the parahippocampus (perirhinal and entorhinal cortices) and the fusiform gyrus. Figure 3 shows the hemispheric asymmetry of activation across tasks, as indicated by the asymmetry ratio computed for each search region. For the face task, bilateral activation was found in the hippocampus, nonsignificantly greater left than right ( $AR = -0.27$ ,  $ns$ ), and in the posterior mTL, significantly greater right than left ( $AR = 0.33$ ,  $p < .05$ ). For the scene task, unilateral activation was found in the left hippocampus ( $AR = -1.0$ ,  $p < .001$ ), and bilateral activation was found in the posterior mTL ( $AR = -0.03$ ,

$ns$ ). For the sentence task, unilateral left hemisphere activation was found both in the hippocampus ( $AR = -1.0$ ,  $p < .001$ ) and in the posterior mTL ( $AR = -1.0$ ,  $p < .05$ ).

## Discussion

Across encoding tasks, the pattern of activation in the posterior mTL is consistent with the material-specific hypothesis. Greater right than left hemisphere activation was found during face encoding, nearly symmetrical bilateral activation during scene encoding, and exclusively left-sided activation during sentence encoding. The hemispheric laterality of activation can be interpreted as "code-specific" (McDermott, et al., 1999): that is, varying with the extent to which the stimuli can be processed using verbal and nonverbal *representational codes* (Paivio, 1991), the neural substrates of which have been shown to be differentially lateralized (Kounios and Holcomb, 1994; Kelly, et al., 1998).

Surprisingly, we observed that within task, the laterality of activation in the posterior mTL was not always consistent with the laterality of activation in the hippocampus. Hippocampal activation was bilateral during face encoding, and exclusively left-sided during scene and sentence encoding. In contrast to the material-specific activation observed in the posterior mTL, activation in the hippocampus during face and scene encoding did not lateralize according to the material-specific hypothesis. Previous studies have reported activation of left mTL structures during intentional encoding across all material types (Martin, 1997; Kelly, et al., 1998). However, the dissociation we observe between the laterality of activation in the hippocampus and posterior mTL structures during face and scene encoding has not been reported previously. It may be possible to account for our findings in terms of the neural connectivity of the mTL and surrounding structures. Hemispheric specialization for verbal and nonverbal materials has been well established in the neocortex. Because the parahippocampus receives direct input from the cortical sensory association areas, whereas the hippocampus receives the majority of its cortical input indirectly, via the parahippocampus (Eichenbaum and Bunsey, 1995), material-specific hemispheric effects may be observed more readily in parahippocampus than in the hippocampus. Furthermore, hemispheric specialization in the hippocampus may be masked due to integration of the right and left hippocampi, which are reciprocally connected via the hippocampal commissure.

Our findings are compatible with the two-component model of mTL memory processing developed by Eichenbaum and colleagues (1994), which suggests a functional dissociation between the hippocampus and posterior mTL structures. Specifically, the model implicates the parahippocampal region in the intermediate-term storage and maintenance of *individual mental representations*, and the hippocampus in the formation of *relations among mental representations*.

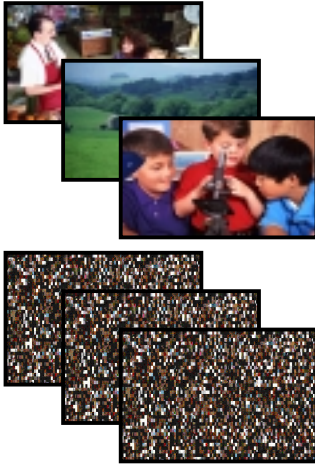


Figure 1a:  
SceneTask and Control Stimuli.

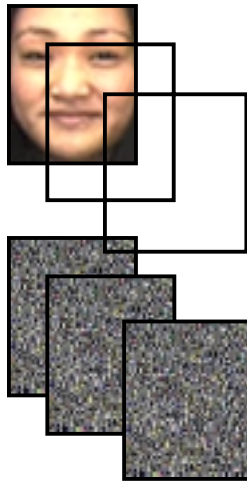


Figure 1b:  
Face Task and Control Stimuli.

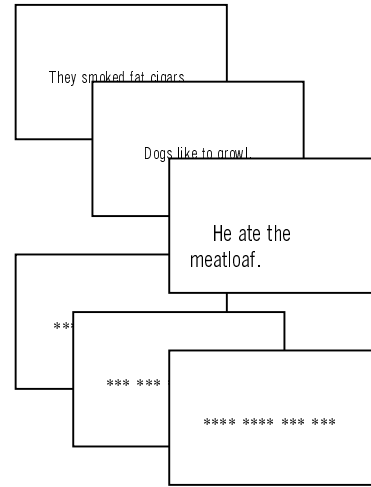


Figure 1c:  
Sentence Task and Control Stimuli.

Table 1: Talairach coordinates and Z-scores of the local maxima within ROI.

Region	x	y	z	Total Volume	Active Volume	Mean Z	Maximum Z
<b>Faces</b>							
Left Hippocampus	-24	-20	-15	64	7	2.37	3.18
Left Parahippocampus	--	--	--	84	0	--	--
Left Fusiform Gyrus	-44	-52	-20	155	12	2.32	3.28
Right Hippocampus	20	-12	-10	69	4	1.98	2.16
Right Parahippocampus	28	-16	-35	77	1	1.96	1.96
Right Fusiform Gyrus	40	-52	-25	134	23	2.02	2.99
<b>Scenes</b>							
Left Hippocampus	-20	-36	-1	64	12	2.25	3.20
Left Parahippocampus	-24	-40	-11	84	6	2.87	4.13
Left Fusiform Gyrus	-40	-48	-21	155	33	2.98	4.26
Right Hippocampus	--	--	--	69	0	--	--
Right Parahippocampus	20	-48	-11	77	22	2.25	3.20
Right Fusiform Gyrus	20	-44	-16	134	14	2.48	3.61
<b>Sentences</b>							
Left Hippocampus	-20	-40	5	64	11	2.00	2.42
Left Parahippocampus	--	--	--	84	0	--	--
Left Fusiform Gyrus	-44	-52	-20	155	4	2.32	3.28
Right Hippocampus	--	--	--	69	0	--	--
Right Parahippocampus	--	--	--	77	0	--	--
Right Fusiform Gyrus	--	--	--	134	0	--	--

Note. Mean Z score indicates the average of all suprathreshold voxels within the ROI. Active volume represents the number of voxels within the ROI exceeding the significance threshold ( $\alpha = .05$ ).

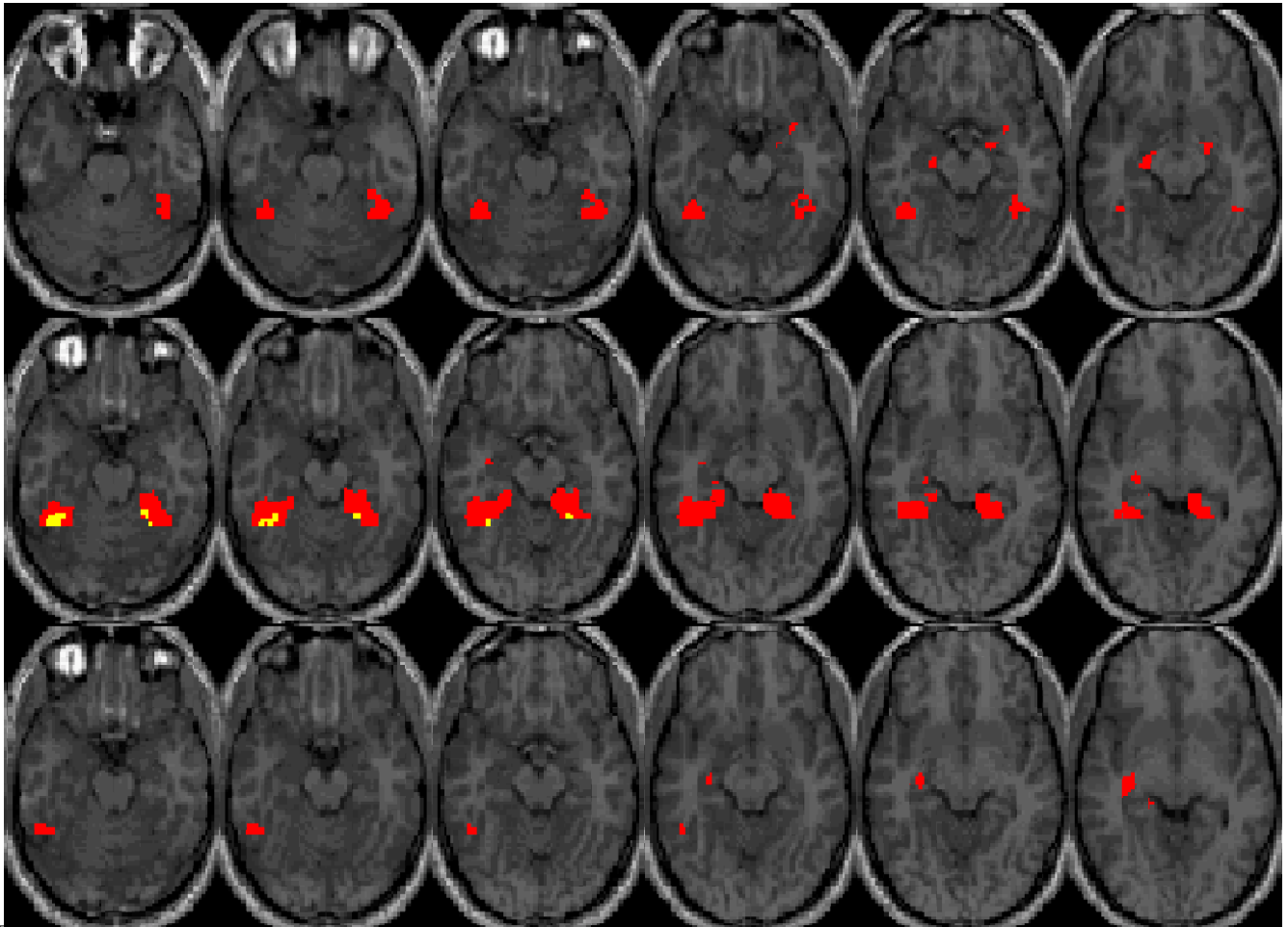


Figure 2: Multisubject statistical parametric maps (SPMs) of functional activation within the mTL region of interest correlating with each encoding task (top row: face encoding; middle row: scene encoding; bottom row: sentence encoding). The left side of each image corresponds to the left side of the brain. Regions demonstrating suprathreshold activation ( $p < .05$ ) during the task - control conditions are displayed in the red-to-yellow color scale.

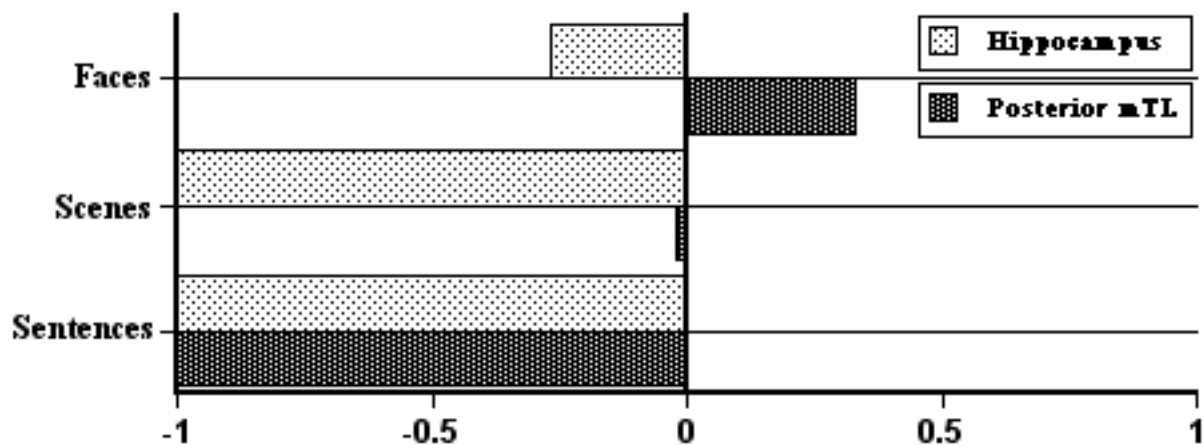


Figure 3: Hemispheric asymmetry of activation across encoding tasks in the hippocampus and posterior mTL. The horizontal axis indicates the Asymmetry Ratio (AR) calculated to show the hemispheric distribution of active suprathreshold voxels ( $p < .05$ ) within each search region ( $AR = \text{Voxels}_R - \text{Voxels}_L / \text{Voxels}_R + \text{Voxels}_L$ ).



This model, based primarily on lesion studies in humans and animals, is supported by recent human electrophysiological data that show a temporal dissociation between parahippocampal and hippocampal activation during encoding (Fernandez, et al., 1999), and by PET data that show increased hippocampal activation during relational vs. non-relational memory processing (Henke, et al., 1999). *Individual* mental representations may be either verbal or nonverbal, whereas *relational* representations may combine verbal and nonverbal codes. Although the Eichenbaum model makes no explicit predictions regarding material-specificity, it provides a theoretical framework in which to consider our finding that the posterior mTL shows greater sensitivity to material type than the hippocampus.

## Conclusions

Whereas hemispheric laterality during memory encoding in the posterior mTL appears to be strongly code-dependent, laterality in the hippocampus may depend upon other variables, as well. Within the frontal lobes, hemispheric effects have been shown to depend upon material type for certain anatomical structures, and upon cognitive set (encoding vs. retrieval) for other nearby structures (McDermott, et al., 1999). This pattern may be extensible to the mesial temporal lobes. Future studies may identify variables affecting the laterality of activation in specific mTL subcomponents, and may help to reconcile neuropsychological findings that suggest material-specific hippocampal involvement in memory processing with conflicting neuroimaging results.

This report was supported in part by NIH grant NS37488.

## References

- Alsop, D. C. (1995). Correction of ghost artifacts and distortion in echoplanar MR with an iterative reconstruction technique. *Radiology*(197P): 338.
- Casasanto, D., W. D. S. Killgore, et al. (2000). *Neural Correlates of Successful and Unsuccessful Verbal Encoding*. Society for Human Brain Mapping 6th Annual Meeting, San Antonio, Academic Press.
- Cohen, J. D., B. MacWhinney, et al. (1993). (*Behavior Research Methods, Instruments, & Computers* 25(2): 257-271.
- Detre, J. A., L. Maccotta, et al. (1998). Functional MRI lateralization of memory in temporal lobe epilepsy. *Neurology* 50(4): 926-32.
- Dobbins, I. G., N. E. Kroll, et al. (1998). Unilateral medial temporal lobe memory impairment: type deficit, function deficit, or both? *Neuropsychologia* 36(2): 115-27.
- Eichenbaum, H., T. Otto, et al. (1994). Two functional components of the hippocampal memory system. *Behavioral & Brain Sciences* 17(3): 449-517.
- Eichenbaum, H., M. Bunsey (1995). On the binding of associations in memory. *Current Directions in Psychological Science* 4(1): 19-23.
- Fernandez, G., A. Effern, et al. (1999). Real-time tracking of memory formation in the human rhinal cortex and hippocampus. *Science* 285(5433): 1582-5.
- Friston, K.J., P. Jezzard, et al. (1994) *Human Brain Mapping*
- Fujii, T., J. Okuda, et al. (1997). Different roles of the left and right parahippocampal regions in verbal recognition: a PET study. *Neuroreport* 8(5): 1113-7.
- Glosser, G., A. J. Saykin, et al. (1995). Neural Organization of Material-Specific Memory Functions in. *Neuropsychology* 9(4): 449-456.
- Grady, C. L., A. R. McIntosh, et al. (1995). Age-related reductions in human recognition memory due to impaired encoding. *Science* 269(5221): 218-21.
- Henke, K., B. Weber, et al. (1999). Human hippocampus associates information in memory. *Proc Natl Acad Sci U S A* 96(10): 5884-9.
- Hinkle, D. E., W. Wiersma, et al. (1988). *Applied Statistics for the Behavioral Sciences*. Boston, Houghton Mifflin.
- Jones-Gotman, M. (1986). Right hippocampal excision impairs learning and recall of a list of abstract designs. *Neuropsychologia* 24(5): 659-70.
- Kelley, W. M., F. M. Miezin, et al. (1998). Hemispheric specialization in human dorsal frontal cortex and medial temporal lobe for verbal and nonverbal memory encoding. *Neuron* 20(5): 927-36.
- Kikinis, R., P. L. Gleason, et al. (1996). Computer-assisted interactive three-dimensional planning for neurosurgical procedures. *Neurosurgery* 38(4): 640-9; discussion 649-51.
- Kounios, J. and P. J. Holcomb (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *J Exp. Psych.: Learning, Memory, & Cognition* 20(4): 804-823.
- Martin, A., C. L. Wiggs, et al. (1997). Modulation of human medial temporal lobe activity by form, meaning, and experience. *Hippocampus* 7(6): 587-93.
- McDermott, K. B., R. L. Buckner, et al. (1999). Set- and code-specific activation in frontal cortex: an fMRI study of encoding and retrieval of faces and words. *J Cogn Neurosci* 11(6): 631-40.
- Milner, B. (1958). Psychological deficits produced by temporal lobe excision. *Research Publications/ Association for research in Nervous and Mental Disease* 36: 244-257.
- Milner, B. (1968). Visual recognition and recall after right temporal lobe excision in man. *Neuropsychologia* 6: 191-209.
- Nyberg, L., R. Cabeza, et al. (1996). PET studies of encoding and retrieval: The Hera model. *Psychonomic Bul* 3(2): 135-148.
- Nyberg, L., A. R. McIntosh, et al. (1996). Activation of medial temporal structures during episodic memory retrieval [see comments]. *Nature* 380(6576): 715-7.
- Saykin, A. J., L. J. Robinson, et al. (1992). Neuropsychological changes after anterior temporal lobectomy. *The Neuropsychology of Epilepsy*. T. L. Bennett. New York, Plenum: 263-290.
- Schacter, D. L., E. Rieman, et al. (1995). Brain regions associated with retrieval of structurally coherent visual information. *Nature* 376: 587-590.
- Scoville WB, M. B. (1957). Loss of recent memory after bilateral hippocampal lesions. *J Neurol Neurosurg Psychiatry* 20: 11-21.
- Squire, L. R. and S. Zola-Morgan (1991). The medial temporal lobe memory system. *Science* 253(5026): 1380-6.
- Stern, C. E., S. Corkin, et al. (1996). The hippocampal formation participates in novel picture encoding: evidence from functional magnetic resonance imaging. *Proc Natl Acad Sci U S A* 93(16): 8660-5.
- Tulving, E. and H. J. Markowitsch (1998). Episodic and declarative memory: role of the hippocampus. *Hippocampus* 8(3): 198-204.
- Tulving, E., H. J. Markowitsch, et al. (1996). Novelty and familiarity activations in PET studies of memory encoding and retrieval. *Cereb Cortex* 6(1): 71-9.
- Wagner, A. D., R. A. Poldrack, et al. (1998). Material-specific lateralization of prefrontal activation during episodic encoding and retrieval. *Neuroreport* 9(16): 3711-7.

# A Connectionist Single-Mechanism Account of Rule-Like Behavior in Infancy

**Morten H. Christiansen** (morten@siu.edu)

**Christopher M. Conway** (conway@siu.edu)

Department of Psychology; Southern Illinois University  
Carbondale, IL 62901-6502 USA

**Suzanne Curtin** (curtin@gizmo.usc.edu)

Department of Linguistics; University of Southern California  
Los Angeles, CA 90089-1693 USA

## Abstract

One of the most controversial issues in cognitive science pertains to whether rules are necessary to explain complex behavior. Nowhere has the debate over rules been more heated than within the field of language acquisition. Most researchers agree on the need for statistical learning mechanisms in language acquisition, but disagree on whether rule-learning components are also needed. Marcus, Vijayan, Rao, & Vishton (1999) have provided evidence of rule-like behavior which they claim can only be explained by a dual-mechanism account. In this paper, we show that a connectionist single-mechanism approach provides a more parsimonious account of rule-like behavior in infancy than the dual-mechanism approach. Specifically, we present simulation results from an existing connectionist model of infant speech segmentation, fitting the behavioral data under naturalistic circumstances without invoking rules. We further investigate diverging predictions from the single- and dual-mechanism accounts through additional simulations and artificial language learning experiments. The results support a connectionist single-mechanism account, while undermining the dual-mechanism account.

## Introduction

The nature of the learning mechanisms that infants bring to the task of language acquisition is a major focus of research in cognitive science. With the rise of connectionism, much of the scientific debate surrounding this research has focused on whether rules are necessary to explain language acquisition. All parties in the debate acknowledge that statistical learning mechanisms form a necessary part of the language acquisition process (e.g., Christiansen & Curtin, 1999; Marcus, Vijayan, Rao, & Vishton, 1999; Pinker, 1991). However, there is much disagreement over whether a statistical learning mechanism is sufficient to account for complex rule-like behavior, or whether additional rule-learning mechanisms are needed. In the past this debate has primarily taken place within specific areas of language acquisition, such as inflectional morphology (e.g., Pinker, 1991; Plunkett & Marchman, 1993) and visual word recognition (e.g., Coltheart, Curtis, Atkins & Haller, 1993; Seidenberg & McClelland, 1989). More recently, Marcus et al. (1999) have presented results from experiments with 7-month-olds, apparently showing that infants acquire abstract algebraic rules after two minutes of exposure to habituation stimuli. The algebraic rules are construed as representing an open-ended relationship between variables for which one can substitute arbitrary values, “such as ‘the first item X is the same as the third item Y,’ or more generally, that ‘item I is the same as item J’” (Marcus et al., 1999, p. 79). Marcus et al. further claim that a connectionist single-mechanism approach based on statistical learning is unable to

fit their experimental data. In this paper, we build on earlier work (Christiansen & Curtin, 1999) and present a detailed connectionist model of these infant data, and provide new experimental data that support a statistically-based single-mechanism approach while undermining the dual-mechanism account.

In the remainder of this paper, we first show that knowledge acquired in the service of learning to segment the speech stream can be recruited to carry out the kind of classification task used in the experiments by Marcus et al. For this purpose we took an existing model of early infant speech segmentation (Christiansen, Allen & Seidenberg, 1998) and used it to simulate the results obtained by Marcus et al. The simulations demonstrate that no rules are needed to account for the data; rather, statistical knowledge related to word segmentation can explain the rule-like behavior of the infants in the Marcus et al. study. We then explore the issue of timing in stimuli presentation and present additional simulations from which empirical predictions are derived that diverge from those of the rule-based account. These predictions are tested in experiments with adults. Experiment 1 replicated the results from Marcus et al. using adult subjects. Experiment 2 confirmed the predictions from our single-mechanism approach, whereas the dual-mechanism approach cannot account for these results without adding extra machinery to complement the statistical and rule-based components. Together, the simulations and the experiments thus suggest that a single-mechanism model provides the most parsimonious account of the empirical data presented here and in Marcus et al., thus obviating the need for a separate rule-based component.

## Simulation 1: Rule-Like Behavior without Rules

Marcus et al. (1999) used an artificial language learning paradigm to test their claim that the infant has two mechanisms for learning language, one that uses statistical information and another which uses algebraic rules. They conducted three experiments which tested infants’ ability to generalize to items not presented in the familiarization phase of the experiment. We focus here on their third experiment because it was controlled for possible confounds found in the first two experiments: differences in phonetic features (Experiment 1) and reduplication<sup>1</sup> (Experiment 2). Marcus et al. claim that

<sup>1</sup>Though the control for reduplication was not entirely complete (see Elman, 1999).

because none of the test items appeared in the habituation part of the experiment the infants would not be able to use statistical information in this task.

The subjects in Experiment 3 of Marcus et al. (1999) were 16 7-month-old infants randomly placed in an AAB or an ABB condition. During a two-minute long familiarization phase the infants were exposed to three repetitions of each of 16 three-word sentences. Each word in the sentence frame AAB or ABB consisted of a consonant-vowel sequence (e.g., “le le we” or “le we we”). The test phase consisted of 12 sentences made up of words to which the infants had not previously been exposed (e.g., “ko ko ga” vs. “ko ga ga”). The test items were broken into two groups for both habituation conditions: consistent (items constructed with the same sentence frame as the familiarization phase) and inconsistent (constructed from the sentence frame the infants were not habituated on). The results showed that the infants preferred the inconsistent test items to the consistent ones (that is, they listened longer to the inconsistent items).

The conclusion drawn by Marcus et al. (1999) was that a single mechanism which relied on statistical information alone could not account for the results. Instead they suggested that a dual mechanism was needed, comprising a statistical learning component and an algebraic rule learning component. In addition, they claimed that a Simple Recurrent Network (SRN; Elman, 1990) would not be able to accommodate their data because of the lack of phonological overlap between habituation and test items. Specifically, they state,

Such networks can simulate knowledge of grammatical rules only by being trained on all items to which they apply; consequently, such mechanisms cannot account for how humans generalize rules to new items that do not overlap with the items that appeared in training (p. 79).

In the first simulation, we demonstrate that SRNs can indeed fit the data from Marcus et al. Other researchers have constructed neural network models specifically to simulate the Marcus et al. results (Altmann & Dienes, 1999; Elman, 1999; Shastri & Chang, 1999; Shultz, 1999). In contrast, we do *not* build a new model to accommodate the results, but take an existing SRN model of speech segmentation (Christiansen et al., 1998) and show how this model—*without additional modification*—provides an explanation for the results.

### The Christiansen et al. Model

The model by Christiansen et al. (1998) was developed as an account of early word segmentation. An SRN was trained on a *single* pass through a corpus of child directed speech. As input the network was provided with three probabilistic cues to word boundaries: (a) phonology represented in terms of 11 features on the input and 36 phonemes on the output, (b) utterance boundary information represented as an extra feature marking utterance endings, and (c) lexical stress coded over two units as either no stress, secondary or primary stress. Figure 1 provides an illustration of the network.

The network was trained on the task of predicting the next phoneme in a sequence as well as the appropriate values for the utterance boundary and stress units. In learning to perform this task the network learned to integrate the cues such

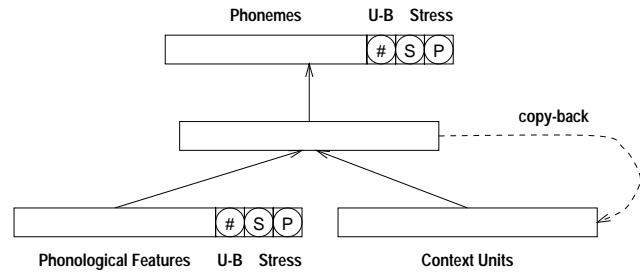


Figure 1: Illustration of the SRN used in Simulations 1 and 2. Solid lines indicate trainable weights, whereas the dashed line denotes the copy-back weights (which are always 1). U-B refers to the unit coding for the presence of an utterance boundary. The presence of lexical stress is represented in terms of two units, S and P, coding for secondary and primary stress, respectively.

that it could carry out the task of segmenting the input into words. This involved activating the boundary unit not only at utterance boundaries, but also at word boundaries occurring inside utterances. The logic behind the segmentation task is that the end of an utterance is also the end of a word. If the network is able to integrate the provided cues in order to activate the boundary unit at the ends of words occurring at the end of an utterance, it should also be able to generalize this knowledge so as to activate the boundary unit at the ends of words which occur *inside* an utterance (Aslin, Woodward, LaMendola & Bever, 1996).

The Christiansen et al. (1998) model acquired distributional knowledge about sequences of phonemes and the associated stress patterns. This knowledge allowed it to perform well on the task of segmenting the speech stream into words. We suggest that this knowledge can be put to use in secondary tasks not directly related to speech segmentation—including the artificial language task used by Marcus et al. (1999). In fact, the experimental procedure used by Marcus et al. was the same as the procedure used by Saffran, Aslin & Newport (1996) to study how word segmentation in infancy can be facilitated by statistical learning. That is, Marcus et al. sought to demonstrate that the statistically-based learning mechanism, which Saffran, Aslin, et al. found to be involved in word segmentation, could not account for their results. It therefore makes sense to investigate whether the comprehensive speech segmentation model by Christiansen et al. can account for the Marcus et al. infant results.

### Method

**Networks** Corresponding to the 16 infants in the Marcus et al. study, we used 16 SRNs similar to the SRN used in Christiansen et al. (1998) with the exception that the original phonetic feature geometry was replaced by a new representation using 18 features. Each of the 16 SRNs had a different set of initial weights, randomized within the interval [0.25;-0.25]. The learning rate was set to 0.1 and the momentum to 0.95. These training parameters were identical to those used in the original Christiansen et al. model. The networks were trained to predict the correct constellation of cues given the current

input segment.

**Materials** Prior to being habituated and tested on the stimuli from Marcus et al., the networks were first exposed to the training corpus used by Christiansen et al. This corpus consists of 8181 utterances extracted from the Korman (1984) corpus of British English speech directed at pre-verbal infants aged 6-16 weeks (a part of the CHILDES database, MacWhinney, 1991). Christiansen et al. transformed each word in the utterances from its orthographic format into a phonological form with accompanying lexical stress using a dictionary compiled from the MRC Psycholinguistic Database available from the Oxford Text Archive.

The materials from Experiment 3 in Marcus et al. (1999) were transformed into the phoneme representation used by Christiansen et al. Two habituation sets were created in this manner: one for AAB items and one for ABB items. The habituation sets used here, and in Marcus et al., consisted of 3 blocks of 16 sentences in random order, yielding a total of 48 sentences in each habituation set. As in Marcus et al. there were four different test sentences: “ba ba po”, “ko ko ga” (consistent with AAB), “ba po po” and “ko ga ga” (consistent with ABB). The test set consisted of three blocks of randomly ordered test sentences, totaling 12 test items. Both the habituation and test sentences were treated as a single utterance with no explicit word boundaries marked between the individual words. The end of each utterance was marked by activating the utterance boundary unit.

**Procedure** The networks were first trained on a single pass through the Korman (1984) corpus as the original Christiansen et al. model. This corresponds to the fact that the 7-month-olds in the Marcus et al. study already have had a considerable exposure to language, and have begun to develop their speech segmentation abilities. Next, the networks were habituated on a *single* pass through the appropriate habituation corpus—one phoneme at a time—with learning parameters identical to the ones used during the pretraining on the Korman corpus. The networks were then tested on the test set (with the weights “frozen”) and the activation of the utterance boundary unit was recorded for every phoneme input in the test set. Finally, the boundary unit activations for test sentences that were consistent or inconsistent with the habituation pattern were separated into two groups. Furthermore, for the purpose of scoring word segmentation performance on the test items, the activation of the boundary unit was also recorded for each habituation condition across all the habituation items and the mean activation was calculated. The networks were said to have postulated a word boundary whenever the boundary unit activation in a test sentence was above the appropriate habituation mean.

## Results and Discussion

To provide a quantitative measure of performance we used completeness scores (Christiansen et al., 1998) to assess segmentation performance.

$$\text{Completeness} = \frac{\text{Hits}}{\text{Hits} + \text{Misses}} \quad (1)$$

Completeness provides a measure of how many of the words in a test set the net is able to discover. With respect to our in-

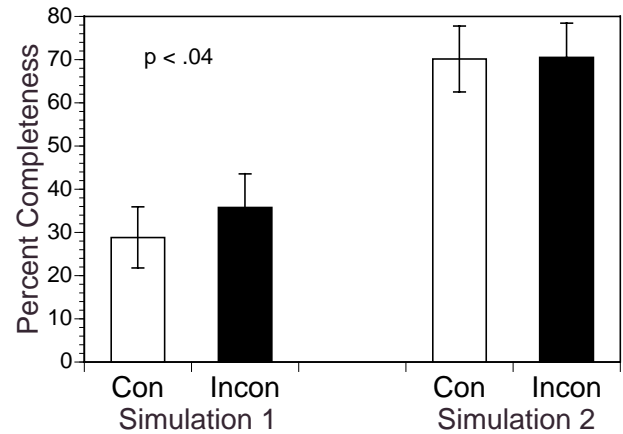


Figure 2: Mean completeness scores for the consistent (CON) and inconsistent (INCON) test items from Simulations 1 (left) and 2 (right).

terpretation of the Marcus et al. data, the completeness score indicates how well networks/infants are at segmenting out the individual words in the test sentences. As an example, consider the following hypothetical segmentation of two test sentences:

# b a b # a # p o # k o # g a g # a #

where ‘#’ corresponds to a predicted word boundary. Here the hypothetical learner correctly segmented out two words, *po* and *ko*, but missed the first and the second *ba* and the first and the second *ga*. This results in a completeness score of  $2/(2+4) = 33.3\%$ .

For each of the sixteen networks, completeness scores were computed across all test items, and submitted to the same statistical analyses as used by Marcus et al. for their infant data. The completeness scores were analyzed in a repeated measures ANOVA with condition (AAB vs. ABB) as between network factor and test pattern (consistent vs. inconsistent) as within network factor. The left-hand side of Figure 2 shows the completeness scores for the consistent and inconsistent items pooled across conditions. There was a main effect of test pattern ( $F(1, 14) = 5.76, p < .04$ ), indicating that the networks were significantly better at segmenting out the words in the inconsistent items (35.76%) compared with the consistent items (28.82%). Similarly to the infant data, neither the main effect of condition, nor the condition  $\times$  test pattern interaction were significant ( $F's < 1$ ). The better segmentation of the inconsistent items suggests that they would stand out more clearly in comparison with the consistent items, and thus explain why the infants looked longer towards the speaker playing the inconsistent items in the Marcus et al. study.

Simulation 1 shows that a separate rule-learning component is not necessary to account for the Marcus et al. (1999) data. An existing SRN model of word segmentation can fit these data without invoking explicit, algebraic-like rules. The pretraining allowed the SRNs to learn to integrate the regularities governing the phonological, lexical stress, and utterance boundary information in child-directed speech. During the habituation phase, the networks then developed weak attrac-

tors specific to the habituation pattern *and* the syllables used. The attractor will at the same time both attract a consistent item (because of pattern similarity) and repel it (because of syllable dissimilarity), causing interference with the segmentation task. The inconsistent items, on the other hand, will tend to be repelled by the habituation attractors and therefore do not suffer from the same kind of interference, making them easier for the network to process.

Importantly, the SRN model—as a statistical learning mechanism—can explain both the distinction between consistent and inconsistent items as well as the preference for the inconsistent items. Note that a rule-learning mechanism by itself only can explain how infants may distinguish between items, but not why they prefer inconsistent over consistent items. Extra machinery is needed in addition to the rule-learning mechanism to explain the preference for inconsistent items. Thus, the most parsimonious explanation is that only a statistical learning device is necessary to account for the infant data. The addition of a rule-learning device does not appear to be necessary.

### Simulation 2: It's about Time

Simulation 1 demonstrated that a statistically-based single-mechanism approach can account for the kind of rule-like behavior displayed by the infants in the Marcus et al. study. However, there may be other cases in which a separate rule-learning component would be required. Here we explore one such case in which our model makes a prediction which is different from what would be predicted from a dual-mechanism approach incorporating a rule-learning component.

Recall that algebraic rules were characterized as abstract relationships between variables, such as item X is the same as item Y. Marcus et al. Experiment 3 was designed to demonstrate that rule learning is independent of the physical realization of variables in terms of phonological features. The same rule, AAB, applies to—and can be learned from—“le le we” and “ko ko ga” (with “le” and “ko” filling the same A slot and “we” and “ga” the same B slot). As the abstract relationships that this rule represents only pertains to the value of the three variables, the amount of time between them should not affect the application of the rule. Thus, just as the physical realization of a variable does not matter for the learning or application of a rule, neither should the time between variables. The same rule AAB, applies to—and can be learned from—“le [250ms] le [250ms] we” and “le [1000ms] le [1000ms] we” (the “le”s should still fill the A slots and the “we”s the B slot despite the increased duration of time between the occurrence of these variables). From this property, one can predict that lengthening the time between variables should not affect the preference for inconsistent items. Indeed, the connectionist implementation of the rule-based approach found in the Shastri & Chang (1999) model would appear to make this prediction.

A lengthening of the pauses between words should, however, have a different effect on our model. In the model, the preference for inconsistent items observed by Marcus et al. is explained in terms of differential segmentation performance. Lengthening the pauses between words would in effect solve the segmentation task for the model, and should result in a disappearance of the preference for inconsistent items. Thus,

we predict that the model should show no difference between the segmentation performance on the consistent and inconsistent items if pauses are lengthened as indicated above. To test this prediction, we carried out a new set of simulations.

### Method

**Networks.** Sixteen SRNs as in Simulation 1.

**Materials.** Same as in Simulation 1 except that utterance boundaries were inserted *between* the words in the habituation and test sentences, simulating a lengthening of pauses between words (from 250 msec to 1000 msec) such that they have the same length as the pauses between utterances.

**Procedure.** Same as in Simulation 1.

### Results and Discussion

Completeness scores were computed as in Simulation 1 and submitted to the same statistical analysis. As illustrated by the right-hand side of Figure 2, the segmentation performance on the test items was improved considerably by the inclusion of utterance boundary-length pauses between words. As predicted, there was no difference between the accuracy scores for consistent (70.14%) and inconsistent items (70.49%) ( $F(1, 14) = .02$ ). As before, there was no main effect of condition, neither was there any interaction between condition and test pattern ( $F's < 1$ ).

Simulation 2 thus confirms the predicted effect of lengthening the pauses between words in stimuli presented to the statistical learning model. This results in diverging predictions derived from the rule-based and the statistical learning models concerning the effect of pause lengthening on human performance on the stimuli. Next, we test these diverging predictions in an artificial language learning experiment using adult subjects.

### Experiment 1: Replicating the Marcus et al. Results

Before testing the diverging predictions from the single- and dual-mechanism approaches we need to first establish whether adults in fact exhibit the same pattern of behavior as the infants in the Marcus et al. study. The first experiment therefore seeks to replicate Experiment 3 from Marcus et al. using adult subjects instead of infants.

### Method

**Subjects.** Sixteen undergraduates were recruited from introductory Psychology classes at Southern Illinois University. Subjects earned course credit for their participation.

**Materials.** For this experiment, we used the original stimuli that Marcus et al. (1999) created for their Experiment 3. Each word in a sentence was separated by 250 msec. The 16 habituation sentences for each condition were created by Marcus et al. using the Bell Labs speech synthesizer. The original habituation stimuli were limited to two predetermined sentence orders. To avoid potential order effects, we used the SoundEdit 16 version 2 software for the MacIntosh to isolate each sentence as a separate sound file. This allowed us to present the habituation sentences in a random order for each subject.

For the test phase, we also used the stimuli from Marcus et al.'s Experiment 3, which consisted of four new sentences that were either consistent or inconsistent with the training grammar. Like the habituation stimuli, each word in a sentence was separated by a 250 msec interval. As before, we stored the test stimuli as separate SoundEdit 16 version 2 sound files to allow a random presentation order for each subject.

**Procedure.** Subjects were seated in front of a G3 Power Macintosh computer with a New Micros button box. Subjects were randomly assigned to one of two conditions, AAB or ABB. The experiment was run using the PsyScope presentation software (Cohen, MacWhinney, Flatt, and Provost, 1993) with all stimuli played over stereo loudspeakers at 75dB. The subjects were instructed that they were participating in a pattern recognition experiment. They were told that in the first part of the experiment their task was to listen carefully to sequences of sounds and that their knowledge of these sound sequences would be tested afterwards. Subjects listened to three blocks of the sixteen randomly presented habituation sentences corresponding either to the AAB or the ABB sentence frame. A 1-second interval separated each sentence as was the case in the Marcus et al. experiment.

After habituation, subjects were instructed that they would be presented with new sound patterns that they had not previously heard. They were asked to judge whether a pattern was "similar" or "dissimilar" to what they had been exposed to in the previous phase by pressing an appropriately marked button. The instructions emphasized that because the sounds were novel, the subjects should not base their decision on the sounds themselves but instead on the patterns derived from the sounds. Subjects listened to three blocks of the four randomly presented test sentences. After the presentation of each test sentence, subjects were prompted for their response. Subjects were allowed to take as long as they needed to respond. Each test trial was separated by a 1000-msec interval.

## Results and Discussion

For the purpose of our analyses, the correct response for consistent items is "similar" while the correct response for inconsistent items is "dissimilar". The mean overall score for correct classification of test items was 8.81 out of a perfect score of 12. A single-sample t-test showed that this classification performance was significantly better than the chance level performance of 6 ( $t(15) = 4.44, p < .0005$ ). Subjects' responses were then subject to the same statistical analysis as the infant data in Marcus et al. (and Simulation 1 and 2 above). The left-hand side of Figure 3 shows the ratings as dissimilar for the six consistent and six inconsistent test items pooled across condition. As expected, there was a main effect of test pattern ( $F(1, 14) = 18.98, p < .001$ ), such that significantly more inconsistent items were judged as dissimilar (4.5) than consistent items (1.69). Neither the main effect of condition, nor the condition  $\times$  test pattern interaction were significant ( $F's < 1$ ).

Experiment 1 shows that adults perform similarly to the infants in Marcus et al.'s Experiment 3, thus demonstrating that it is possible to replicate their findings using adult subjects instead of infants. This result is perhaps not surprising given that Saffran and colleagues were able to replicate statistical learning results obtained using adults subjects

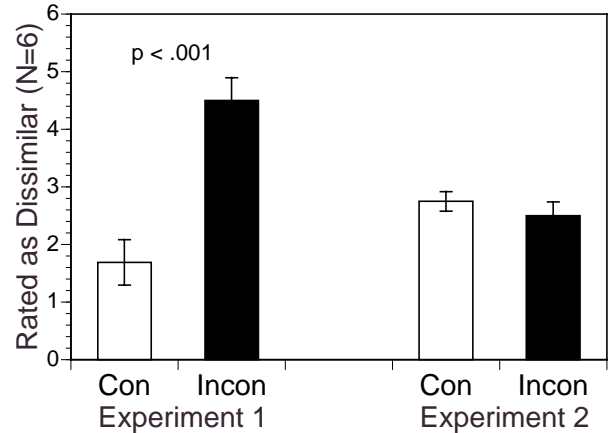


Figure 3: The mean proportion of consistent (con) and inconsistent (incon) test items rated as dissimilar to the habituation pattern in Experiments 1 (left) and 2 (right).

(Saffran, Newport & Aslin, 1996) in experiments using 8-month-olds (Saffran, Aslin, et al., 1996). More generally, these results and ours suggest that despite small differences in the experimental methodology used in infant and adult artificial language learning studies, both methodologies appear to tap into the same learning mechanisms. Also from a dual-mechanism approach, one would expect that the same learning mechanisms—statistical and rule-based—would be involved in both infancy and adulthood, and that similar results should be expected in both infant and adult studies of the kind of material used here.

## Experiment 2: Testing the Diverging Predictions

Having replicated the Marcus et al. (Experiment 3) infant data with adult subjects, we now turn our attention to the diverging predictions concerning the effect of pause length on the preference for the inconsistent items.

### Method

**Subjects.** Sixteen additional undergraduates were recruited from introductory Psychology classes at Southern Illinois University. Subjects earned course credit for their participation.

**Materials.** The training and test stimuli were the same as in Experiment 1 except that the 250 msec interval between words in a sentence was replaced by a 1000 msec interval using the SoundEdit 16 version 2 software. The 1000 msec interval between sentences remained the same as before.

**Procedure.** The procedure and instructions were identical to that used for Experiment 1.

### Results and Discussion

The mean overall classification score was 5.75 out of 12. This was not significantly different from the chance level performance of 6, as indicated by a single-sample t-test ( $t < 1$ ). The responses of the subjects were submitted to the same further

analysis as in Experiment 1. The right-hand side of Figure 3 shows the ratings as dissimilar for the consistent and inconsistent test items averaged across condition. As predicted by Simulation 2, there was *no* main effect of test pattern in this experiment ( $F(1, 14) = .56$ ), suggesting that subjects were unable to distinguish between consistent and inconsistent items. As in Experiment 1, both the main effect of condition and the interaction between condition and test pattern interaction were not significant ( $F's = 0$ ).

These results show that preference for inconsistent items disappears when the pauses between words are lengthened. This corroborates the prediction from the statistically-based single-mechanism model, but not the prediction from the rule-learning component of the dual-mechanism account. It may be objected that the rules need to work over specific domains, and that by lengthening the pauses between words the input is no longer chunked into sentences at a pre-specified length (three words). Hence, the rule can no longer be expected to apply. Note, however, that this requires additional machinery to pre-process the input prior to the learning or application of a rule. This would require a separate account of how this pre-processing ability was acquired and how it was applied in the specific case of Marcus et al.'s original experiment. Of course, this makes the rule-based account even less parsimonious in comparison with the statistical learning model. The latter model can account for both the preference for inconsistent items in the Marcus et al. Experiment 3 (and our Experiment 1) as well as the lack of preference in our Experiment 2 *without requiring any extra machinery*. Thus, a language learning device that exploits the statistical properties of language and integrates these multiple cues can account for the Marcus et al. data, thereby removing the need to posit a dual-learning mechanism.

## Conclusion

Infants possess powerful learning mechanisms that allow them to acquire language rapidly. Saffran, Aslin, et al. (1996) showed that infants can use statistical regularities to discover word boundaries in fluent speech. Marcus et al. (1999) found that infants exhibit rule-like behavior. Because both studies used the same experimental paradigm, a plausible null hypothesis is that both types of behavior should rely on the same learning mechanism. Based on unreported SRN simulations, Marcus et al. rejected this null hypothesis. In contrast, Simulation 1 demonstrated that that an existing SRN model of early infant word segmentation (Christiansen et al., 1998) could utilize statistical knowledge acquired in the service of speech segmentation to fit the infant data from Marcus et al. under very naturalistic circumstances. Experiment 2, which investigated the effect of "variable" timing on rule-like behavior, provided additional support for the single-mechanism approach. The results confirmed the predictions from our model (Simulation 2), but do not appear to fit the dual-mechanism approach because the amount of time between variables should not affect their abstract rule-based relationship. We note that the dual-mechanism account could possibly be augmented to account for these data, but that this would require the addition of extra machinery. Our single-mechanism model, on the other hand, can account for the data from Saffran, Aslin, et al. and Marcus et al. as well as

the results from Experiment 2 *without any modifications*, obviating the need for a separate rule-learning component. We therefore conclude that a connectionist single-mechanism approach provides the most parsimonious account of both statistical learning and rule-like behavior in infancy.

## Acknowledgments

We would like to thank Gary Marcus for making his stimuli available to us for our experiments, and Jeff Elman for his comments on an earlier version.

## References

- Altmann, G.T.M. & Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. *Science*, 284, 875.
- Aslin, R.N., Woodard, J.Z., LaMendola, N.P., & Bever, T.G. (1996). Models of word segmentation in fluent maternal speech to infants. In J.L. Morgan & K. Demuth (Eds.), *Signal to syntax*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Christiansen, M.H., Allen, J., & Seidenberg, M.S. (1998). Learning to segment using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13, 221-268.
- Christiansen, M.H. & Curtin, S.L. (1999). The power of statistical learning: No need for algebraic rules. In *The Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 114-119). Mahwah, NJ: Lawrence Erlbaum Associates.
- Cohen J.D., MacWhinney B., Flatt M., & Provost J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers*, 25, 257-271.
- Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, 100, 589-608.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.
- Elman, J. (1999). *Generalization, rules, and neural networks: A simulation of Marcus et al. (1999)*. Ms., University of California, San Diego.
- Korman, M. (1984). Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First Language*, 5, 44-45.
- MacWhinney, B. (1991). *The CHILDES Project*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Marcus, G.F., Vijayan, S., Rao, S.B., & Vishton, P.M. (1999). Rule learning in seven month-old infants. *Science*, 283, 77-80.
- Pinker, S. (1991). Rules of language. *Science*, 253, 530-535.
- Plunkett, K., & Marchman, V. (1993). From rote learning to system building. *Cognition*, 48, 21-69.
- Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month olds. *Science*, 274, 1926-1928.
- Saffran, J., Newport, E., & Aslin, R. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35, 606-621.
- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523-568.
- Shastri, L., & Chang, S. (1999). *A spatiotemporal connectionist model of algebraic rule-learning* (TR-99-011). Berkeley, California: International Computer Science Institute.
- Shultz, T. (1999). Rule learning by habituation can be simulated by neural networks. In *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 665-670). Mahwah, NJ: Lawrence Erlbaum Associates.

# Committing to an Ontology: A Connectionist Account

**Eliana Colunga** (ecolunga@cs.indiana.edu)

Computer Science Department; Lindley Hall 215  
Bloomington, IN 47405 USA

**Linda B. Smith** (smith4@indiana.edu)

Department of Psychology; 1101 East Tenth Street  
Bloomington, IN 47405 USA

## Abstract

Young children generalize nouns in systematic ways. They generalize names for solid things by shape and names for non-solid things by material. Recent evidence suggests that the source of these biases is in the children's lexicon: the bias becomes apparent only after they know names for things that are solid and have a similar shape and they know names for things that are non-solid and similar in material. In Experiment 1, we train a simple connectionist network with the regularities present in early noun vocabularies and show that this network shows generalization patterns comparable to those of young children. In Experiment 2 we look for other possible biases coming from statistical regularities and find that the network predicts that children will not cross ontological boundaries in their word generalizations. In Experiment 3 we test this prediction in 30-36 month-old children. We explain this finding in terms of the statistical regularities present in young children's noun vocabularies.

## Introduction

Young children are excellent learners of object names. After hearing a noun used once to name one object, they seem to know the scope of the whole category. To explain this proficiency people have proposed several mechanisms in the form of constraints or biases (Landau, Smith & Jones, 1988; Markman, 1989; Soja, Carey & Spelke, 1991). This paper is about the shape and material biases and about a new "bias", what one might call an "ontology bias". In the end, we propose that all these biases and constraints reduce to associative learning and generalization by similarity.

Our starting point is a recent study by Samuelson & Smith (1999). They examined the similarity structure of 300 object categories, the names of which are typically known by 30 month-olds. They found many nouns that name things that are solid and similar in shape and fewer nouns that refer to non-solid substances similar in material. They also showed that children do generalize novel nouns for solids by shape and for non-solids by material, but only after they know many of these words. These results suggest that these biases may be the product of statistical learning. In other words, children's noun generalizations are themselves generalizations over the nouns the child already knows.

In this paper we show that a simple statistical learner, when trained with the regularities present in early noun vo-

cabularies, generalizes novel nouns like children do. In Experiment 1 we train connectionist networks on the regularities found in early vocabulary by Samuelson and Smith (1999) and show that, like children, the networks generalize by shape for solid objects and by material for non-solid substances. In Experiment 2 we examine this early lexicon for other regularities that might create biases in a statistical learner and find that networks trained on this set exhibit what we call an "ontology bias". In Experiment 3 we test for this bias in children.

## Experiment 1

The goal of Experiment 1 is to determine if the regularities present in early noun vocabularies are sufficient to create word learning biases in a simple associative learner. If this is the case, it would support the idea that the biases are learned as part of learning the regularities in the lexicon. To do this we trained simple connectionist networks with a vocabulary organized using the regularities found in early lexicon by Samuelson and Smith (1999) and then we tested the network's performance on an adaptation of the novel noun generalization task.

## Architecture

We used a Hopfield network, which is a simple settling network. The network was trained using Contrastive Hebbian Learning (Movellan, 1990), an algorithm which adjusts weights on the basis of correlations between unit activations. Figure 1. shows the architecture of the network. The network has a Word Layer, in which words are represented locally. That is, each unit corresponds to one word in the network's vocabulary. Individual objects are represented on what we call the Dimension layer. Activation patterns on this layer represent the shape and material of each individual object or substance presented to the network. More specifically, the shape and material of an object (say the roundness of a particular ball and its yellow rubbery material) are represented by an activation pattern along the whole layer, in a distributed fashion. In the Solidity layer one unit stands for Solid and another for Non-Solid. Finally, there is a hidden layer that is connected to all the other layers and recurrently with itself. Note that the Word Layer and the Dimension and Solidity layers are only connected through the hidden layer, there are no direct connections among them.



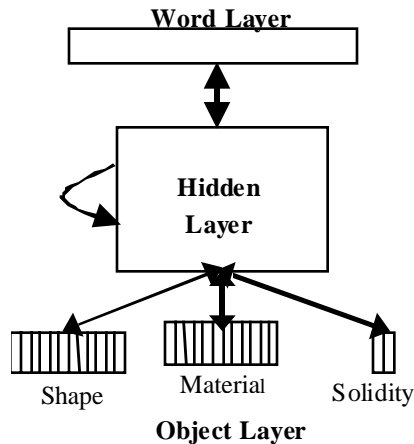


Figure 1. Architecture of the network used in Experiments 1 and 2.

### Training

The goal of the training phase was to mimic in the network the vocabulary learning that a child brings into a novel noun generalization experiment. We trained the networks on a subset of the nouns studied by Samuelson and Smith (1999). We specifically selected the names for objects and substances, excluding names for people, animals, places and abstract objects (e.g. wind). There were 149 training nouns. For each of these noun categories we used the adult judgments from Samuelson and Smith (199) to construct category exemplars. Importantly, although adults judged most solid things to be categorized by shape, there were exceptions and complications – e.g. muffins are judged to be alike in both shape and material and bubbles are judged to be non-solid but similar in shape. Our training instantiated the structures attributed to these words by adults.

More specifically, the statistical regularities across the noun vocabularies were built into the network’s training set in the following way. First, for each word that the network was to be taught, a pattern was generated to represent its value along the *relevant* dimension -- the dimension on which objects named by that noun were found to be similar. Second, at each presentation of the word, the value along the *irrelevant* dimension was varied randomly. For example, the word “ball” was judged to refer to things that are similar in shape; thus, a particular pattern of activation was randomly chosen and then assigned to represent ball-shape. All balls presented to the network were defined as having this shape, although each ball presented to the network also consisted of a unique and randomly generated pattern defining the material. So, each time the unit representing the word “ball” was turned on during training, the pattern representing ball-shape was presented along the shape dimension and a different, randomly generated pattern was presented along the material dimension.

Solid objects were assumed to have a bigger range of values along the shape dimension. This assumption is in line with the fact that solid things can hold more varied and complex shapes than non-solid things.

### Testing

We tested the networks in an analog of the novel noun generalization task used with children. Our approach is based on our conceptualization of the novel noun generalization task. In that task, the child sees an exemplar and hears its name. If, for example, the child attends exclusively to the shape of the named exemplar, then a test object that matches the exemplar in shape (although different from the exemplar in material) should be perceived as highly similar to the exemplar. Thus, we asked if the network’s internal representations – the patterns of activations on the hidden layer -- of a named exemplar and a test object were similar.

The novel noun generalization task used with children is typically a forced choice task in which the child must choose between an object matching the named exemplar in shape and one matching in material. Accordingly, on each simulated test trial, we measured the similarity of the internal patterns of representation for two test objects –one matching the exemplar in shape and one matching the exemplar in material.

More specifically, on each test trial, we created a novel exemplar object by randomly generating an activation pattern along the shape and material dimensions. Then we combined the exemplar’s shape pattern with a novel randomly generated material pattern to create a novel shape-matching test object. A similarity measure of the exemplar and the shape match was computed using the Euclidean distance between the activation patterns in the Hidden Layer after the exemplar and its shape match had been presented.

Similarly, we generated a novel material-matching test object by combining the exemplar’s material pattern with a new randomly generated shape pattern and then computed the similarity between exemplar and material match. Finally, we used these similarity measures between the emergent patterns of activation on the hidden layer to calculate the probability of choosing the shape and the material matches using Luce’s Forced Choice Rule.

In this way, we trained 10 networks (with 10 different randomly generated initial connection weights) with the object and substance terms young children know. During training, we presented multiple instances of each trained noun until the network stably produced the right noun when presented an instance of each kind. We then tested each of these networks in the novel noun generalization task using 20 novel exemplars. Half of these exemplars were defined by patterns of activation representing solid things and half by patterns representing nonsolid things. If the statistical regularities in early child vocabulary are sufficient to create learning biases then the networks should present a shape bias when the exemplar is solid and a material bias when the exemplar is non-solid.

### Results

Figure 2 shows the networks’ performance in the novel noun generalization task. As is apparent, the connectionist networks prefer the shape match in the solid trials and the material match in the non-solid trials. This supports the idea that the statistical regularities in the lexicon are sufficient to

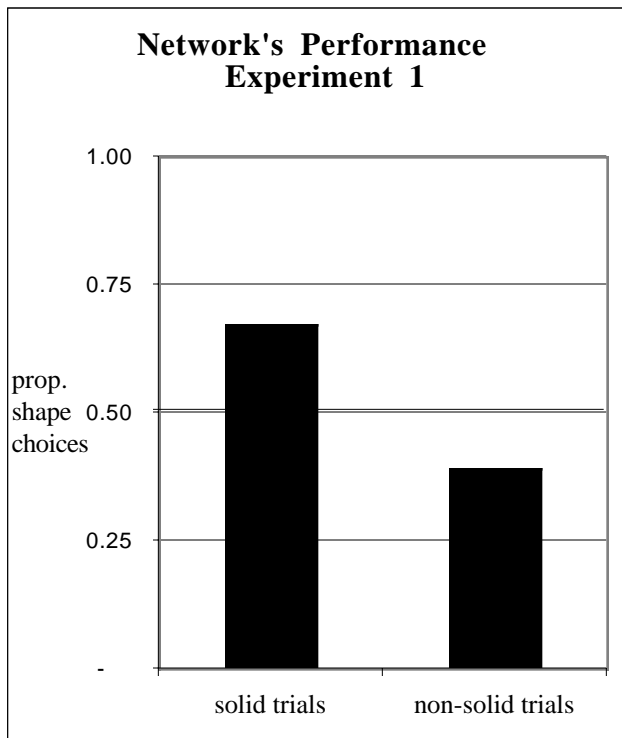


Figure 2. Network's performance in Experiment 1. The networks show shape and material biases comparable to those of children.

create word-learning biases in a statistical learner. If this is true, then other regularities present in the language should create their own "biases". One ubiquitous regularity that became obvious to us is that things that share a name share their solidity value. In other words, names do not refer to categories that span across ontological boundaries. This is true for all words in children's vocabulary except one – egg, which adults judged to have both solid and non-solid forms. If noun generalizations by the network are generalizations over the structures of already learned noun categories, then the network's generalizations of new names for novel things should adhere to this constraint. Given a solid exemplar, sameness in shape should not count if the test object is non-solid; given a non-solid exemplar, sameness in material should not count if that material is now solid. In Experiment 2 give this tests to the networks.

### Experiment 2

The goal of Experiment 2 is to test the network on the ontology bias. The network architecture and training procedure were the same as in Experiment 1. Ten networks were trained using the same testing procedure as in Experiment 1 except for the kinds of test objects used.

As in Experiment 1, on each test trial, we created a novel exemplar object by randomly generating an activation pattern along the shape and material dimensions and then created shape and material matches combining the exemplar's shape and material patterns with novel randomly generated material and shape patterns. Again, the networks were tested

on 20 novel exemplars; half of them defined as solid and half of them defined as non-solid. However, to make the ontology violating test, the shape match for solid exemplars was defined as non-solid and the material match for non-solid exemplars was defined as solid. So for the solid trials, we computed forced choice probability between a non-solid shape match and a solid material match, while in non-solid trials we compared a non-solid shape match with a solid shape match.

### Results

Figure 3 shows the proportion of shape choices predicted by the networks for solid exemplar trials and for non-solid exemplar trials. As predicted from the regularities in the training set, the networks chose the test item that matches the exemplar in solidity. That is, when the exemplar is solid the network prefers the solid test object, (even though it does not match in shape) and when the exemplar is non-solid the network prefers the non-solid test item (even though it does not match in material). Thus, the pattern of generalization observed in Experiment 1 (and typical in experimental tests of children) is now reversed: the networks exhibit a shape bias in non-solid trials and a material bias in solid trials. In Experiment 3 we look for this effect in children.

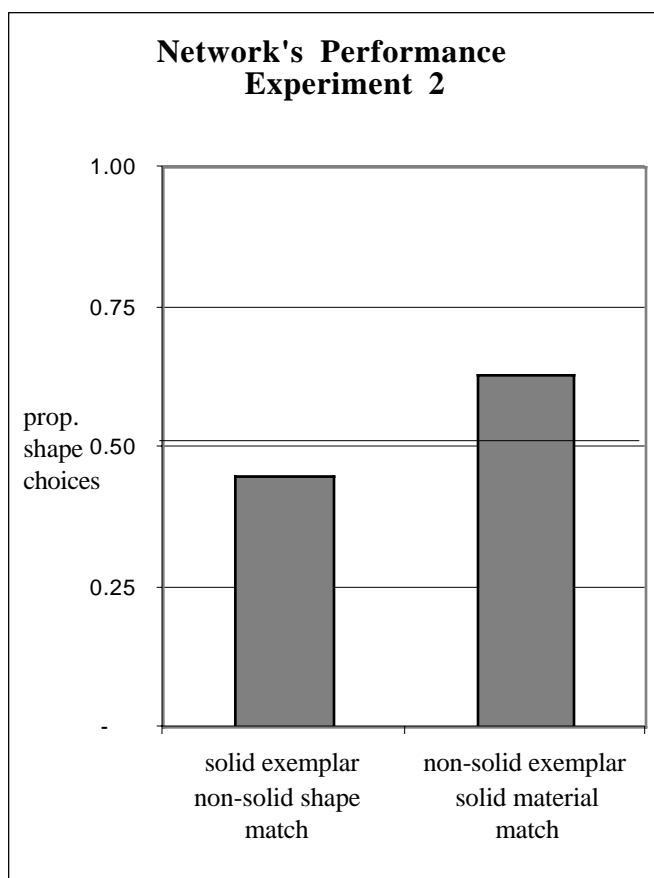


Figure 3. Network's performance in Experiment 2. The networks preferences are reversed when the shape match for the solid exemplar is made non-solid and the material match for the non-solid exemplars is made solid.

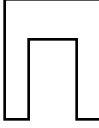



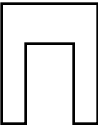
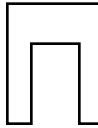

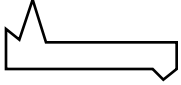
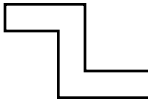



	Exemplar	Material Match	Shape Match	
			Ontology Violating	Traditional
TEEMA	 red sand paint	red sand paint   	 toothpaste with glitter shaving cream purple hair gel	 metallic blue clay burlap
WAZZLE	 blue cheese cloth	blue cheese cloth   	 toothpaste with glitter shaving cream purple hair gel	 pink felt with bumps purple foam green fur

Figure 4. Stimuli for Experiment 3A..

### Experiment 3

The goal of Experiment 3 is to test the prediction made by the network in Experiment 1. Given a solid object, will children refuse to generalize its name to an object of the same shape if the test object is not solid? Given a non-solid object, will children refuse to generalize its name to a material match if the test object is solid? Experiment 3A tests the first question using solid exemplars and Experiment 3B tests the second question using non-solid exemplars. Constructing stimuli for Experiment 3A (shape matches that differ in solidity) was easy; we can create the same shape out of hardened clay and shaving cream. Constructing stimuli for the second question (material matches that differ in solidity) required more creativity. What we did was use translucent gel and translucent hardened plastic for one set and off-white hand lotion and off-white hardened paint for the other. In both cases the material looked to be the same and was judged by adults to be the non-solid and hardened versions of the same material.

### Method

**Subjects** Twenty-four children between the ages of 30 and 36 months participated in this study. Half of them were ran-

domly assigned to Experiment 3A and half of them were assigned to Experiment 3B.

**Stimuli** The stimuli for Experiment 3A are shown in Figure 4. There were two exemplar objects. The exemplar for one set, the Teema, was a “U” shape covered with red sand-paint. The exemplar for the other set, the Wazzle, was an irregular “M” shape covered with blue cheese-cloth. For each exemplar there were three objects matching in material and two sets of items matching in shape. The Traditional set consisted of three solid objects that matched the exemplar in shape and differed in material (e.g. metallic clay, styrofoam covered with fur). The Ontology Violating set consisted of shape matches made out of non-solid materials (e.g. shaving cream, hair gel).

The stimuli for Experiment 3B are shown in Figure 5. There were two exemplar objects. The exemplar for one set, the Teema, was a “V” shape made out of translucent gel. The exemplar for the other set, the Wazzle, was an irregular “M” shape made out of hand lotion. For each exemplar there was a set of shape matches made out of three different non-solid substances. For the Teema, the shape matches were made out of wax, glitter and noxzema mixed with sand; for the Wazzle, the shape matches were made out of green sand, toothpaste with glitter and shaving cream. For each exemplar there were also two sets of “material” matches: a Traditional set and an Ontology Violating set. For the Teema the


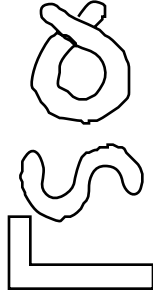
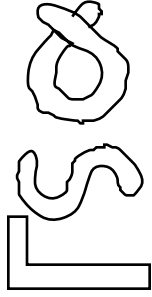


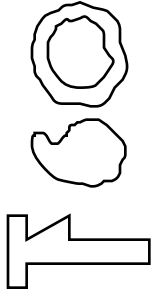
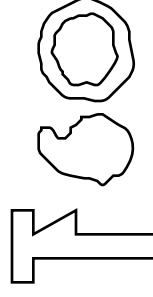

	Exemplar	Material Match		Shape Match
		Ontology Violating	Traditional	
TEEMA	 translucent gel	translucent plastic 	translucent gel 	 noxzema with sand glitter wax crystals
WAZZLE	 hand lotion	hardened paint 	hand lotion 	 toothpaste with glitter fine sand shaving cream

Figure 5. Stimuli for Experiment 3B.

Traditional set consisted of shapes made out of translucent hair gel and the Ontology Violating set consisted of shapes made out of translucent hard plastic. For the Wazzle the Traditional set consisted of shapes made out of off-white hand lotion and the Ontology Violating set consisted of shapes made out of off-white hardened fabric paint.

**Procedure** The procedure used was a forced choice task. The child were shown an exemplar (i.e., the Teema) and told its name (“this is the Teema”). The child was then presented with pairs of objects, a shape match and a material match, and asked “Can you show me the Teema?”. Each child was presented with the Traditional set of one exemplar and the Ontology Violating set of the other. Half of the children were assigned at random to judge the Traditional version of one exemplar and the Ontology Violating version of the other. The two exemplars were presented in separate blocks. Each shape-match/material-match pair was presented twice in random order for a total of 12 trials. The order of the sets was counterbalanced across subjects; the position of the choices was counterbalanced across trials.

### Results

Figure 6 shows the proportion of shape choices for the solid exemplar (Experiment 3A) and for the non-solid exemplar (Experiment 3B) Ontological Violating and Traditional sets respectively. In the Traditional sets, children’s performance replicates previous findings: they present a clear shape bias for the trials with solid exemplars (Experiment 3A) and

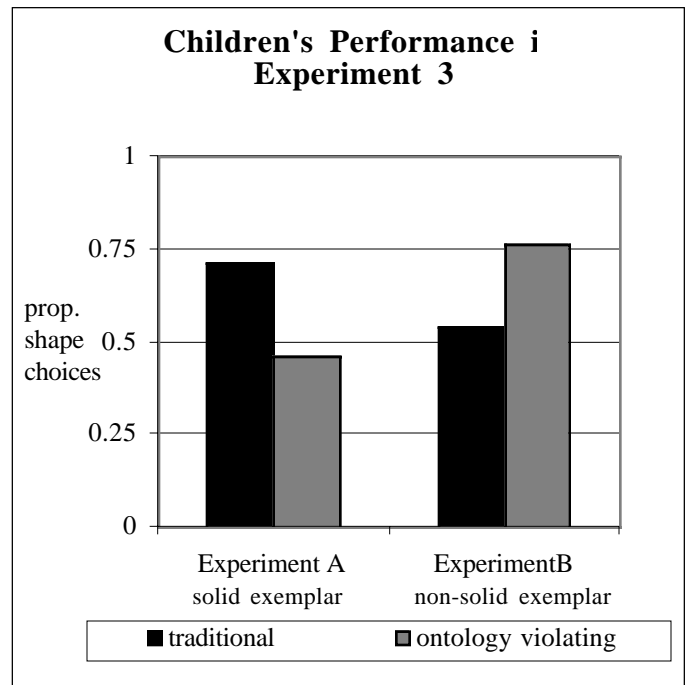


Figure 6. Results of Experiment 3.

show increased attention to the material of non-solid exemplars (Experiment 3B). In the Ontology Violating sets, as

the network simulations predicted, children's shape bias decreased to chance levels in solid trials and increased to above chance in the non-solid trials.

Although these results are consistent with a bias in children to extend category names only within ontological boundaries, there is an alternative explanation. Children's preference for the same-solidity item could be a result of the way the ontological violating choices alter the exemplar-test items' similarity. For example, in the case of the solid exemplar, the material match matches in both material and solidity, while the shape match now only matches in shape (and imperfectly at that, given the change of solidity). While we can't be sure of which explanation is the case in children, we know for a fact that it is more than just similarity for the networks.

## Conclusions

Learning a first language is a hard problem. However, the task appears less daunting when we consider that the kinds of words children know early present an organized structure. A smart learner could learn to exploit this structure to its advantage. In this paper we have shown that a simple statistical learner, with no other mechanisms than associative learning and generalization by similarity, will learn shape and material biases to match the systematic regularities present in its training set. We have also documented a new bias, one which could be taken as evidence of an underlying ontology, but that also makes sense in terms of the statistical regularities present in the language. This suggests that word-learning biases and constraints could be a product of learning. While the evidence presented here is consistent with this account, it does not provide conclusive proof; the regularities found in children's vocabularies could be a product of pre-existing biases. However, the fact that we have demonstrated the computational plausibility of the learning account and simple parsimony suggest that this is not the case.

## References

- Landau, B., Smith, L.B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3, 299—321.
- Markman, E. M. (1989). *Categorization and naming in children.*, Cambridge, MA: MIT Press.
- Movellan, J. R. (1990). Contrastive Hebbian Learning in the Continuous Hopfield Model. Proceedings of the 1990 Connectionist Models Summer School (pp. 10—17). San Mateo, CA: Morgan Kaufmann.
- Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: do ontology, category structure and syntax correspond? *Cognition*, 73, 1--33.
- Soja, N.N., Carey, S., & Spelke, E.S. (1991). Ontological categories guide young children's inductions of word meaning: object terms and substance terms. *Cognition*, 38, 179—211.

# An Exemplar Model of Classification in Single and Combined Categories.

Fintan Costello ([fintan@compapp.dcu.ie](mailto:fintan@compapp.dcu.ie))

School of Computer Applications,  
Dublin City University,  
Glasnevin, Dublin 9, Ireland.

## Abstract

This paper describes an exemplar-based model of people's classification and typicality judgements in both single and combined categories. This model, called the diagnostic evidence model, explains the observed family resemblance structure of single categories; the productive nature of category combination; the observed overextension of typicality judgments in some combined categories; and the situations in which that overextension occurs. The model also gives a close fit to quantitative results from a representative single-category classification data-set.

Models of categorisation need to explain two basic aspects of human cognition: our ability to classify items as members of single categories such as *fish* or *cat*, and our ability to classify items as members of combinations of categories such as *wild cat* or *pet fish*. A successful model should account for the graded structure of classification in single categories: the fact that people's judgements of membership typicality for items in categories are proportional to the items' family resemblance to members of those categories (Rosch, 1978; Rosch & Mervis, 1975). A successful model should also account for the productivity of category combination: the fact that people are able to understand and judge membership in new combinations of categories, even if no already-existing examples of those combinations are known. This combinatorial productivity is important because it underlies our ability to think new thoughts and understand new expressions. In many currently popular models of categorisation (e.g. the context theory; Medin & Schaffer, 1978), an item's membership in a category is proportional to its similarity to the stored exemplars of that category. While this approach gives a good account for the graded structure of single categories, it has difficulty explaining the productivity of category combination, which involves classification in combinations for which no stored exemplars are available (Rips, 1995). This paper describes an exemplar-based model of classification in single and combined categories which explains the family resemblance structure of single categories, the productivity of category combination, and other specific results in both domains. The model, called the diagnostic evidence model, extends a successful earlier theory (Costello & Keane, 1997, in-pressA, in-pressB).

The first part of the paper presents the diagnostic evidence model of categorisation in single and combined categories, and gives its account for family resemblance and productivity in combination. The second part demonstrates the model by showing how it explains the observed overextension of typicality in some combined categories.

Overextension occurs when people rate an item as a poor member of both constituents of a combination, but as a good member of the combination as a whole; for example, when goldfish are rated as untypical members of the categories *pet* and *fish*, but as typical members of the combination *pet fish* (Hampton, 1988). Overextension has posed problems for a number of theories of category combination. The diagnostic evidence model accounts for results on overextension, and explains why overextension occurs in some combinations but not in others. The third part of the paper demonstrates this model further by showing how it gives a good fit to quantitative results from a representative classification data-set (Nosofsky, Palmeri, & McKinley, 1994); a fit as close as that given by exemplar-similarity models such as the context theory.

## The diagnostic evidence model

The diagnostic evidence model extends an earlier theory of the interpretation of noun-noun combined phrases, called the constraint theory (Costello & Keane, 1997, in-press-A). That theory set out to explain the diversity of interpretations which people produce for noun-noun combinations: the fact that people sometimes interpret combinations by forming conjunctions between the combining categories (as in the interpretation "*pet bird*: a parrot or some other bird which is also a pet"), sometimes by asserting relations between the categories (as in "*jungle bird*: a bird that lives in jungles"), and sometimes by transferring properties from one concept to the other (as in "*skunk bird*: a bird that smells bad"). Constraint theory explains this diversity by describing a combination process that forms mental representations satisfying three constraints of diagnosticity, plausibility and informativeness. Each interpretation type represents a different way of satisfying these constraints. The theory has been tested in a computer program which simulates the interpretation of noun-noun combinations, producing each interpretation type and generating results that agreed with people's interpretations of those combinations (Costello & Keane, in-press-A). Further, Costello & Keane (in-press-B) have provided direct experimental evidence for diagnosticity's role in the formation of combined categories.

Where the Constraint theory gave a qualitative account of noun-noun interpretation, the diagnostic evidence model aims to give a quantitative account of people's classification of items in single and combined categories. The model focuses on the diagnosticity constraint. The model assumes that people represent categories by storing sets of category exemplars in memory. From these sets, diagnostic attributes for categories are computed: these attributes serve to identify category members. An item's membership

typicality in a single or combined category is a function of the diagnosticity of its attributes for that category or for the constituent categories of that combination. An item has high membership typicality in a category if it has attributes that are highly diagnostic for that category. An item has high typicality in a combination if it has some attributes highly diagnostic for one constituent of the combination, and other attributes highly diagnostic for the other. Two novelties in this model are its method for computing attribute diagnosticity, and its logic for combining the diagnosticity of multiple attributes to compute membership in single or combined categories. I describe these below.

### Attribute Diagnosticity

Diagnostic attributes are attributes which occur frequently in stored instances of a category, but rarely in that category’s contrast set (the set of stored instances which are not members of the category). These attributes serve to identify members of a category: a new item possessing a attribute which is highly diagnostic for a given category is likely to be a member of that category. The diagnosticity of attribute  $x$  for category  $C$  is defined in Equation 1. Let  $K$  be the contrast set for  $C$ . Let  $j_x$  be 1 if instance  $j$  possesses attribute  $x$ , and 0 otherwise.  $D(x/C/K)$ , the diagnosticity of  $x$  for  $C$  relative to  $K$ , is equal to the number of instances in  $C$  that possess  $x$ , divided by the total number of instances in  $C$  plus the number of instances in  $K$  that possess  $x$ :

$$D(x/C/K) = \frac{\sum_{j \in C} j_x}{|C| + \sum_{j \in K} j_x} \quad (1)$$

If the attribute  $x$  occurs in all instances in category  $C$ , but no instances in  $C$ ’s contrast set, then  $x$  is fully diagnostic for  $C$  ( $D(x/C/K) = 1$ ). Such an attribute is a perfect guide to membership of  $C$ : every instance possessing  $x$  is a member of  $C$ , every instance not possessing  $x$  is not a member. An attribute which does not occur in all members of  $C$ , or which occurs in some members of  $C$ ’s contrast set, will be less diagnostic for the category. Such an attribute will be a poorer guide to membership of  $C$ : not every instance possessing  $x$  will be a member of  $C$ , not every instance not possessing  $x$  will be a non-member.

An important novelty in the diagnostic evidence model is that the diagnosticity of an attribute for a category can change depending on whether the category occurs singly or as part of a category combination. This change in diagnosticity arises because the contrast set used for computing diagnosticity is different in single and combined categories. For single categories, the contrast set consists of all instances which are not members of the category in question. For combined categories, however, the contrast set consists of instances which are not members of any of the constituents of the combination. The contrast set for a combination is thus a subset of the contrast sets for the single categories which make it up. This change in contrast set means that some attributes which are not diagnostic for a category when it occurs singly (because they occur frequently in that category’s contrast set), will be diagnostic

**Table 1.** An illustrative array of exemplars

Exemplar category labels		Attributes			
		FOUND	KEPT-IN	COLOR	HAS-PART
1	lobster	sea	-----	pink	claws
2	lobster	aquarium tank		pink	claws
3	fish goldfish	house	tank	gold	scales
4	fish guppy	house	tank	silver	skin
5	fish salmon	sea	-----	silver	scales
6	fish shark	sea	-----	silver	skin
7	pet dog spaniel	house	basket	brown	tail
8	pet dog doberman	house	kennel	black	tail
9	pet dog bulldog	house	basket	brown	-----
10	pet terrapin	house	tank	green	skin

for that category when it occurs in a combination (if they occur only rarely in that combination’s contrast set).

The computation of attribute diagnosticity can be demonstrated using an illustrative set of stored exemplars of categories such as *pet*, *fish*, *dog* and *lobster*, shown in Table 1. These exemplars are described in attribute-value pairs on four dimensions: FOUND, KEPT-IN, COLOUR, and HAS-PART. Consider the diagnosticity of the attribute FOUND:HOUSE for the single category *fish*, which has 4 stored exemplars (exemplars 3, 4, 5, 6).  $K_{fish}$ , the contrast set for the category *fish*, contains exemplars 1, 2, 7, 8, 9, 10. FOUND:HOUSE occurs in 2 of the 4 *fish* exemplars in Table 1, and in 4 exemplars in the contrast set  $K_{fish}$ . The diagnosticity of FOUND:HOUSE for the *fish* is thus

$$D(\text{FOUND} : \text{HOUSE}/\text{fish}/K_{fish}) = \frac{2}{4 + 4} = 0.25 \quad (2)$$

This attribute has a low diagnosticity for the single category *fish*: FOUND:HOUSE does not identify members of category *fish* well. In the context of the combination *pet fish*, however, the attribute has a higher degree of diagnosticity for *fish*.  $K_{petfish}$ , the contrast set for the combination *pet fish*, consists of exemplars that are members neither of *pet* nor of *fish* (exemplars 1 and 2). FOUND:HOUSE does not occur in any exemplars in the contrast set  $K_{petfish}$ . The diagnosticity of FOUND:HOUSE for *fish* relative to the contrast set  $K_{petfish}$  is thus

$$D(\text{FOUND} : \text{HOUSE}/\text{fish}/K_{petfish}) = \frac{2}{4 + 0} = 0.5 \quad (3)$$

The attribute thus gives a greater degree of diagnostic evidence for membership in the *fish* constituent of *pet fish*; in other words, the attribute FOUND:HOUSE is more diagnostic in identifying an item as a pet fish than it is identifying an item as a fish. This effect of contrast set on diagnosticity is central to the diagnostic evidence model’s account for overextension in combined categories, and is discussed in the section on overextension, below.

### A Continuous-valued Logic for Evidence

Diagnostic attributes, then, give evidence for an instance’s classification in a category. Instances usually

contain a number of different attributes, however, which may be more or less diagnostic for the category in question, or diagnostic for other categories. How is the diagnostic evidence from an item's attributes combined to produce an overall measure of evidence for category membership? The diagnostic evidence model uses a continuous-valued logic to combine diagnostic evidence from multiple attributes. This logic assumes continuous variables with values between 0 and 1, and uses the following logical operations:

$$NOT A = 1 - A \quad (4)$$

$$A AND B = AB \quad (5)$$

$$A OR B = 1 - (1 - A)(1 - B) \quad (6)$$

These equations derive from standard probability theory, and can be justified by considering the operations *AND*, *OR*, and *NOT* for samples of independently distributed variables. Suppose variables *A* and *B* have 0.75 and 0.5 probability of being true, respectively. Then the probability of *NOT A* being true is 0.25 ( $1 - 0.75$ ). The probability of *A AND B* being true is 0.375 ( $0.75 \times 0.5$ ): of the 75% of cases in which *A* is true, 50% of those are cases in which *B* is also true. Finally, the probability of *A OR B* being true is 0.875 ( $1 - (1 - 0.75) \times (1 - 0.5)$ ): of the 25% of cases in which *A* is false, 50% of those are cases in which *B* is also false; thus *A OR B* is true in 87.5% of cases. Similar (though often more complex) logics have been used in various areas (e.g. in models of evidence-based reasoning; Shafer, 1976). The current model is unique in using this approach to compute the contribution which different attributes make in people's classification of items in single or combined categories.

### Combining Diagnosticity of Multiple Attributes.

To combine the diagnostic evidence from multiple attributes for membership in a category, the diagnostic evidence model uses the equation for *OR*. An instance *i* with a set of attributes  $x_1, x_2, x_3$ , will be a member of category *C* if  $x_1$  or  $x_2$  or  $x_3$  serves to identify the instance as a member of *C* (if  $x_1 OR x_2 OR x_3$  is diagnostic for *C*). This is formalised in Equation 7, which has the form of the equation for *OR* (Equation 6, above). Let *A* be the set of attributes of instance *i* and  $D(x|C/K)$  be the diagnosticity of attribute *x* for *C*. Then  $E(i|C/K)$ , the overall evidence for classifying instance *i* as a member of *C*, is

$$E(i|C|K) = 1 - \prod_{x \in A} (1 - D(x|C|K)) \quad (7)$$

This equation accounts for people's classification in both strictly defined and "family resemblance" categories. If an attribute *x* strictly defines a category *C* (occurs in all instances of *C* and never occurs outside *C*), then *x* is perfectly diagnostic of *C* ( $D(x|C/K) = 1$ ). If any item *i* possesses attribute *x*, then by Equation 7  $E(i|C/K)$  will be 1, and the instance *i* will definitely be a member of *C*. In categories which have no single perfectly diagnostic attribute but rather have a range of attributes of medium diagnosticity, Equation 7 combines evidence from different attributes in computing evidence for category membership: the more diagnostic attributes the instance has, the higher its degree of membership will be. In other words, the more of a family resemblance an instance has to the members of a

category, the higher its membership typicality will be. This relationship between diagnostic attributes and membership has specific support in Rosch & Mervis' (1975) finding that people's judgements of an instance's typicality in a single category rose reliably with the number of diagnostic attributes for that category which the instance possessed.

The combination of diagnostic evidence can be illustrated using the exemplars in Table 1. For example, consider the evidence for exemplar 5 (*salmon*) as a member of the category *fish*. This exemplar has attributes LIVES:SEA, COLOUR:SILVER, and HAS-PART:SCALES. The diagnosticities of these attributes for *fish* are relatively high (0.4, 0.75 and 0.5 respectively, as computed from Equation 1). From Equation 7, these diagnostic evidence values are combined to obtain an overall measure of evidence for exemplar *salmon*'s typicality in category *fish* as follows:

$$E(\text{salmon} | \text{fish} | K_{\text{fish}}) = 1 - (1 - 0.4)(1 - 0.75)(1 - 0.5) = 0.925 \quad (8)$$

The exemplar *salmon* has good evidence for membership in the category *fish* because it possesses highly diagnostic attributes for that category: in other words, *salmon* is a highly typical *fish*. Other exemplars have less diagnostic attributes for the category *fish*, and thus have lesser degrees of evidence and are less typical category members. For example, the exemplar *shark* has the less diagnostic attribute HAS-PART:SKIN and is a less typical member of the category *fish* ( $E(\text{shark} | \text{fish} | K_{\text{fish}}) = 0.91$ , computed as above); the exemplar *goldfish* has two less diagnostic attributes LIVES:HOUSE and COLOUR:GOLD and is less typical again ( $E(\text{goldfish} | \text{fish} | K_{\text{fish}}) = 0.813$ ); the exemplar *spaniel* has no diagnostic attributes and is a poor member of the category ( $E(\text{spaniel} | \text{fish} | K_{\text{fish}}) = 0.25$ ).

### Diagnostic Evidence for Combined Categories

The diagnostic evidence model of classification, then, is consistent with observed patterns of typicality in single categories. The model extends easily to account for classification in category combinations: an item will be a member of a combined category if it gives diagnostic evidence for membership in each constituent in that combination. In computing an item's membership in a combined category, the model uses the continuous-valued *AND* described above (Equation 5) to combine the item's evidence for membership in each constituent of the combination. An instance *i* will be classified as a member of a combined category  $C_1 \dots C_N$  if it gives evidence for membership in  $C_1$  AND evidence for membership in  $C_2$  AND evidence for membership in  $C_3$  and so on. More formally,  $E(i|C_1 \dots C_N | K_{1 \dots N})$ , the evidence for classifying *i* as a member of combination  $C_1 \dots C_N$ , is

$$E(i|C_1 \dots C_N | K_{1 \dots N}) = \prod_{n=1}^N E(i|C_n | K_{1 \dots N}) \quad (9)$$

where the contrast set  $K_{1 \dots N}$  is the set of instances not in any of the categories  $C_1 \dots C_N$ . Note that an instance *i* will give evidence for membership in each constituent of a combination if it has some attributes diagnostic for each constituent: some attributes diagnostic for one constituent, other attributes diagnostic for others.



Because the diagnostic evidence model computes evidence for membership in a combination by combining evidence for membership in its constituent categories, it can explain people’s ability to classify items in new combinations, even if they have no stored exemplars of those combinations. An item is classified as a member of a combination, even one with no stored exemplars, if the item has diagnostic attributes for each constituent category in the combination. For example, in Table 1, there are no stored exemplars of the combination *pet lobster*. However, an item could be classified as a good member of the combination *pet lobster* if it possessed the attribute HAS-PART:CLAWS (perfectly diagnostic for *lobster* in Table 1) and the attribute FOUND:HOUSE (highly diagnostic for *pet*).

In accounting in this way for the productivity of category combination, the model goes beyond theories such as the context theory, in which classification is based on similarity to stored exemplars of a category. Such theories cannot account for classification in new combinations for which there are no stored exemplars. For example, in an exemplar-similarity based model, people would judge membership in *pet lobster* by computing an item’s similarity to stored exemplars of that combination (by comparing the item to previously seen examples of pet lobsters). Since *pet lobster* has no stored exemplars, this computation would be meaningless (see Rips, 1995).

### Accounting for Overextension

Various studies have examined overextension of classification in combined categories. Overextension occurs when people rate an item as a poor member of both constituents of a combination, but as a good member of the combination as a whole. For example, people might rate goldfish as typical members of the combination *pet fish*, but as untypical members of the single categories *pet* and *fish*. Hampton (1988) found that overextension was more likely for some combinations than for others: the lower the degree of overlap between combining categories (the fewer exemplars the categories had in common) the more likely the combinations were to be overextended. For example, the constituents of *pet fish* have low overlap (many fish are not pets; many pets are not fish), and that combination was often overextended. By contrast, combinations of categories with many common members were usually not overextended. For example, the constituents of *pet dog* have high overlap (most dogs are also pets), and that combination was usually not overextended.

Overextension poses a challenge for theories of category combination (Osherson & Smith, 1981). In the diagnostic evidence model, overextension arises because of changes in attribute diagnosticity: because some attributes may have low diagnosticity for a category when it occurs singly, but high diagnosticity for that category when it occurs as part of a combination. As we saw earlier, the attribute FOUND:HOUSE was less diagnostic for the single category *fish*, but was more diagnostic for the category in the context of the combination *pet fish* (because the attribute occurred often in the contrast set for the category *fish*, but not in the contrast set for the combination *pet fish*). This change in

**Table 2.** Overextension of exemplar *goldfish* in *pet fish*

Evidence for membership in	Exemplar	Attribute Diagnosticity			
		FOUND	KEPT-IN	COLOR	HAS-PART
	<i>goldfish</i> :	house	tank	golden	scales
<i>pet</i> singly :	0.714	0.67	0.14	0	0
<i>fish</i> singly:	0.813	0.25	0.33	0.25	0.5
<i>pet fish</i> :	0.89				
<i>pet</i>	1	1.0	0.2	0	0
<i>fish</i>	0.89	0.5	0.4	0.25	0.5

diagnosticity means that an item with that attribute could give good evidence for membership in the combination *pet fish* (and therefore high typicality in that combination), but poor evidence for membership in the single categories *fish* and *pet* (low typicality in those single categories).

Table 2 illustrates this account of overextension, showing computed evidence for the exemplar *goldfish* as a member of the single categories *pet* and *fish*, and the combination *pet fish*. Note that *goldfish* gives higher evidence for membership in *pet fish* (0.89) than in either *pet* (0.714) or *fish* singly (0.813). *Goldfish* would thus be judged a highly typical *pet fish* but a less typical *pet* or *fish*. This is because the exemplar’s attributes have higher diagnosticity for the combination than for the single categories. For example, FOUND:HOUSE has a diagnosticity of 0.67 for the single category *pet* and of 0.25 for the single category *fish*. In the context of *pet fish*, however, FOUND:HOUSE has a higher diagnosticity both for the constituent *pet* (1.0) and the constituent *fish* (0.5). (In Table 2, evidence for membership in the single categories is computed by combining attribute diagnosticity as in Equation 7. Evidence for membership in the combination is obtained by computing evidence for membership in each constituent category as in Equation 7, and combining that evidence as in Equation 9).

In this account, overextension arises from a difference between the contrast sets for single categories and those for a combination, which leads to a difference in diagnostic evidence for membership in the single categories and the combination. If there is little difference between these contrast sets, overextension won’t occur. Table 3 illustrates this for the combination *pet dog*. *Pet dog* is not overextended: the exemplar *spaniel* gives more evidence for membership in the single categories *pet* (0.96) and *dog* (0.98) than in the combination *pet dog* (0.95). Because the categories *pet* and *dog* have a high overlap (in Table 1, all

**Table 3.** Non-overextension of exemplar *spaniel* in *pet dog*

Evidence for membership in	Exemplar	Attribute Diagnosticity			
		FOUND	KEPT-IN	COLOR	HAS-PART
	<i>spaniel</i> :	house	basket	brown	tail
<i>pet</i> singly :	0.96	0.67	0.5	0.5	0.5
<i>dog</i> singly:	0.98	0.5	0.67	0.67	0.67
<i>pet dog</i> :	0.95				
<i>pet</i>	0.96	0.67	0.5	0.5	0.5
<i>dog</i>	0.98	0.6	0.67	0.67	0.67

**Table 4.** Predicted and observed probability of classification of exemplars in Nosofsky, et al., (1994) Experiment 1.

Exemplar labels	Exemplars	Diagnostic evidence ( $C(i/A), W = 8$ )	Predicted classification probability (linear transform of $C(i/A)$ )	Classification probability observed in Experiment
A1	1 1 1 2	0.69	0.77	0.77
A2	1 2 1 2	0.65	0.74	0.78
A3	1 2 1 1	0.75	0.83	0.83
A4	1 1 2 1	0.52	0.6	0.64
A5	2 1 1 1	0.52	0.6	0.61
B1	1 1 2 2	0.37	0.46	0.39
B2	2 1 1 2	0.37	0.46	0.41
B3	2 2 2 1	0.13	0.23	0.21
B4	2 2 2 2	0.07	0.17	0.15
T1	1 2 2 1	0.45	0.54	0.56
T2	1 2 2 2	0.31	0.4	0.41
T3	1 1 1 1	0.78	0.86	0.82
T4	2 2 1 2	0.31	0.4	0.4
T5	2 1 2 1	0.16	0.26	0.32
T6	2 2 1 1	0.45	0.54	0.53
T7	2 1 2 2	0.07	0.17	0.2

dogs are also pets) there is little difference between the contrast sets for the single categories *pet* and *dog* and the contrast set for the combination *pet dog*. There is thus little difference in the diagnosticity of attributes for the single categories and for the combination; hence, there is no overextension.

This account explains Hampton’s (1988) finding that overextension is rare for combinations whose constituent categories have a high degree of overlap. The greater the overlap between the constituent categories of a combination, the less of a difference there is between the contrast sets for those categories occurring singly, and the contrast set for that combination. The less of a difference between contrast sets, the less of a difference between diagnostic evidence for membership in the single categories and in the combination; the less chance of overextension.

### Fitting Classification Data-sets

As described above, the diagnostic evidence model can explain various results in natural-language categorisation and category combination. In this section I fit the model to results obtained in a study of classification in artificial laboratory-learned categories: Nosofsky, Palmeri, & McKinley’s (1994) replication of Medin & Schaffer’s (1978) study. In Nosofsky, Palmeri, & McKinley’s experiment, participants learned to classify drawings of rocketships as coming from planet A (category A) or planet B (category B). The rocketships varied on four dimensions (shape of tail, wings, nose, and porthole) each with two values, represented by 1 and 2. Rockets from planet A had values of 1 on most dimensions, while rockets from planet B had values of 2. An abstract representation of this category structure is shown in Table 2. In an initial training phase, participants learned 9 training items: *A1...A5* from category A and *B1...B4* from category B. In the test phase

participants categorised the 9 training items and 7 new test items *T1...T7*. Test item T3 was the prototype for category A (having a value 1 on all dimensions).

In this experiment participants classified items into one of only two possible categories (A or B). Classification in this two-category task is different from classification in natural-language categories: when only two categories are available, an item’s membership in a category depends both on evidence that the item is a member of the category, and on evidence that the item is not a member of the other category. In applying the diagnostic evidence model to this two-category task, the model was extended (using the continuous-valued logic described above) to take account of both sources of evidence: an item was classified in category A if it gave evidence for membership in A, OR did NOT give evidence for membership in B. Formally,  $C(i/A)$ , the classification score for  $i$  as a member of category A, is

$$C(i/A) = E(i/A/K_A) \text{ OR } (\text{NOT } E(i/B/K_B)) \quad (10)$$

$$= 1 - (1 - E(i/A/K_A))^W (1 - (1 - E(i/B/K_B)))$$

where  $E(i/A/K_A)$  and  $E(i/B/K_B)$  give measures of evidence for membership in A and B respectively (computed according to Equation 7), and where parameter  $W$  represents the relative importance of evidence for membership in A versus evidence for membership in B in classification.

The diagnostic evidence model was applied to the data-set using only the training stimuli (exemplars *A1...A5* and *B1...B4*). These training exemplars were used to compute the diagnosticity of the values 1 and 2 on each dimension for the categories A and B. These diagnosticities were then used to compute the diagnostic evidence score  $C(i/A)$  for both training and test exemplars as members of category A. These scores are shown in the “diagnostic evidence” column in Table 4. These scores are those for the value of  $W$  for which the correlation between predicted and observed

classification was highest ( $W = 8$ ;  $r = .99$ ,  $\%var = 98\%$ ). The model's predicted classification probabilities (shown in the next column in Table 4) were obtained by a linear transformation of the diagnostic evidence scores, mapping the mean diagnostic evidence score onto the mean observed classification probability, and the standard deviation of the diagnostic evidence score onto the standard deviation of observed classification probabilities. (This transformation introduces no extra degrees of freedom into the model's fit to the data; it simply allows direct comparison between computed evidence for classification and the classification probabilities observed in the experiment). The diagnostic evidence model's computed classification scores for items closely follow people's classifications of those items, as comparison of the predicted and observed classification probability columns in Table 4 shows. The model accounts for the qualitative finding that the test exemplar T3 (the prototype for category A) gets a higher classification probability than all other test exemplars. The percentage of variance explained by the diagnostic evidence model (98%) is in the same range as that produced by other models (the context model explains 96% of variance in these results; the Rulx model explains 98%; see Nosofsky, Palmeri, & McKinley, 1994). However, those models used four free parameters to fit the data (varying the selective attention paid to the 4 dimensions on which exemplars were described), as opposed to the single parameter used by the diagnostic evidence model.

### Conclusions and Future work

The diagnostic evidence model of classification described here goes beyond other theories of classification in giving an account for both single and combined categories. The model explains the family resemblance structure of single categories, the productivity of category combination, and the occurrence of overextension in some combined categories. That the model is exemplar-based is significant: a number of results have shown that exemplars are important both for simple and combined categories (e.g. Gray & Smith, 1995). Some argue that exemplar-based models cannot account for the productivity of combination (Rips, 1995); the current model provides evidence against this argument. The model fits a representative classification data set as closely as Medin & Schaffer's (1978) context theory, while using fewer free parameters.

There are, however, various classification results which the diagnostic evidence model cannot currently explain. Because the model does not provide a mechanism for learning, it cannot address the role of learned attribute correlations in classification. A number of studies (e.g. Medin, Altom, Edelson, & Freko, 1982) show that people learn to associate correlated pairs of attributes with categories, and to use those correlated attributes in classification. The diagnostic evidence model as it currently stands cannot account for this result because it treats all attributes independently. Extending the model with a learning mechanism which can recognise attribute correlations and use those correlations to form new "composite" attributes may allow the model to account for

the role of correlation in classification. In an initial test of this approach, in which composite attributes were hard-coded into the representation used, the diagnostic evidence model was able to give a good fit to Medin et al.'s results. In future work I aim to develop the diagnostic evidence model in this direction.

### References

- Costello, F. J., & Keane, M. T. (in press A). Efficient creativity: Constraint guided conceptual combination. *Cognitive Science*.
- Costello, F. J., & Keane, M. T. (in press B). Testing two theories of conceptual combination: Alignment versus diagnosticity in the comprehension and production of combined concepts. *Journal of Experimental Psychology: Learning, Memory & Cognition*.
- Costello, F. J., & Keane, M. T. (1997). Polysemy in conceptual combination: Testing the constraint theory of combination. In *Nineteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Gray, K. C. & Smith, E. E. (1995). The role of instance retrieval in understanding complex concepts. *Memory & Cognition*, 23(6), 665-674.
- Hampton, J. A. (1988). Overextension of conjunctive concepts: Evidence for a unitary model of concept typicality and class inclusion. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15, 55-71.
- Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 8, 37-50.
- Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3), 207-238.
- Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, 101(1), 53-79.
- Osherson, D. N. & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, 9, 35-58.
- Rips, L. J. (1995). The current status of research on concept combination. *Mind & Language*, 10(1/2), 72-104.
- Rosch, E. & Mervis, C. D. (1975). Family resemblance studies in the internal structure of categories. *Cognitive Psychology*, 7, 573-605.
- Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.) *Cognition and Categorization*. Hillsdale, NJ: Erlbaum.
- Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

# A Six-Unit Network is All You Need to Discover Happiness

Matthew N. Dailey   Garrison W. Cottrell  
{mdailey, gary}@cs.ucsd.edu  
UCSD Computer Science and Engineering  
9500 Gilman Dr., La Jolla, CA 92093-0114 USA

Ralph Adolphs  
ralph-adolphs@uiowa.edu  
University of Iowa Department of Neurology  
220 Hawkins Dr., Iowa City, IA 52242 USA

## Abstract

In this paper, we build upon previous results to show that our facial expression recognition system, an extremely simple neural network containing six units, trained by backpropagation, is a surprisingly good computational model that obtains a *natural* fit to human data from experiments that utilize a forced-choice classification paradigm. The model begins by computing a biologically plausible representation of its input, which is a static image of an actor portraying a prototypical expression of either Happiness, Sadness, Fear, Anger, Surprise, Disgust, or Neutrality. This representation of the input is fed to a single-layer neural network containing six units, one for each non-neutral facial expression. Once trained, the network's response to face stimuli can be subjected to a variety of "cognitive" measures and compared to human performance in analogous tasks. In some cases, the fit is even better than one might expect from an impoverished network that has no knowledge of culture or social interaction. The results provide insights into some of the perceptual mechanisms that may underlie human social behavior, and we suggest that the system is a good model for one of the ways in which the brain utilizes information in the early visual system to help guide high-level decisions.

## Introduction

In this paper, we report on recent progress in understanding human facial expression perception via computational modeling. Our research has resulted in a facial expression recognition system that is capable of discriminating prototypical displays of Happiness, Sadness, Fear, Anger, Surprise, and Disgust at roughly the level of an untrained human. We propose that the system provides a good model of the perceptual mechanisms and decision making processes involved in a human's ability to perform forced-choice identification of the same facial expressions. The present series of experiments provides significant evidence for this claim.

One of the ongoing debates in the psychological literature on emotion centers on the structure of emotion space. On one view, there is a set of discrete basic emotions that are fundamentally different in terms of physiology, means of appraisal, typical behavioral response, etc. (Ekman, 1999). Facial expressions, according to this categorical view, are universal signals of these basic emotions. Another prominent view is that emotion concepts are best thought of as prototypes in a continuous, low-dimensional space of possible emotional states, and that facial expressions are mere clues that allow an observer to locate an approximate region in this space (e.g. Russell, 1980; Carroll and Russell, 1996).

One type of evidence sometimes taken as support for categorical theories of emotion involves experiments that

show "categorical perception" of facial expressions (Etcoff and Magee, 1992; Young et al., 1997). Categorical perception is a discontinuity characterized by sharp perceptual category boundaries and better discrimination near those boundaries, as in the bands of color in a rainbow. But as research in the classification literature has shown (e.g. Ellison and Massaro, 1997), seemingly categorical effects naturally arise when an observer is asked to employ a decision criterion based on continuous information. Neural networks also possess this dual nature; many networks trained at classification tasks map continuous input features into a continuous output space, but when we apply a decision criterion (such as "choose the biggest output") we may obtain the *appearance* of sharp category boundaries and high discrimination near those boundaries, as in categorical perception.

Our model, which combines a biologically plausible input representation with a simple form of categorization (a six-unit softmax neural network), is able to account for several types of data from human forced-choice expression recognition experiments. Though we would not actually propose a localist representation of the facial expression category decision (we of course imagine a more distributed representation), the evidence leads us to propose 1) that the model's input representation bears a close relationship to the representation employed by the human visual system for the expression recognition task, and 2) that a dual continuous/categorical model, in which a continuous representation of facial expressions coexists with a discrete decision process (either of which could be tapped by appropriate tasks), may be a more appropriate way to frame human facial expression recognition than either a strictly categorical or strictly continuous model.

## The Expression Classification Model

For an overview of our computational model, refer to Figure 1. The system takes a grayscale image as input, computes responses to a lattice of localized, oriented spatial filters (Gabor filters) and reduces the resulting high dimensional input by unsupervised dimensionality reduction (Principal Components Analysis). The resulting low-dimensional representation is then fed to a single-layer neural network with six softmax units (whose sum is constrained to be 1.0), each corresponding to one expression category. We now describe each of the components of the model in more detail.

## The Training Set: Pictures of Facial Affect

The model's training set is Ekman and Friesen's Pictures of Facial Affect (POFA, 1976). This database is a good

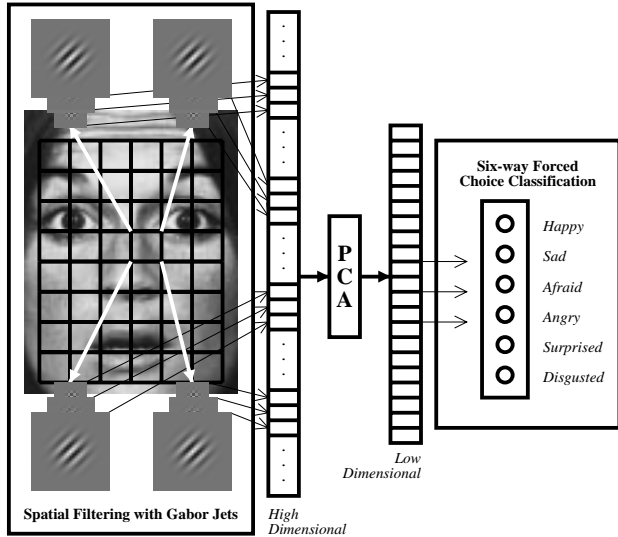


Figure 1: Facial Expression Classification Model.

training set because the face images are reliably identified as expressing the given emotion by human subjects (at least 70% agreement), and the images are commonly used in psychological experiments. We digitized the 110 POFA slides by scanning them at 520x800 pixels, performing a histogram equalization, aligning the eyes and mouths to the same location in every image by a linear transformation, and cropping off most of the background. The result is a set of 110 240x320 grayscale images of 14 actors portraying prototypical expressions of six basic emotions and neutral.

### Feature Extraction: The Gabor Jet Lattice

The system represents input stimuli using a lattice of responses of 2-D Gabor wavelet filters (Daugman, 1985). The Gabor filter, essentially a sinusoidal grating localized by a Gaussian envelope, is a good model of simple cell receptive fields in cat striate cortex (Jones and Palmer, 1987). It provides an excellent basis for recognition of facial identity (Wiskott et al., 1997), individual facial actions (Donato et al., 1999), and facial expressions (Dailey and Cottrell, 1999; Lyons et al., 1999). We use phase-invariant Gabor magnitudes with a parameterization of the filter at five scales ranging from 16–96 pixels in width and eight orientations ranging from 0 to  $\frac{7\pi}{8}$  as described by Donato et al. (1999). Thus, at each point in the lattice (in our representation a  $29 \times 36$  grid of filter locations placed at regular 8-pixel intervals over the face), we extract a 40-element vector of Gabor magnitudes (sometimes called a “jet”) that characterizes a localized region of the face. A few of the filters are displayed graphically in Figure 1. To extract the  $29 \times 36 \times 40 = 41,760$  filter responses, we first convolve the entire image with each filter and take the magnitude of each complex valued response. We then (globally) divisively normalize the vector of responses at each filter scale to unit length. By equalizing the contribution of each filter size to the final representation, we overcome

the problem that most of an image’s power lies in lower spatial frequency ranges, without destroying information possibly present in the relative magnitude of response at each orientation. Since even the smallest filters in our representation overlap with their neighbors, and Gabor magnitudes are mildly invariant to slight translation, we lose very little of the information in the higher spatial frequency ranges, with a small price paid (due to ignoring phase information) in loss of precise feature localization and a larger price paid in that the resulting representation is very high dimensional (41,760 elements).

**Evaluation of the representation** In this section, we examine the representation’s utility and plausibility.

Donato et al. (1999) found that a nearest neighbor classifier with a cosine similarity metric applied directly to a Gabor grid-based representation achieved 95.5% correct classification of image *sequences* containing *individual facial actions* (Ekman and Friesen, 1978), e.g. facial action 1, the inner brow raiser. We evaluated this type of classifier on our task, classification of full-face expressions in static images. Nearest neighbor classification of the 96 expressive faces in POFA using leave-one-actor-out cross validation and a cosine similarity metric achieves an expected generalization accuracy of 74.0%. There are several possible reasons for this sub-par performance: the need to simultaneously integrate information from multiple facial actions, the small size of the POFA database, and/or the lack of information on the dynamics of facial movement. But the simple system’s performance is well above chance (16.7% correct), giving an indication that a more complicated (and more psychologically plausible) model such as a neural network could do much better.

One way of visualizing the effectiveness of a representation, and gaining insight into how an agent might use the representation to support decision-making, is to apply discriminant analysis.<sup>1</sup> For the Gabor magnitude components at a given location and spatial frequency, we find Fisher’s Linear Discriminant (Bishop, 1995), the projection axis  $\vec{w}$  that maximizes the criterion  $J(\vec{w})$ , the ratio of between-class to within-class scatter along  $\vec{w}$ .  $J(\vec{w})$  is a measure (invariant to linear transformations) of the diagnosticity of that portion of the representation for determining the class of the stimulus. That is, we can determine exactly how well (in the linear sense) the representation separates individual facial expressions.

We applied this method to the 85 expressive faces of a 12-actor subset of the POFA database. The results for Fear, the most difficult to recognize expression in POFA (for both humans and machines), are shown in Figure 2. The size of the dots placed over each grid location in the face is proportional to how easy it is to separate Fear from all of the other expressions based on the 8 Gabor filter responses extracted at that position of the grid. There are two interesting aspects to the result. First, the lowest spatial frequency channel (using filters about

<sup>1</sup>We introduced this visualization method for the Gabor representation in a recent technical report (Dailey and Cottrell, 1999), and Lyons et al. (1999) have independently introduced a similar technique.

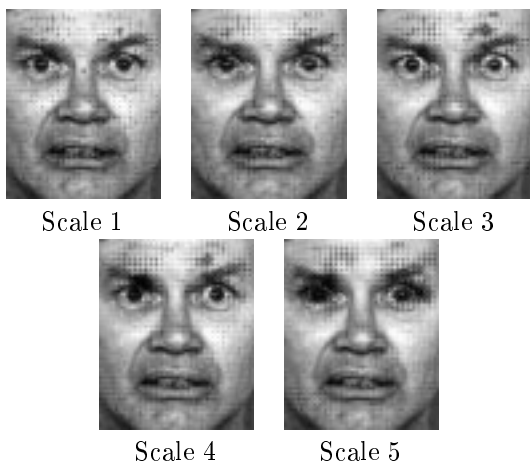


Figure 2: Diagnosticity of Gabor filter locations for Fear discrimination, separated by filter spatial frequency, from scale 1 (highest SF) to scale 5 (lowest).

96 pixels in width, compared to the total image width of 240) is best for this expression, implying that improvement might be obtained by dropping the smaller scales from the representation and even increasing the filter size. Second, the technique hints at which facial actions are most reliable for distinguishing expressions from one another, readily making predictions for psychological experiments. According to Ekman and Friesen (1978), prototypical displays of Fear include facial action 1 (inner brow raise), 2 (outer brow raise), 4 (scrunching together of the eyebrows), and 5 (upper eyelid raise) in the upper face, along with 25 (lips part) and some combination of 20 (lip stretch), 26 (jaw drop), or 27 (mouth stretch) in the lower face. Although some discriminability can be obtained in the higher spatial frequencies in the region of the mouth (presumably detecting facial action 25), our model finds that the best regions are in the lower spatial frequencies around the eyes, especially around the upper eyelids.

### Principal Components Analysis for Dimensionality Reduction

We use Principal Components Analysis (PCA) as a simple, unsupervised, linear method to reduce the dimensionality of the network’s input patterns by projecting each 41,760-element pattern onto the top  $k$  eigenvectors of the training set’s covariance matrix. This speeds up classifier training and improves generalization. We experimented with various values of  $k$  and achieved the best generalization results with  $k = 35$ , so in all experiments reported here we project training and test patterns onto the top 35 principal component eigenvectors of the training set, then use the standard technique of “z-scoring” each input to a mean of 0 and a standard deviation of 1.0 (Bishop, 1995).

### Classification by a Six Unit Network

The classification portion of the model is a six-unit neural network. Each unit in the network first com-

putes its net input, a weighted sum of the input pattern  $\vec{x}$ :  $a_i = b_i + \sum_j w_{ij}x_j$ . Then the softmax function  $y_i = e^{a_i} / \sum_k e^{a_k}$  is applied to the net inputs to produce a 6-element output vector  $\vec{y}$ . The network is trained with the relative entropy error function (Bishop, 1995). Since the outputs of this network must sum to 1.0, we use a constant target vector of  $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})^T$  for the neutral training stimuli.

With no hidden layer and just 35 elements in its input, the network is very small, but its number of parameters, 216, is still large compared to the number of training examples (88-99). Therefore, we must avoid overtraining the network; we have found that too-fast optimization techniques lead to poor generalization. We have obtained the best results using stochastic gradient, momentum, weight decay, and early stopping using a hold-out set. For the experiments reported here, we used a learning rate  $\eta = 0.0017$  (the number of units divided by the number of inputs times 0.01), a momentum  $\alpha = 0.9$ , and weight decay rate  $\nu = 0.01$ .

The early stopping technique bears some explanation. We obtain expected generalization results by leave-one-actor-out cross validation. For POFA, this means a network is trained on the images of 13 actors and tested on generalization to the 14th. Rather than training on the full 13 actors, we leave one out as a holdout set to help determine when to stop training. After each epoch of training on the remaining 12 actors’ faces, we test the network’s performance on the 13th actor (the holdout set). If classification accuracy on the holdout set has not improved in 6 epochs, we stop training and restore the weights from the best epoch. Training time under this paradigm varies greatly; it ranges anywhere from 60 to 300 epochs depending on which partition into training, holdout, and test set is used.

### Evaluation of the Network’s Performance

How does the network perform the expression recognition task? An examination of the trained network’s representation provides some insight. The idea is to project each unit’s weight vector back into image space in order to visualize what the network is sensitive to in an image. But this is not a trivial task; though PCA is linear and easily inverted, the Gabor magnitude representation, besides being subsampled, throws away important phase information. Normalization of the power in each spatial frequency channel could also be problematic for inversion. Current techniques for inverting Gabor magnitude representations (C. von der Malsburg, personal communication) are computationally intensive and make several assumptions that do not apply here. So we instead take a simpler approach: learning the function from the 35-element input space into facial image space with linear regression, then using the regression formula to produce an image that visualizes each network unit’s weight vector.

The results for one network trained on an arbitrary 12-actor subset of POFA are shown in Figure 3. In each image, each pixel value is the result of applying the regression formula predicting the value of the pixel at that



Figure 3: Images reconstructed by linear regression from a trained network’s weight vectors.

location as a linear function of the 35-element weight vector for the given network output unit. Dark and bright spots indicate the features that excite or inhibit a given output unit depending on the relative gray values in the region of that feature. Note that the representations are very much like one might predict given the linear discriminant analysis described earlier: each unit combines evidence based upon the presence or absence of a few local features; for Fear, the salient criteria appear to be the eyebrow raise and the eyelid raise, with a smaller contribution of parted lips.

An important factor not shown in Figure 3 is the effect output units have on each other. Due to the divisive normalization of the softmax function, an active output unit can effectively inhibit other units that are only mildly activated. Nevertheless, it seems clear from the reconstructions that the network’s effective strategy is to learn how the combination of facial actions involved in each prototypical expression can be reliably detected in a static image. We hypothesize that, when faced with a forced choice expression recognition task, humans must use similar representations and classification strategies. In the next two sections, we provide some indirect support for this hypothesis with both qualitative and quantitative comparisons between the model’s performance and human performance on the same stimuli.

### Modeling Forced-Choice Classification

Ekman and Friesen (1976) presented subjects with the task of 6-way forced choice classification of the expressive stimuli in POFA and provide the results of their experiment with the dataset. Their criterion for admission into the final database was that at least 70% of subjects should agree on each face’s classification into one of the six POFA expression categories. On average, the proportion of agreement (or chance of correct classification) was 91.7%.

### Classification accuracy comparison

We trained  $14 \times 13 = 182$  networks, one for each of the possible partitions of the database into a training set of 12 actors, a holdout set of one actor, and a test set of one actor. After training using the method described earlier, we tested each network’s classification accuracy on its generalization (test) set and averaged their performance. The 182 networks, on average, obtain a classification accuracy of 85.9% (compared to a human accuracy of 91.7%), and interestingly, the rank order of expression category difficulty, Happy – Disgusted – Surprised – Sad – Angry – Afraid, is *identical to that of the humans*. We also find that the humans and networks

show the same rank order. We have also found that it is possible to boost classifier accuracy on this task if the classifier is given the opportunity to “peek” at the test set (without labels) before actually classifying it. This “batch mode” classification technique is a plausible model for familiarizing subjects with the stimuli in an experiment prior to testing them. It boosts classifier accuracy to up to 95%; details are available in a technical report (Dailey and Cottrell, 1999).

### Visualization with Multidimensional Scaling

Multidimensional Scaling (MDS) is a frequently-used technique for visualizing relationships in high-dimensional data. It aims to embed stimuli in a low dimensional space (usually two or three dimensions) while preserving, as best possible, observed distances or similarities between each pair of stimuli. MDS has long been used as a tool for exploring the psychological structure of emotion. Russell has proposed a “circumplex” model of affect (Russell, 1980) that describes the range of human affective states along two axes, pleasure and arousal. Russell and colleagues have found support for their theory in a wide range of studies for which MDS consistently yields two-dimensional solutions whose axes resemble pleasure and arousal.

A similar technique can be applied to Ekman and Friesen’s forced-choice data. We computed a  $96 \times 96$  Euclidean distance matrix from the 6-dimensional response vectors supplied by Ekman and Friesen and used non-metric MDS<sup>2</sup> to find a 2-dimensional configuration of the 96 stimuli. This configuration, shown in the first graph of Figure 4, yielded a Kruskal stress  $S = 0.205$ . The circumplex embedded in Ekman and Friesen’s data, Happiness – Surprise – Fear – Sadness – Anger – Disgust, or HSFMAD (using M for Maudlin in place of Sadness to distinguish it from Surprise), is different from that typically reported by Russell and colleagues. This is not surprising, however, because a large portion of Russell’s circumplex (affective states that are negative on the arousal dimension and positive or neutral on the pleasure dimension, such as sleepiness, content, and relaxation) is simply not represented in POFA. The HSFMAD circumplex *is* the same, however, reported by Katsikitis (1997), who used the same set of expressions, a similar forced-choice arrangement, but an entirely different set of photographs in which the actors were not instructed on how to portray each expression.

Does the facial expression similarity structure induced by the network resemble the human psychological similarity structure in any way? We have performed MDS analyses at three levels in this network: at the input layer (on the Gabor/PCA representation), at the net inputs to the network’s output units (the units’ un-softmaxed activations  $a_i$ ), and at the softmax output layer. As one might expect, at the input layer, the patterns form

<sup>2</sup>There are many varieties of MDS; we implemented the Guttman-Lingoes SSA-1 algorithm as described in Borg and Lingoes (1987). Put briefly, the algorithm iteratively derives a configuration  $\mathbf{X}$  that minimizes Kruskal’s stress  $S$ , which is the proportion of variance in a monotonic regression unexplained by  $\mathbf{X}$ .

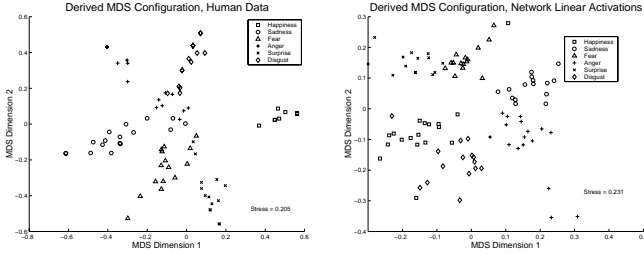


Figure 4: MDS configurations derived from human classification data and the linear activations of the units in the network model. The circumplex (order of stimuli around the graph) is the same: H-S-F-M-A-D (M=Maudlin/Sadness).

a cloud in the plane with little structure. At the network’s output, the responses on the training set tend to be so nearly binary that there is very little similarity structure. But using the net inputs to the softmax units, averaged over all 182 networks, we obtain a solution (stress = 0.231) that orders the expressions in the same way as the human circumplex, as shown in the second graph of Figure 4.

With the caveat that this only occurs in the linear part of the network, the fact that the human and network MDS solutions contain the same ordering is striking. It is very unlikely ( $p = 0.017$  for a single trial and  $p = 0.033$  for two trials) that we would obtain the same ordering if the human and network similarity structure were in fact unrelated.

### Correlation of network and human errors

MDS analysis is useful as a visualization tool, but the correspondence between the human circumplex and network circumplex is not a formal test of the model. Is the correspondence between the human and network MDS solutions simply a fortuitous coincidence? One way to address this concern is with a direct comparison of the confusion matrices for the humans and networks. For the humans and networks, we computed the  $6 \times 6$  confusion matrix whose  $ij$ -th entry gives the probability that when a face from class  $i$  is present, the humans or networks (on the training set) respond with expression  $j$ . Since the network was explicitly trained to produce label  $i$  for members of class  $i$ , we removed the diagonal elements from each confusion matrix and compared the network and human *error patterns*, i.e. the 30 off-diagonal terms of the confusion matrices. Note that it is not “cheating” to use the network’s responses on the training set here; the network was never biased in any way to make errors similar to humans. We found that the correlation between the off-diagonal elements of the confusion matrices for the humans and networks is  $r = 0.567$ . An  $F$ -test ( $F(1, 28) = 13.3; p = 0.0011$ ) confirms the significance of this result. These results lead us to claim that much of the facial expression similarity structure observable in forced-choice experiments is due to direct perceptual similarity, and that our model does an excellent job of capturing that structure.

## Modeling Perception of Morphs

Beyond the forced-choice classification data provided by Ekman and Friesen, the literature on categorical perception of facial expressions transitions is a treasure trove of data for modeling. Previous work (Padgett and Cottrell, 1998) compared a somewhat different facial expression recognition model to human behavior in a large study by Young et al. (1997) (henceforth referred to as “Megamix”). In the Megamix study, the researchers created morph stimuli interpolating each of the 21 possible transitions between six expressive images and one neutral image of POFA actor “JJ.” They then tested subjects on forced-choice identification of the perceived expression in the morphs (they also measured response times, discrimination, and the subjects’ ability to detect mixed-in expressions in the morph stimuli). Padgett and Cottrell (1998) simulated the Megamix morph stimuli with *dissolves*, or linear combinations of each source image and target image. Their linear feature extraction technique (projection of eye and mouth regions onto a Local PCA basis) and neural network classifier applied to the linear dissolves produced good results. However, when we created true morphs and attempted to apply the same techniques, we found that the model no longer fit the human data — there were large intrusions of unrelated expressions along the morph transitions, indicating that linear feature extraction is unable to produce a smooth response to nonlinear changes in the image. One might expect that the Gabor magnitude representation, with its built-in invariance to phase, might better capture the smooth, categorical transitions observed in the Megamix study on nonlinear morphs. In this section, we very briefly show that this is indeed the case: the Gabor/PCA-based model does produce smooth transitions between expression categories without intrusions and a very good fit to the human identification data without any free parameters.

### Network training

We used a slightly different methodology for modeling this data because we wanted to model each human subject with one trained network. This requires as much between-subject variability as possible (although variability is difficult to achieve given POFA’s small size). We trained 50 networks on different random partitions of the 13 non-JJ actors’ images into training and holdout sets. Each network’s training set consisted of 7 examples of each expression plus neutrality, with the remaining data used as a holdout set. As before, neutral stimuli were assigned the uniform target vector  $[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]^T$  and the expressive faces were assigned binary target vectors.

After training each network until holdout set classification error was minimized, we tested its performance on JJ’s prototypes as well as all morphs between them. We then extracted identification, response time, discrimination, and faint morph detection response variables from the model.



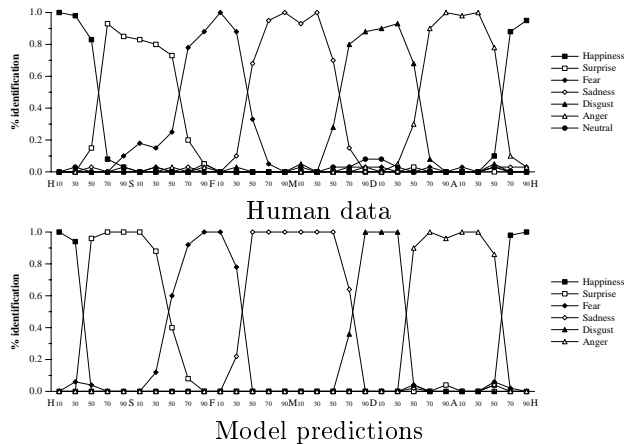


Figure 5: Human and network responses to JJ morphs along the transitions HSFMDA.

### Model fit

Using the same response variable measurements as Padgett and Cottrell (1998), we do find the Megamix pattern of sharp categorical transitions, scallop-shaped response time curves, improved discrimination near category boundaries, and a close correspondence between humans and networks on detection of the secondary expression in morph transitions. Due to space limitations, we cannot report all of the Megamix modeling results here, but we do show the model’s fit to the human responses on one series of morph transitions. Forced-choice identification results for the Happy – Surprised – Afraid – Sad – Disgusted – Angry – Happy transition series are shown in Figure 5. The human data and model prediction are quite similar, but the networks appear to place slightly sharper boundaries between expressions; this is because there is not as much variation in our population of network “subjects” as that occurring in the Megamix data. Nevertheless, the correspondence ( $r^2 = 0.846$ ) is remarkable considering that the networks were never trained on images of JJ or morph stimuli and that there are absolutely no free parameters involved in fitting the model to the data.

### Discussion

We have shown that a simple, mechanistic computational model obtains a natural fit to data from several psychological studies on classification of human facial expressions. Exploring the space of possible expression classification models has led us to reject several alternative models (including local PCA-based input representations and more complicated ensembles of networks containing hidden layers). Since one simple model, despite its lack of culture and social experience, explains so much data without any free parameter fitting, we claim that it is a strong model for how the human visual system perceives facial expressions in static images. To the extent that performance in the controlled forced-choice psychological experiments cited here generalizes to more naturalistic social situations (an admittedly big assump-

tion to make), we suggest that the model captures the essentials of the visual processing used to make many social judgments.

### Acknowledgments

We thank Gary’s Unbelievable Research Unit (GURU) for valuable comments on this research, Curtis Padgett for laying the foundation for the work, and Andrew Young for data obtained in his “Megamix” study. The research was funded by NIH grant MH57075 to GWC.

### References

- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press, Oxford.
- Carroll, J. M. and Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology*, 70(2):205–218.
- Dailey, M. N. and Cottrell, G. W. (1999). PCA = Gabor for expression recognition. UCSD CSE TR CS-629.
- Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Optical Society America A*, 2:1160–1169.
- Donato, G., Bartlett, M. S., Hager, J. C., Ekman, P., and Sejnowski, T. J. (1999). Classifying facial actions. *IEEE PAMI*, 21(10):974–989.
- Ekman, P. (1999). Basic emotions. In Dagleish, T. and Power, M., editors, *Handbook of Cognition and Emotion*. Wiley, New York.
- Ekman, P. and Friesen, W. (1976). *Pictures of Facial Affect*. Consulting Psychologists, Palo Alto, CA.
- Ekman, P. and Friesen, W. (1978). *Facial Action Coding System*. Consulting Psychologists, Palo Alto, CA.
- Ellison, J. W. and Massaro, D. W. (1997). Featural evaluation, integration, and judgment of facial affect. *JEP: HPP*, 23:213–226.
- Etcoff, N. L. and Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, 44:227–240.
- Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258.
- Katsikitis, M. (1997). The classification of facial expressions of emotion: A multidimensional scaling approach. *Perception*, 26:613–626.
- Lyons, M. J., Budynek, J., and Akamatsu, S. (1999). Automatic classification of single facial images. *IEEE PAMI*, 21(12):1357–1362.
- Padgett, C. and Cottrell, G. W. (1998). A simple neural network models categorical perception of facial expressions. In *Proc. 20th Cognitive Science Conference*, pages 806–807, Mahwah, NJ. Erlbaum.
- Russell, J. A. (1980). A circumplex model of affect. *J. Personality and Social Psych.*, 39:1161–1178.
- Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE PAMI*, 19(7):775–779.
- Young, A. W., Rowland, D., Calder, A. J., Etcoff, N., Seth, A., and Perrett, D. I. (1997). Facial expression megamix. *Cognition*, 63:271–313.

# Addressing the Learnability of Verb Subcategorizations with Bayesian Inference

**Mike Dowman** (Mike@cs.usyd.edu.au)  
Basser Department of Computer Science, F09,  
University of Sydney, NSW2006, Australia

## Abstract

Elman (1993) has shown that simple syntactic systems can be learned solely on the basis of distributions of words in text presentation. However Pinker (1989) has proposed that children must make use of verbs' semantic representations in order to infer their syntactic subcategorizations (semantic bootstrapping). Results reported here demonstrate how Bayesian statistical inference can provide an alternative, and much simpler, account of how subcategorizations are learned. The acquisition mechanism described here suggests that syntactic acquisition may involve a much larger component of learning, and less innate knowledge, than is presumed within mainstream generative theory.

## Introduction

This paper investigates how children learn their first language, and in particular the syntactic system of that language. It conceives of the problem in the following way: when exposed to utterances in that language, how is it possible to infer the grammatical system which produced those utterances. Further, the learner is assumed not to know the meanings of the words, have access to prosodic cues to structure, or to receive feedback about which sentences are not grammatical.

Currently the major paradigm within which language acquisition is explained is the parameter setting framework (Chomsky, 1995). Within this framework it is proposed that knowledge of language is largely specified innately, and learning consists of identifying word tokens and setting a limited number of parameters according to the syntactic structures to which the child is exposed. Chomsky argues that this position is necessary because 'even the most superficial look reveals the chasm that separates the knowledge of the language user from the data of experience.' (p. 5).

Gold (1967) investigated this problem more formally, and proved that without negative evidence (explicit information about which sentences are ungrammatical) languages are not 'learnable in the limit' unless the class of languages which the learner may consider is restricted *a priori*, for example by innate knowledge. Below I will discuss an alternative result by Feldman, Gips, Horning and Reder (1969) which suggests that Gold's result is not relevant to the circumstances under which children learn languages.

Redington, Chater and Finch (1998) investigated to what extent syntactic categories could be inferred based on distributions alone, without knowing *a priori* what syntactic categories existed in the language. They formed vectors by taking the two preceding and two following context words for each occurrence of each target word in a large corpus of transcribed speech, and recorded how often each context

word occurred in each position. Only the 150 most frequent words were used as context, and so this resulted in 600 dimensional vectors for each word (there being one entry for each of the 150 context words in each of four positions). Clustering those words whose vectors were most similar in terms of Spearman's rank correlation resulted in clusters which corresponded to appropriate word classes for most of the 1,000 target words. While this system was good in that it could be applied to naturally occurring speech, it was necessary to decide at what level of dissimilarity to form separate classes, and so it doesn't completely solve the problem of recovering the syntactic classes used by the original speakers.

Elman (1993) demonstrated that not only word classes, but also syntactic patterns in which words belonging to those classes appeared, could be learned without much innate syntactic knowledge, at least for simple languages. He trained a recurrent neural network to predict the following word in artificially generated sentences conforming to a simple syntactic system containing 23 words, and syntactic features such as number agreement and recursion in relative clauses. Once trained on 50,000 sentences in this simple language, the network performed at near optimum accuracy at predicting the subsequent word at any stage in a sentence, showing that the network had internalized the structural constraints implicit in the data.

While both Redington et al (1998) and Elman (1993) demonstrate that much of syntactic structure can be learned by making statistical inferences based on the distributions of words, Pinker (1989) suggests that some aspects of syntax cannot be learned in this way. He proposes that, in order to determine verbs' subcategorizations in the absence of negative evidence, children must rely on complex innate rules combined with knowledge of the verbs' semantic representations.

Verbs such as *give* can appear in both the prepositional dative construction (1a), and the double object dative construction (1b), but there is a class of verbs such as *donate* which can only appear in the prepositional construction, (1c and 1d). However Gropen et al (1989) observe that, based on the alternation between (1a) and (1b), children generalize this alternation to verbs such as *donate*, and so produce ungrammatical sentences such as (1d). They also demonstrated that when presented with novel, nonce, verbs in the prepositional construction, children will productively use them in the double object construction in appropriate contexts. However, ultimately children do learn which verbs cannot occur in the double object construction, and so we need a theory which can explain why children first make such gen-

eralizations, and then subsequently learn the correct sub-categorizations.

- (1) a. John gave a painting to the museum.
- b. John gave the museum a painting.
- c. John donated a painting to the museum.
- d. \*John donated the museum a painting.

While the main point of Pinker (1989) is that syntax cannot be learned from distributions alone, he acknowledges that the fact that certain syntactic structures do not occur could be used as indirect negative evidence that these structures were ungrammatical. However, he notes that children can neither consider that all sentences which they have not heard are not grammatical, and nor do they rule out all verb argument structure combinations which they haven't heard. He notes that it is necessary to identify 'under exactly what circumstances does a child conclude that a nonwitnessed sentence is ungrammatical?' (p.14). The computational model presented in this paper is able to do just this, and so predict that a verb such as donate cannot occur in the double object construction, while at the same time predicting that a novel verb encountered only in the prepositional construction will follow the regular pattern and also appear in the double object construction.

### Bayesian Grammatical Inference

Most work in syntactic theory assumes that grammars are not statistical, that is that they specify allowable structures, but do not contain information about how frequently particular words and constructions occur. However, if grammars were statistical, it appears that it would be much easier to account for how they were learned. Feldman et al (1969) proved that as long as grammars were statistical, and so utterances were produced with frequencies corresponding to the grammar, then languages are learnable. They note that proofs that language isn't learnable rely on the possibility of an unrepresentative distribution of examples being presented to the learner. While under Feldman et al's learning scheme it is not possible to be certain when a correct grammar has been learned, as more data is observed it becomes more and more likely that the correct grammar will be identified.

Feldman et al's proof uses Bayes' theorem, which relates the probability of a hypothesis given observed data to the *a priori* probability of the hypothesis and the probability of the data given the hypothesis. For a fixed set of data the best hypothesis is that for which the product of the *a priori* probability of the hypothesis and the probability of the data given the hypothesis is greatest. Feldman et al relate the probability of a grammar (seen as a hypothesis about language) to its complexity – more complex grammars are less probable *a priori*. As grammars are statistical, it is also possible to calculate the probability of the data given a grammar. This leads to an evaluation criterion for grammars where the complexity of a grammar is weighed off against how much data it has to account for, and how well it fits that data. A more complex grammar can be justified if it accounts for regularities in the data, but otherwise a simpler grammar will be preferred.

Minimum coding length provides an efficient implementation of Bayesian inference, using information theory (Shannon, 1948), which allows us to quantify the amount of information in a formal description of a grammar. The amount of information conveyed by an event (or symbol in a grammar) is equal to the negative logarithm of its probability. It is conventional to take logarithms to base two, resulting in the units of quantity of information being bits. Within this framework the best grammar is that which, together with a description of a corpus of data in terms of the grammar, can be specified using the least amount of information.

While Feldman et al (1969) showed that, given two or more grammars, it is possible to decide which is the best given a corpus of data, they did not show how these grammars could be created. For any reasonably complex grammar, the number of possible, but incorrect, grammars of equal or simpler complexity is so large that it is not plausible that a child could consider each in turn. However, in the next section, I describe computational models which are able to learn grammars by starting with a simple grammar, and then making small iterative changes which gradually lead towards the correct grammar. This avoids the need to consider every single possible grammar, and so allows grammars to be learned within a reasonable amount of time.

### Computational Models of Syntactic Acquisition

Langley (1995) and Stolcke (1994) used simplicity metrics to learn simple syntactic systems, while Goldsmith (submitted) has applied this approach to the acquisition of morphology. Both Langley and Stolcke's systems produced similar results to those found by Dowman (1998) using the model described in the next section, although Langley's (1995) system did not incorporate considerations of how well the grammar fitted the data. It is shown below how Dowman's (1998) model was used to obtain new results concerning the acquisition of verb subcategorizations.

### Description of Model

Dowman's (1998) model learned grammars for simple subsets of several languages, including the English data given in Table 1, which corresponds to the grammar given in Table 2. The only *a priori* knowledge of the structure of the corpus which was available to the model was implicit in the grammatical formalism with which grammars were specified. This formalism restricted the model to using binary branching or non-branching phrase structure rules, introducing each word with a non-branching rule, and using no more than eight non-terminal symbols. The non-terminal symbols were all equivalent arbitrary symbols, except that each grammar would contain one special symbol, *S*, with which each top down derivation would begin.

The frequency, and hence probability, with which each symbol (including words) appeared in the grammar was specified, and so the amount of information required to specify each symbol in a grammar could be calculated (using Shannon's (1948) information theory). A specification of a grammar would consist of a list of groups of three symbols, one for a rule's left hand side, and two for its right

hand side (a special null symbol being incorporated for use in non-branching rules). As the grammar was statistical, it was also necessary to record how often each rule was used in parsing the corpus. It was assumed that a fixed amount of information could be used to specify these probabilities, and so 5 bits of information was added to the evaluation of the grammar per rule. (The assumption of 5 bits of information is fairly arbitrary, but sufficient for the purposes described here.) The total cost of the grammar was the amount of information needed to specify each symbol in the grammar, and each rule's frequency.

Table 1: Data for English

John hit Mary	Ethel thinks John ran
Mary hit Ethel	John thinks Ethel ran
Ethel ran	Mary ran
John ran	Ethel hit Mary
Mary ran	Mary thinks John hit Ethel
Ethel hit John	John screamed
Noam hit John	Noam hopes John screamed
Ethel screamed	Mary hopes Ethel hit John
Mary kicked Ethel	Noam kicked Mary
John hopes Ethel thinks Mary hit Ethel	

Table 2: Grammar Describing English Data

$S \rightarrow NP VP$	$V_s \rightarrow$ thinks
$VP \rightarrow$ ran	$V_s \rightarrow$ hopes
$VP \rightarrow$ screamed	$NP \rightarrow$ John
$VP \rightarrow V_t NP$	$NP \rightarrow$ Ethel
$VP \rightarrow V_s S$	$NP \rightarrow$ Mary
$V_t \rightarrow$ hit	$NP \rightarrow$ Noam
$V_t \rightarrow$ kicked	

Given such grammars, the data was then parsed left to right, bottom up, with only the first parse found for each sentence being considered, and an ordered list of rules needed to derive the sentence obtained. This list allows us to make a probabilistic encoding of the data in terms of the grammar. Given the probabilities of the rules, and always knowing the current non-terminal symbol being expanded (starting with  $S$ , and always expanding the left most unexpanded non-terminal), it is only necessary to specify which of the possible expansions of that symbol to make at each stage. Hence, if a grammar accounts well for regularities in the data, little information will be required to specify the data. If a symbol can only be expanded by a single rule (such as  $S$  in the grammar above), then no information is necessary to specify that that rule is used.

By summing the amount of information needed to specify the grammar rules, the frequencies of those rules, and the data given that grammar, we obtain an evaluation for each grammar, with lower evaluations corresponding to better grammars. However, in order to complete the model of acquisition, it is necessary to describe the search mechanism that was used for generating and testing grammars.

The model started learning with a simple grammar of the form given in Table 3, with a rule introducing each word. This grammar is very simple, hence having a good evalua-

tion itself, but it does not describe any regularities in the data, and so has a very bad evaluation in that respect, resulting in a poor overall evaluation.

Table 3: Form of Initial Grammars

$S \rightarrow X S$	$S \rightarrow X$
$X \rightarrow$ John	$X \rightarrow$ thinks
$X \rightarrow$ screamed	$X \rightarrow$ Ethel

The model would begin learning by making one of four random changes to the grammar, either adding a new rule (which would be the same as an old rule, but with one of the symbols changed at random), deleting a randomly chosen rule, changing one of the symbols in one of the rules, or the order of the rules, or adding a pair of rules in which one non-terminal symbol occurring on the left hand side of one and the right hand side of another was changed to a different non-terminal symbol. These changes are slightly simpler than those described in Dowman (1998), but further investigations have revealed that this learning system works well, and it was able to reproduce the results obtained with the more complex system, so it was used for deriving the new results presented in this paper.

After each change the evaluation of the new grammar with respect to the data would be calculated. If the change improved the evaluation of the grammar then it would be kept, but if the new grammar was unable to parse the data, it would be rejected. If the change made the evaluation of the grammar worse, then the probability that it would be kept would be inversely proportional to the amount by which it made the evaluation worse, and also throughout learning the probability that changes resulting in worse evaluations would be accepted was gradually reduced. This is an implementation of annealing search, which enables the system to learn despite finding locally optimal grammars in the search space. The program learned in two stages, in the first only taking account of the evaluation of the data in terms of the grammar (making it easier to find the grammatical constructions which best fitted the data), and in the second taking account of the overall evaluation (and so removing any parts of the grammar which could not be justified given the data). After a fixed number of changes had been considered (less than 18,000 in the case of the above data) learning would finish with the current grammar, no improvements usually having been found for a long time. For efficiency reasons, there were also limits placed on how deeply the parser could search for correct parses, and on the maximum number of rules which the grammar could contain at any stage of the search. Because the search strategy is stochastic, it is not guaranteed to always find the optimal grammar every time, so the learning mechanism would run the search several times, and select the grammar with the best overall evaluation.

## Results

When used to learn from the English data in Table 1, the system learned a grammar which corresponded exactly to that in Table 2 in structure. (As linguistic categories are not known *a priori*, the system simply used a different arbitrary

symbol to represent each learned category.) Table 4 shows that this grammar was preferred because, while the grammar itself is more complex than the initial one, and so receives a worse evaluation, it captures regularities in the data, and so improves the evaluation of the data with respect to the grammar by a greater amount. Dowman (1998) used this same learning system (without any modifications except to the maximum number of non-terminal symbols) to learn aspects of French, Japanese, Finnish and Tigak.

Table 4: Evaluations for English Grammar

	Initial state of learning	Learned Grammar
Overall Evaluation	406.5 bits	329.5 bits
Grammar	160.3 bits	199.3 bits
Data	246.2 bits	130.3 bits

### Learning Verb Subcategorizations

Given Dowman's (1998) success in learning simple syntactic systems, it was decided to investigate whether the same model could be used to learn some of the kinds of phenomena which it has been argued are especially problematic for theories of learning. In particular it was investigated whether the distinction between sub-classes of ditransitive verbs such as *gave* and *donated* could be learned.

There were three key results which the model aimed to replicate. Firstly, children eventually learn a distinction between verbs which can appear in both the double object and prepositional dative constructions, and those which do not show this alternation. Secondly, when children encounter a previously unseen verb they use it productively in both constructions. Finally, during learning, before children have seen many examples of an irregular verb which only occurs in a subset of the possible constructions of other verbs, they use that verb productively in constructions in which it is not grammatical.

### Data Used for Learning

The same model was used as in Dowman (1998), but this time the data consisted of two types of sentences, prepositional datives such as (2a) and (2b), containing one of the verbs *gave*, *passed*, *lent*, or *donated*, and double object datives such as (2c), containing *gave*, *passed* or *lent*, but not *donated*. Each of these four verbs occurred with roughly equal frequency, and the alternating verbs were just as likely to appear in either construction. In addition the sentence (2d) was added, containing the only example of the verb *sent*. Noun phrases consisted of either one of two proper nouns, or one of the two determiners *a* or *the*, followed by either *painting* or *museum*. There were no biases as to which noun phrase was most likely to occur in which position, and overall the data consisted of 150 sentences.

No modifications were made to the model of Dowman (1998), except that in order to cope with the more complex data set the maximum number of non-terminals was increased to 14, and the number of iterations in the search was also increased.

- (2) a. John gave a painting to Sam.
- b. Sam donated John to the museum.
- c. The museum lent Sam a painting.
- d. The museum sent a painting to Sam.

### Results

The initial and final evaluations of the grammars are given in Table 5. Again a more complex grammar has been learned which accounts better for regularities in the data than the original grammar. Examination of the learned grammar showed that the verbs had been divided into two classes (they have different symbols on the left hand sides of the rules producing them). *gave*, *passed*, *lent* and *sent* had all been placed in one class, while *donated* appeared in a class of its own. The grammar is able to generate only grammatical sentences, so *gave*, *passed*, *lent* and *sent* may appear in both double object and prepositional constructions, while *donated* may occur only in the prepositional dative construction. This has been learned even though there was no data explicitly indicating that *donated* did not follow the regular pattern, and even though *sent* only occurred once, and in the prepositional structure.

Table 5: Evaluations for Ditransitive Verbs Data

	Initial state of learning	Learned Grammar
Overall Evaluation	3445.6 bits	1703.4 bits
Grammar	190.3 bits	321.0 bits
Data	3255.3 bits	1382.3 bits

The results above account both for eventual learning of the distinction between syntactically distinct verbs such as *gave* and *donated*, and the productive use of novel verbs in regular constructions. The final phenomenon which we aimed to demonstrate was that, at earlier stages of learning, children overgeneralize and use verbs such as *donated* productively in constructions in which they are ungrammatical. In order to investigate this phenomenon, the total amount of data was reduced, to simulate a stage of acquisition where children had not been exposed to so many examples of each kind of verb. When the model learned from this data it failed to maintain a distinction between sub-classes of verbs, allowing all verbs to occur in both constructions. This was because there were not enough examples of *donated* to justify making the grammar more complex by creating a separate syntactic class, and so it was simply placed in the regular class.

### Discussion

These results on the acquisition of regular and irregular verb subcategorizations show that an aspect of syntax is learnable which many other theories would have difficulty accounting for. In particular it is interesting to compare the performance of the model described here to that of connectionist models of syntactic acquisition such as Elman (1993).

Elman's network learned a language containing only 23 words, and yet 50,000 sentences were used to train the net-

work. This means that every word could have been observed in every syntactic position many times over, greatly reducing the need to form generalizations. Christiansen and Chater (1994) investigated to what extent this kind of model was able to generalize to predict that a word observed in one syntactic position would also be grammatical in another position. In order to do this, they trained a similar connectionist network on a more complex language containing 34 words, again using 50,000 sentences. In the training data they did not include *girl* and *girls*, in any genitive contexts, and, *boy* and *boys* in any noun phrase conjunctions. After training they found that the network was able to generalize so that it would allow *boy* and *boys* to appear in noun phrase conjunctions, but it didn't generalize to allow *girl* and *girls* to occur in genitive contexts. Christiansen and Chater considered the learning to have been successful in the case of *boy* and *boys*, but not in the case of *girl* and *girls*.

However, the account of the acquisition of verb subcategorizations presented in this paper relies on statistical properties of the data, and in particular the non-occurrence of certain forms. So, given 50,000 sentences of a language with only 34 words, in which two words did not appear in a given construction, it would seem that a learner would predict that this could not simply be due to chance. Given this perspective, it seems that Christiansen and Chater's network has learned correctly in the case of *girl* and *girls*, but not in the case of *boy* and *boys*.

In order to account for distinctions between *gave* and *donated*, it seems that neural networks must be more sensitive to quantitative information in language. The degree to which recurrent neural networks generalize is partly dependent on the fixed architecture of the network, and in particular on the number of hidden nodes. Bayesian learning methods for neural networks (MacKay, 1995) should be able to solve this problem, by placing a prior probability distribution on network structures and parameter values, although I am not aware of any applications of such networks to models of language acquisition.

Redington et al's (1998) system for learning word classes is capable of making very fine distinctions between subclasses of verbs, but unlike the system described here it is not able to decide when the distributions of two words are dissimilar enough that they should be placed into separate classes, and when the difference in distributions is simply due to chance variation within a class. However Boulton (1975) describes a program which does incorporate a Bayesian based metric into this kind of clustering system, and so demonstrates that it is possible to learn discrete classes automatically.

Certainly evaluation procedures based on simplicity metrics are not new to linguistic theory. Chomsky's (1965) theory of syntactic acquisition relied on such a measure to choose between alternative grammars. However, it is possible to identify some key differences which make Chomsky's theory very different to the Bayesian approach suggested here. Firstly Chomsky considered syntax to be fundamentally non-statistical. He had earlier argued that 'Despite the undeniable interest and importance of semantic and statistical studies of language, they appear to have no direct rele-

vance to the problem of determining or characterizing the set of grammatical utterances....[P]robabilistic models give no particular insight into some of the basic problems of syntactic structure.' (Chomsky, 1957, p17). It seems hard to explain how any system which didn't monitor the frequencies with which verbs such as *donated* and *gave* are used would be able to account for how the different subcategorizations of these verbs could be acquired.

Probably an even more important difference between the kind of simplicity measure proposed in Chomsky (1965) and the kind used here, is that Chomsky did not incorporate a measure of goodness of fit to data into his simplicity metric. Chomsky's metric simply looked for the grammar which was shortest, in terms of the number of symbols which it contained. The theory relied on innate constraints on what forms grammar could take in order that 'significant considerations of complexity and generality are converted into considerations of length, so that real generalizations shorten the grammar and spurious ones do not.' (p42). Ultimately any notion of a simplicity metric was dropped from syntactic theory, because little progress seemed to be being made in understanding grammar selection in this way.

Interestingly however, Chomsky's (1965) theory shows that simplicity metrics are not necessarily incompatible with theories which postulate very strong innate constraints on grammar. It seems that even within a parameter setting model of language acquisition, statistical inferences would make the task of learning much easier, especially given the presence of noise in the data from which people learn (due primarily to grammatical errors, and exposure to data from children who have not mastered certain aspects of grammar). Showing that Bayesian inference can be useful in explaining language acquisition does not necessarily mean that it is actually used. Essentially it allows us to return the degree to which language is determined by innate principles of grammar to an empirical question, allowing the possibility of a much greater degree of learning in the process of syntactic acquisition.

However, postulating that a Bayesian mechanism is used in acquiring syntax results in very different predictions about what form syntactic knowledge will take than if we presume that language is largely determined by universal principles. Chomsky (1995) has argued that the language faculty of the mind should satisfy 'general conditions of conceptual naturalness that have some independent plausibility, namely, simplicity, economy, symmetry, nonredundancy, and the like' (p. 1). While Chomsky notes this is 'a surprising property of a biological system' (p. 5) he argues that this view is justified because throughout the history of syntactic research systems conforming to this kind of principle have turned out to be the right ones. However, if language is learned with a Bayesian system we would not expect it to conform to such principles. Grammars could contain a lot of irregular rules if these accounted well for regularities in observed language. Even the principle of lexical minimization is not so clear cut within a Bayesian based account of learning, as Bayesian metrics will favor grammars which associate a lot of information with individual words if this allows them to account better for regularities in the data. Hence, one prediction of Bayesian theory is that

the most commonly occurring words may be very idiosyncratic and irregular in their behavior, while very rare ones must conform to regular patterns.

It is interesting to compare the Bayesian account of acquisition of subcategorizations presented here to Pinker's (1989) theory. Pinker's theory predicts that universal innate principles relate the meaning of a word to its syntactic subcategorization. Instead of the syntactic subcategorization of a verb being determined empirically by a learner based on observations of patterns of occurrence, it is determined by the meaning of that verb. Certainly Gropen et al (1989) have shown that children are sensitive to correlations between semantic and phonological characteristics of verbs, and which subcategorization frames they are most likely to occur in. However, it is quite possible that these patterns were learned by the child in much the same way as we have proposed that syntactic subcategorizations may be learned. It would be interesting to investigate empirically whether children or adults could be influenced to prefer verbs in one construction or another by controlling the exemplars of these verbs to which they were exposed, perhaps by using artificial language experiments or nonce verbs integrated into natural languages. This kind of experiment should be able to resolve to what extent children make use of innate principles versus learning in determining verbs' subcategorizations.

The main limitation of the computational model described here is that it can only learn from small artificial data sets. There is no reason in principle why it cannot operate on naturally occurring language, it is simply that it would take an extremely long time to run on this kind of corpus. This is clearly a limitation which is shared with connectionist approaches, though Redington et al (1998) demonstrate impressive results learning from real language corpora. Current research is investigating ways in which the search procedure could be made more efficient, so that learning from more realistic corpora is possible, though it seems worth acknowledging that we are modeling a process which takes place over many years, and that the human brain is much more powerful than any computer.

## Conclusion

This paper has shown that Bayesian inference is able to provide a simple and plausible account of how a number of aspects of syntax could be learned. In particular the computational model described here can learn verb subcategorizations where one verb is grammatical in only a subset of the structures in which another can appear, and yet predicts that newly encountered verbs are used productively in regular patterns. The model also accounts for overgeneralization and hence the use of irregular items in regular constructions during early stages of acquisition. While it is not logically necessary that children must make use of Bayesian inference in learning language, it has the potential to be incorporated into theories as diverse as recurrent neural networks and universal grammar.

## Acknowledgments

I would like to thank Jeff Elman, David Powers, Brett Baker, Hong Liang Qiao, Adam Blaxter Paliwala, Cassily Charles, and two anonymous reviewers, for helpful comments on this paper. This research was supported by ARC and IPRS scholarships.

## References

- Boulton, D. M. (1975). *The Information Measure for Intrinsic Classification*. Ph.D. Thesis, Monash University, Melbourne.
- Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton & Co.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.
- Christiansen, M. H. & Chater, N. (1994). Generalization and Connectionist Language Learning. *Mind and Language*, 9, 273-287.
- Dowman, M. (1998). *A Cross-linguistic Computational Investigation of the Learnability of Syntactic, Morphosyntactic, and Phonological Structure* (Research Paper EUCCS-RP-1998-6). Edinburgh, UK: Edinburgh University, Centre for Cognitive Science.
- Elman, J. L. (1993). Learning and Development in Neural Networks: The Importance of Starting Small. *Cognition*, 48, 71-99.
- Feldman, J. A., Gips, J., Horning, J. J., & Reder, S. (1969). *Grammatical Complexity and Inference* (Tech. Rep. CS 125). Stanford, CA: Stanford University: Computer Science Department.
- Gold, E. M. (1967). Language Identification in the Limit. *Information and Control*, 16, 447-474.
- Goldsmith, J. (submitted). Unsupervised Learning of the Morphology of a Natural Language.
- Gropen, J., Pinker, S., Hollander, M., Goldberg, R. & Wilson, R. (1989). The Learnability and Acquisition of the Dative Alternation in English. *Language*, 65, 203-257.
- Langley, P. (1995), *Simplicity and Representation Change in Grammar Induction* (Unpublished Manuscript). Palo Alto, CA: Institute for the Study of Learning and Expertise.
- MacKay, D. J. C. (1995). Bayesian Methods for Supervised Neural Networks. In Arbib, M. A. (Ed.) *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press.
- Pinker, S. (1989), *Learnability and Cognition The Acquisition of Argument Structure*. Cambridge, MA: MIT Press.
- Redington, M., Chater, N. & Finch, S. (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science*, 22, 425-469.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27, 379-423 & 623-656.
- Stolcke, A. (1994). *Bayesian Learning of Probabilistic Language Models*. Doctoral dissertation, Department of Electrical Engineering and Computer Science, University of California at Berkeley.

## Hidden Markov Models for Coding Story Recall Data

**Michael A. Durbin (golden@utdallas.edu)**

Cognitive Science Program (Attention: Professor Golden)  
University of Texas at Dallas, GR4.1, Box 830688  
Richardson, TX 75083-0688

**Jason Earwood (golden@utdallas.edu)**

Psychology Program (Attention: Professor Golden)  
University of Texas at Dallas, GR4.1, Box 830688  
Richardson, TX 75083-0688

**Richard M. Golden (golden@utdallas.edu)<sup>1</sup>**

Psychology and Cognitive Science Programs, GR4.1, Box 830688  
University of Texas at Dallas  
Richardson, TX 75083-0688

### Abstract

Current methods of coding recall, summarization, talk-aloud, and question-answering data are inherently unreliable and not effectively documented. If the process of coding protocol data could even be partially automated, this would be an important scientific advance in the field of text comprehension. Twenty-four human subjects read and recalled each of four short texts. Half of the human recall data (the "training data") was coded by a human coder and then used to estimate the parameters of a set of Hidden Markov Models (HMMs) where each HMM was associated with a particular complex proposition in the text. The Viterbi algorithm was then used to assign the "most probable" complex proposition to human-coder specified text segments in the remaining half of the human recall data (the "test data"). The HMM algorithm made coding decisions which agreed well with a human coder's decision on the test data indicating that the HMM is indeed capable of formally representing a human coder's "theory" of how text segments should be mapped into complex propositions for simple texts.

### Introduction

Theories and experiments in the field of text comprehension often require mapping recall (e.g., Golden, 1997), summarization (e.g., van den Broek & Trabasso, 1986), talk-aloud (e.g., Trabasso & Magliano, 1996), and question-answering (e.g., Graesser & Franklin, 1990) protocol data into a semantic model of the implicit and explicit information in text clauses. This semantic model of the information in the text clauses has been referred to by Kintsch (1998) as the *textbase microstructure*. Typically this initial *coding procedure* of mapping the protocol data into a textbase microstructure is done using human coders. Inter-coder reliability measures are then used to establish the reliability of the coding procedure.

This widely used coding procedure methodology, however, has several problems. First, such coding procedures

are typically not well documented. Second, the reliability of such procedures is often highly dependent upon "human coders", who despite their best intentions, are prone to inconsistent coding behaviors (especially over very large coding tasks). Third, such coding procedures are typically not readily accessible to other researchers. And fourth, coding procedures across research labs located in different parts of the world are not standardized in any particular manner.

An ideal solution to these problems would be to develop an automated approach to coding human protocol data (as advocated by Ericsson and Simon, 1993). Although important progress in this area has been made (see especially Kintsch, 1998, Chapter 3), additional work is required. It should also be emphasized that the task of coding human protocol data is not nearly as complex as the full-fledged natural language understanding problem. Consider a typical experiment where a group of human subjects are asked to recall the same story from memory. Although the resulting protocol data will be extremely rich and varied, typically the text comprehension researcher is only interested in detecting a relatively small number of complex propositions. This dramatically simplifies the pattern recognition problem.

The main goal of this research is to develop and empirically evaluate a new theoretical framework for reliably mapping protocol data into a textbase microstructure. Specifically, a Hidden Markov Model (HMM) (see Allen, 1995; Charniak, 1993; Jelinek, 1997; for relevant reviews) is constructed for each complex proposition in each of four short stories. The stories, based upon classic fables, each consisted of approximately 10-15 short sentences with each sentence corresponding roughly to a complex proposition (Golden, 1997). Twenty-four human subjects read and recalled each of the four short texts (see Golden, 1997, for additional details). Half of the human recall data (the "training data") was coded by a human coder and then used to estimate the parameters of the HMM associated with each

<sup>1</sup> The order of the authors is arbitrary. Please address all correspondence to Richard M. Golden.



complex proposition. The prior probability that a particular complex proposition was used by the human coder was also recorded. Next, the Viterbi algorithm (Viterbi, 1967; see Allen, 1995; Charniak, 1993; Jelinek, 1997) was used to assign the "most probable" complex proposition to human-coder specified text segments in the remaining half of the human recall data (the "test data"). Measures of agreement between the human coder and AUTOCODER were then computed using only the test data. A high measure of agreement indicates that the HMM is indeed capable of formally representing a human coder's "theory" of how text segments should be mapped into complex propositions.

## Method

### Human Protocol Data

**Texts.** The human protocol data used consisted of recall data associated with four texts collected by Golden (1997). The four texts ("Cuckoo", "Miser", "Eagle", and "Doctor") were especially written to have approximately similar levels of syntactic and semantic complexity. Each sentence in the text was written to conform approximately to: (1) a standard subject-verb-object form, and (2) such that each sentence corresponded roughly to one complex proposition. For example, the "Miser" text read by the human subjects is shown below.

#### The "Miser" Text (Golden, 1997)

A miser bought a lump of gold using all of his money. The miser buried the gold in the ground. The miser looked at the buried gold each day. One of the miser's servants discovered the buried gold . The servant stole the gold . The miser , on his next visit , found the hole empty . The miser was very upset . The miser pulled his hair . A neighbor told the miser not to be upset . The neighbor said , " Go and take a stone , and bury it in the hole . "The neighbor said , " And imagine that the gold is still lying there ." The neighbor said , " The stone will be as useful to you as the gold . " The neighbor said, " When you had the gold , you never used it . "

**Recall Protocol Data.** Twenty-four college students read and verbally recalled each of four texts ("Miser", "Cuckoo", "Doctor", and "Eagle") from memory as described in Golden (1997). The recall data was then transcribed. Text segments in all of the recall protocol data corresponding to complex propositions were then identified by human coders. The recall data from twelve of the college students was designated as *training data*, while the recall data from the remaining twelve college students was designated as *test data*.

To provide some insights into the richness and complexity of the statistical pattern recognition problem considered

in this paper. Here is an example recall protocol extracted from the training data set.

#### Subject 1 recall of "Miser Text" (training data set)

someone that a servant that knew that discovered the money# and took it# and then the miser saw that the money was gone# and he was upset# and complained to a neighbor# and the neighbor said well just get a stone and bury your money# dig a hole and bury the money# because it'll do you just as much good as your real money your gold is doing you#

The symbol # in the above recall protocol associated with subject 1 refers to the marking of text segments by an experienced human coder. Text segments corresponding to complex propositions were marked by experienced human coders for both the training data and test data sets. Here is a representative recall protocol from subject 12 who was assigned to the test data set. The complexity of the recall data (even when a human coder has already identified text segments) is readily apparent (compare recall data of Subject 1, Subject 12 with one another and the original "Miser" text).

#### Subject 12 recall of "Miser Text" (test data set)

and he buried it in the ground # and he went over every day to look at where the money was where the lump of gold was buried# and one day when the miser was- n't there a thief came and dug up the lump of gold# and so the miser goes and he sees the hole in the ground# and he's very upset by that# and a bystander tells the miser to take a rock and bury it in the ground# and the miser says why# and the bystander says well all you ever did was look at the ground anyway# you never did use the gold# so there might as well be a rock there#

### Parameter Estimation (Learning Algorithm)

The learning process involves a specially designed graphical user-interface which is referred to as AUTOCODER. Figure 1 shows a typical AUTOCODER display. A subject's recall data (in this case, the recall data for Subject 12) is displayed. The human coder first segments the text so that each word sequence in each text segment corresponds to a complex proposition. Beneath each word is a pull-down menu consisting of a series of concepts. The human coder decides which words (or word sequences) should be assigned concepts, and then uses the pull-down menu to assign a concept to each selected word within a given text segment. Another pull-down menu is then used to

assign a complex proposition to a given sequence of concepts within a text segment.

**Probabilistic Modeling Assumptions.** Let  $W_1, \dots, W_M$  be the ordered sequence of words (or more generally word phrases) within a particular text segment which an experienced human coder has decided should be assigned concepts. Let  $C_i$  denote the concept assigned to the  $i$ th word,  $W_i$ . Let  $F$  be the complex proposition assigned to the concept sequence  $C_1, \dots, C_M$ .

After the human coder has completed the coding task, AUTOCODER has stored the following items for the human coder. First, a concept dictionary consisting of the concepts created by the human coder. Second, a complex proposition dictionary consisting of the complex propositions created by the human coder. Third, the percentage of times that a particular complex proposition  $F$  has been used (denoted by  $p(F)$ ). Fourth, the percentage of times that a word (or word phrase)  $W_i$  is used to express the concept  $C_i$  (denoted by  $p(W_i | C_i)$ ) is computed (this is referred to as the "emission probability" in the HMM literature). And fifth, the percentage of times that one concept follows another concept given a particular complex proposition  $F$  (denoted by  $p(C_{i+1} | C_i, F)$ ) (this is referred to as the "transition probability" in the HMM literature). Given the usual conditional independence assumptions of an HMM, these statistics in conjunction with the concept and complex proposition dictionaries correspond to a particular type of probabilistic theory of how the human coder codes the recall data.

For example, consider the text segment "*He buried his life savings deeply in the ground*". The human coder might choose to model this text segment as an ordered sequence of word phrases: ( $W_1 = \text{"He"}$ ,  $W_2 = \text{"buried"}$ , \*,  $W_3 = \text{"life savings"}$ , \*, \*, \*, \*) might be associated with the ordered sequence of concepts: ( $C_1 = \text{"MISER"}$ ,  $C_2 = \text{"BURY"}$ , \*,  $C_3 = \text{"GOLD"}$ , \*, \*, \*, \*) where the notation \* is used to refer to a word (or word phrase) which is not assigned a concept for the purposes of coding the protocol data. The complex proposition  $F = \text{"BURY(MISER, GOLD)"}$  would be assigned to the concept sequence ( $C_1 = \text{"MISER"}$ ,  $C_2 = \text{"BURY"}$ , \*,  $C_3 = \text{"GOLD"}$ , \*, \*, \*, \*).

Once the assignments have been made, statistics are computed. Specifically, the probability that one concept follows another given a particular complex proposition (e.g.,  $P(\text{BURY} | \text{MISER, BURY(MISER, GOLD)})$ ) is estimated from the observed relative frequencies. In addition, the probability of a word given a concept is estimated (e.g.,  $P(\text{"life savings"} | \text{GOLD})$ ). The probability that a given complex proposition is used is also estimated from the coder's behavior (e.g.,  $P(\text{BURY(MISER, GOLD)})$ ). Instead of assigning a zero probability to transition and emission probabilities whose corresponding observed relative frequencies were equal to zero, a small "smoothing" probability was used to facilitate processing of novel word sequences. Figure 2 shows a possible HMM representation for the complex proposition **BURY(MISER, GOLD)**.

### Protocol Data Coding Algorithm

The Viterbi algorithm (Viterbi, 1967) as described in Allen (1995, p. 202) was then used to construct the "most probable" concept sequence associated with each possible

complex proposition for a particular text segment. The "information content" in bits (i.e., a normalized log-likelihood measure)  $I$  of a complex proposition  $F$  consisting of  $M$  concepts  $C_1, C_2, \dots, C_M$  and represented by  $M$  word phrases  $W_1, \dots, W_M$  is computed using the formula:

where  $\log[x]$  denotes the logarithm base 2.

$$I = -(1/M) \log \left[ p(F) \prod_{i=1}^M p(C_i | C_{i-1}, F) p(W_i | C_i) \right]$$

Next, the complex proposition which was "most probable" (i.e., had the smallest information content score  $I$ ) was selected. Complex propositions whose information content exceeded some maximum critical value were discarded and those text segments were defined as "incomprehensible" to AUTOCODER. This threshold was set sufficiently high, however, so that the occurrence of "incomprehensible" complex propositions was very rare. Notice that unlike the usual HMM approach to syntactic and semantic parsing, a unique HMM is constructed for each complex proposition rather than trying to construct a general HMM applicable to all possible complex propositions which could occur in the text.

### Procedure

Three human coders jointly coded the recall data from the training data set using AUTOCODER. The human coders were careful not to examine the test data, so the dictionaries created as a result of coding the training data were likely to not contain all concepts, complex propositions, and statistics necessary to code the test data set. Text segments in the test data were then identified by the three human coders as well. AUTOCODER then assigned the "most probable" complex proposition to each text segment using the information content score described in the previous section. The three human coders then coded the test data without the use of AUTOCODER and measures of agreement between AUTOCODER's performance and the human coder performance on the test data set were recorded.

### Results and Discussion

In order to compare performance of AUTOCODER and the human coder on the test data set, three different measures of agreement were used. All measures were computed individually for each text across all relevant subject data. It is important to emphasize that AUTOCODER always codes the same set of protocol data in exactly the same manner with 100% reliability. Thus, the agreement measures actually are measures of the validity as opposed to the reliability of AUTOCODER's coding performance.

### Agreement Measures

The first measure was *percent agreement* which is defined as the percentage of times the two coders agree that a proposition was mentioned in the recall protocol plus the percentage of times the two coders agree that a proposition was not mentioned. One difficulty with the percent agreement measure is that percent agreement can be artificially increased by simply increasing the number of complex

propositions in the proposition dictionary! Accordingly, other agreement measures were considered.

The second measure of agreement was Cohen's Kappa score (Cohen, 1960) which essentially corrects for agreement by chance. The formula for Cohen's Kappa is given by:  $\kappa = (p - p_c) / (1 - p_c)$  where  $p$  is the *percent agreement* described in the previous paragraph and  $p_c$  is the expected agreement between the two coders if the coding strategy of one coder provided no information (i.e., was statistically independent of the coding strategy of the other coder). The performance of the model for the percent agreement and kappa agreement measures on the training data set is provided in Table 1. The quantity  $N$  denotes the number of opportunities for agreement. Typically, in the text comprehension literature, Percent agreement scores for coding data which are above 90% and kappa scores which are above 70% are deemed acceptable for publication. The data was also analyzed using a third more stringent agreement measure we call *sequential agreement*. Sequential agreement is typically not computed. But since the same coder has identified the text segments in both the training and test data, the percentage of times both the human coder and AUTOCODER agreed upon the coding of a particular text segment across recall protocols could be computed. This coding criterion thus takes into account the sequential structure of the recall data unlike the previously described agreement measures which are typically reported in the literature.

### Analysis of Training Data

Table 1 shows the performance of AUTOCODER on the training data set using standard agreement measures, while Table 2 shows the performance of AUTOCODER using the sequential agreement measure. As can be seen from Tables 1 and 2, AUTOCODER's performance clearly demonstrates that it is picking up on a sufficient number of statistical regularities from the skilled human coder's data to almost completely reconstruct the skilled human coder's decisions.

Table 1: Performance of Autocoder on Training Data (Standard Agreement Measures)

Text	N	Percent Agreement	Cohen Kappa
"Miser"	192	95%	91%
"Cuckoo"	336	93%	84%
"Doctor"	228	99%	97%
"Eagle"	384	97%	93%

Table 2: Performance of Autocoder on Training Data (Sequential Agreement Measures)

Text	N	Percent Agreement
"Miser"	111	90%
"Cuckoo"	111	86%
"Doctor"	105	98%
"Eagle"	150	92%

### Analysis of Test Data

Tables 3 and 4 show the performance of AUTOCODER on the test data set using the standard agreement measures and the sequential agreement measure. As can be seen from Tables 3 and 4, AUTOCODER's performance is almost comparable to experienced human coders keeping in mind the limitation that the test data set was parsed into text segments corresponding to complex propositions by a human coder. On the other hand, the AUTOCODER methodology has the important advantage that it is entirely well-documented and can be reliably implemented by computer software (unlike coding schemes implemented by human coders).

Table 3: Performance of Autocoder on Test Data (Standard Agreement Measures)

Text	N	Percent Agreement	Cohen Kappa
"Miser"	192	83%	65%
"Cuckoo"	336	88%	71%
"Doctor"	228	88%	75%
"Eagle"	384	84%	66%

Table 4: Performance of Autocoder on Test Data (Sequential Agreement Measures)

Text	N	Percent Agreement
"Miser"	111	69%
"Cuckoo"	111	67%
"Doctor"	105	76%
"Eagle"	150	68%

To provide a qualitative feeling regarding AUTOCODER's performance, Table 5 shows AUTOCODER's "coding" of the protocol data of Subject 12 who was assigned to the test data set.

It is extremely encouraging (despite the simple texts considered in this initial study) that the performance of the AUTOCODER algorithm was so effective on the test data. In almost all cases, AUTOCODER automatically and reliably coded the data at an almost publishable agreement level using completely documented and accessible algorithms. We are excited and pleased with these preliminary results even though the text segments in the test data had to be pre-

parsed by a human coder. Future work in this area is currently being pursued.

Table 5: AUTOCODER's "coding" of novel recall data

Human Recall Data	AUTOCODER Interpretation
"and he buried it in the ground"	<b>BURY</b> <b>AGENT: MISER</b> <b>OBJECT: GOLD</b>
"and he went over every day to look at where the money was where the lump of gold was buried"	<b>ATTEND</b> <b>AGENT: MISER</b> <b>OBJECT: GOLD</b>
"and one day when the miser wasn't there a thief came and dug up the gold"	<b>ATTEND</b> <b>AGENT: MISER</b> <b>OBJECT: GOLD</b> [Disagrees with Human Coder!]
"and so the miser goes and he sees the hole in the ground"	<b>BURY</b> <b>AGENT: MISER</b> <b>OBJECT: GOLD</b> [Disagrees with Human Coder!]
"and he's very upset by that"	<b>MISER</b> <b>STATE: PLEASED</b> [Disagrees with Human Coder!]
"and a bystander tells the miser to take a rock and bury it in the ground"	<b>TELLS-INFO</b> <b>FROM: NEIGHBOR</b> <b>TO: MISER</b> <b>INFO: BURY(STONE)</b>
"and the miser says why"	<b>ATTEND</b> <b>AGENT: MISER</b> <b>OBJECT: GOLD</b> [Disagrees with Human Coder!]
"and the bystander says well all you ever did was look at the ground anyway"	<b>TELLS-INFO</b> <b>FROM: NEIGHBOR</b> <b>TO: MISER</b> <b>INFO: ATTEND (MISER, GROUND)</b>
"you never did use the gold"	<b>TELLS-INFO</b> <b>FROM: NEIGHBOR</b> <b>TO: MISER</b> <b>INFO: NOTUSE (MISER, GOLD)</b>
"so there might as well be a rock there"	<b>TELLS-INFO</b> <b>FROM: NEIGHBOR</b> <b>TO: MISER</b> <b>INFO: ASGOOD (STONE, GOLD)</b>

### References

Allen, J. (1995). *Natural language understanding*. Redwood City, CA: Benjamin/Cummings.

Charniak, E. (1993). *Statistical language learning*. Cambridge, MA: MIT.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Earwood, J. (2000). *AUTOCODER: An intelligent assistant for coding protocol data*. Psychology Program Senior Honors Thesis. School of Human Development. University of Texas at Dallas. Richardson, TX.

Ericsson, K. & Simon, H. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT.

Golden, R. M. (1997). Causal network analysis validation using synthetic recall protocols. *Behavior Research Methods, Instruments, and Computers*, 29, 15-24.

Graesser, A. & Franklin, S. (1990). Quest: A cognitive model of question-answering. *Discourse Processes*, 13, 270-304.

Jelinek, F. (1997). *Statistical methods for speech recognition*. Cambridge, MA: MIT.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge.

Trabasso, T. & Magliano, J. (1996). Conscious understanding during comprehension. *Discourse Processes*, 21, 255-287.

van den Broek, P. & Trabasso, T. (1986). Causal networks versus goal hierarchies in summarizing texts. *Discourse Processes*, 9, 1-15.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, 13, 260-269.

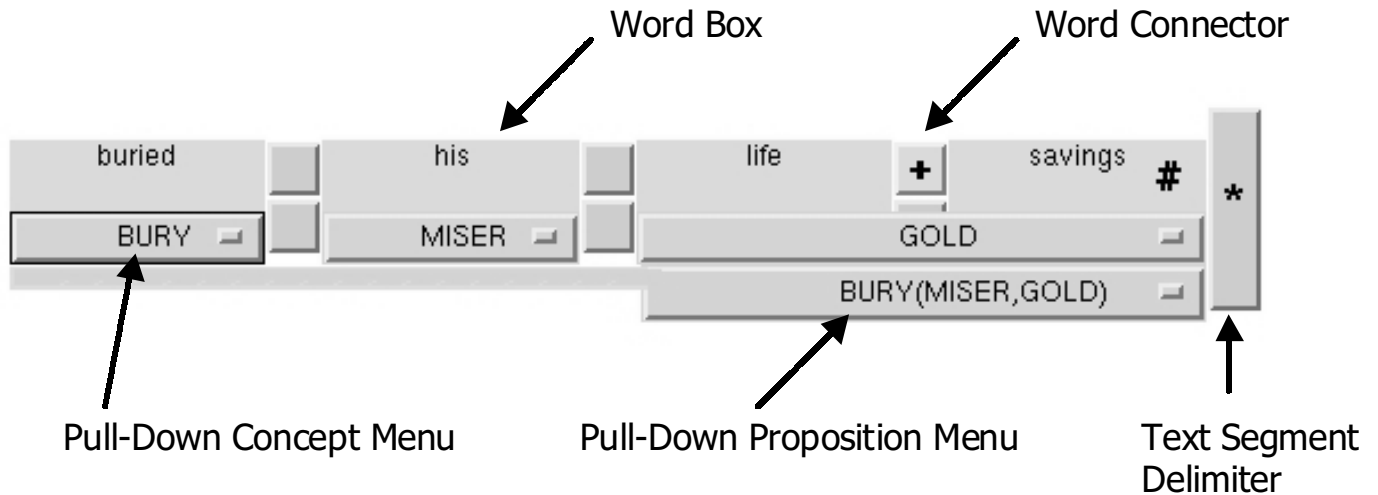


Figure 1. A portion of the AUTOCODER user-interface associated with the coding of the phrase "buried his life savings". Each word in the text appears in a particular window called the *word box*. Word boxes can be connected to form word phrases using the *connector button*. Beneath each word phrase is a pull-down *concept menu*. Another pull-down *proposition menu* which lists the set of available complex propositions which can be assigned to the phrase is also displayed to the user. Both concept and proposition menus provide facilities for the addition of new concepts and propositions by the skilled human coder. Menu choices are made by a skilled human coder for the purposes of providing training data for the Hidden Markov Models (HMMs). The HMMs are then used to automatically make "most probable" menu selections without the aid of a skilled human coder through the use of the Viterbi algorithm for HMMs as described in the text.

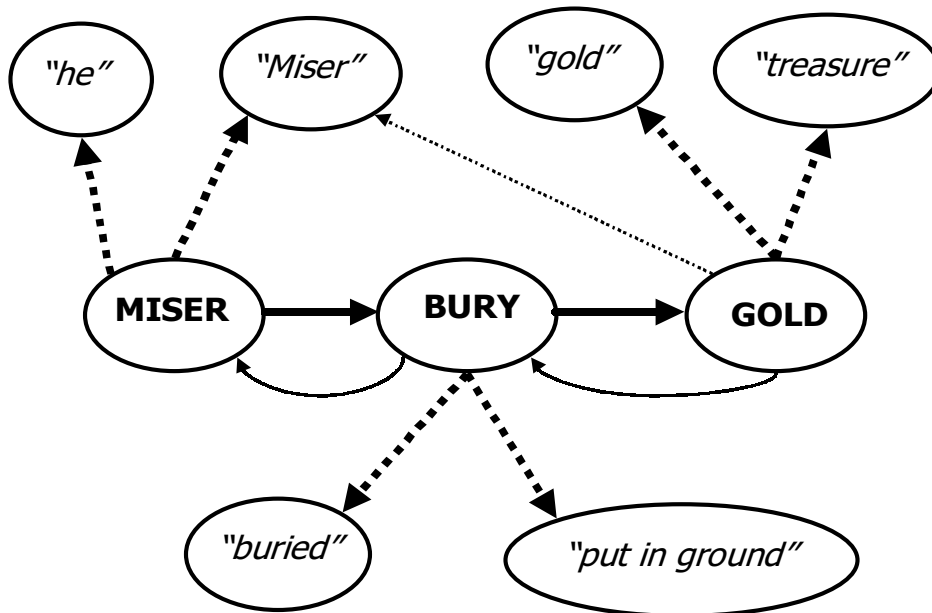


Figure 2. Each complex proposition is represented by its own HMM (Hidden Markov Model). In this figure, the HMM for the proposition **BURY(MISER, GOLD)** is graphically displayed. Transition probabilities are represented by solid arrows while emission probabilities are represented by dashed arrows. Line thickness indicates the relative magnitude of the corresponding transition or emission probability. Thus, the line thicknesses for the emission probability  $P(\text{Word} = \text{"gold"} \mid \text{Concept} = \text{GOLD})$  and transition probability  $P(\text{Concept} = \text{GOLD} \mid \text{Concept} = \text{BURY}, \text{Proposition} = \text{BURY(MISER, GOLD)})$  are both much thicker than the line thicknesses for the emission probability  $P(\text{Word} = \text{"Miser"} \mid \text{Concept} = \text{GOLD})$  and transition probability  $P(\text{Concept} = \text{BURY} \mid \text{Concept} = \text{GOLD}, \text{Proposition} = \text{BURY(MISER, GOLD)})$ .

# A Rarity Heuristic for Hypothesis Testing

**Aidan Feeney**

Department of Psychology  
University of Durham  
Science Laboratories  
South Road  
Durham DH1 3LE  
United Kingdom  
aidan.feeney@durham.ac.uk

**Jonathan St.B.T. Evans & Simon Venn**

Centre for Thinking and Language  
Department of Psychology  
University of Plymouth  
Drake's Circus  
Plymouth PL4 8AA  
United Kingdom  
{jevans, svenn}@plymouth.ac.uk

## Abstract

This paper presents the results of two experiments designed to investigate the processes underlying the effects of beliefs about probabilities on an hypothesis testing task. Both experiments demonstrate that although such effects exist, they are inflexible in the face of explicit statistical and implicit contextual manipulations of the likely information to be gained from selecting evidence concerning rare features. It is argued that these results suggest the operation of a rarity heuristic in hypothesis testing whilst possible adaptive functions for such a heuristic are discussed.

## Probabilities and Hypothesis Testing

Over the last ten or more years, human hypothesis testing, which previously had been viewed as being prone to bias (Wason, 1960; Doherty et al, 1979), has been rehabilitated. One of the central claims to be made during this process of rehabilitation is that human hypothesis testing is in some way adapted to the probabilistic structure of our environment. For example, Klayman and Ha (1987) have argued that confirmation, verification and matching biases, amongst others, may be viewed as the result of a generalised positive test strategy in hypothesis testing. Klayman and Ha demonstrated how such a strategy could be a good heuristic in environments with a realistic probabilistic structure. Their claim is that hypothesis testing tasks such as Wason's 2 4 6 task lead to non-normative behaviour because they encourage participants to adopt a generally sensible strategy in an experimental situation whose probabilistic structure does not match the strategy.

More recently, Oaksford and Chater (1994), in the spirit of Anderson's (1990) more general 'rational analysis' of cognition, have proposed a decision-theoretic account of Wason's selection task (Wason, 1966). This account is based on both the probabilistic structure of the task itself and assumptions about people's understanding of abstract conditional hypotheses. Our aim in this paper is to extend the study of probabilistic effects to another hypothesis testing task and to gain some insight into the mechanisms underlying such effects.

## Probabilities and Pseudodiagnosticity

Feeney, Evans and Clibbens (1997) have considered the role of background probabilities in determining performance on the pseudodiagnosticity (PD) task. Pseudodiagnosticity (Doherty et al, 1979) is the tendency to select information relevant to just one of a pair of hypotheses when trying to decide between them. An example of the standard paradigm used to investigate pseudodiagnosticity is taken from Mynatt, Doherty and Dragan (1995):

Your sister has a car she bought a couple of years ago. It's either a car X or a car Y but you can't remember which. You do remember that her car does over 25 miles per gallon and has not had any major mechanical problems in the two years she's owned it.

You have the following information:

A. 65% of car X's do over 25 miles per gallon.

Three additional pieces of information are also available:

B. The percentage of car Y's that do over 25 miles per gallon.

C. The percentage of car X's that have had no major mechanical problems for the first two years of ownership.

D. The percentage of car Y's that have had no major mechanical problems for the first two years of ownership.

Assuming you could find out only one of these three pieces of information (B, C or D) which would you want in order to help you to decide which car your sister owns? Please circle your answer.

In the standard PD task, as above, an anchor is provided (item A) which provides some potentially supportive evidence for one of the two hypotheses presented in the scenario. This we term the *focal* hypothesis. According to Doherty and his colleagues, the normatively correct answer to this problem is to choose item B which provides - in Bayesian terms - a completed likelihood ratio and allows the diagnosticity of the evidence to be assessed. For example,

we might discover that only 25% of Y's do over 25 mpg, favouring X or that 90% of Y's do over 25 mpg, favouring the Y hypothesis. However, the more common response is for people to choose item C, thus learning more about X. In the study quoted, 28% of participants chose B (deemed correct), 59% chose C and 13% chose D. In the absence of information about Y, however, items A and C provide only pseudodiagnostic evidence for X.

The original interpretation of this apparent error by Doherty et al (1979) - with general support in the later literature - was that it constituted a form of confirmation bias similar to that observed on other tasks such as the Wason 2 4 6 problem (see Evans, 1989, Klayman, 1995 for extended discussion of confirmation bias effects). It is supposed that people think only about the focal hypothesis, fail to consider alternatives and try to find evidence to confirm their favored hypothesis.

However, the analysis of the task becomes more complex if one takes into account background beliefs that the participant may bring to the experiment. Suppose, for example, that you were told that your sister's car had a radio and a top speed of over 165 miles per hour. If the information provided was then that most X's have a radio, according to the standard normative analysis people ought to choose to discover whether most Y's also have a radio. However, since they know a priori that most cars have radios, the participants could reason that this will be true of most Y's as well and that nothing will be learned by choosing this option. On the other hand discovering whether X does over 165 miles an hour (a rare feature among cars) would provide good evidence relative to background beliefs about the likelihood of this feature. Given these beliefs, such evidence could be regarded as being implicitly diagnostic rather than as being pseudodiagnostic. In this case, one can actually argue that the PD choice is correct, because its expected information gain (Oaksford, Chater & Larkin, 1999) or epistemic utility (Evans & Over, 1996) is higher, relative to background beliefs.

Feeney, Evans and Clibbens (1997) have shown that when the initial piece of evidence concerns a rare feature and the second piece concerns a common feature, then people will seek to discover a second piece of evidence about the rare feature, leading to a large drop in the usual PD choice rate. This tendency, to make diagnostic selections when evidence concerning a rare feature is available has been replicated on three different variants of the task (Feeney, Evans and Venn, 2000). In a separate version of the paradigm in which participants rate their degree of belief in the focal hypothesis after one or two pieces of 'pseudodiagnostic' information, we also found (Feeney, Evans and Clibbens, in press) that people are significantly more confident in a hypothesis supported by rare rather than common evidence. These findings support the view that rare information is taken to be implicitly diagnostic.

Whilst the experiments described above have established a robust influence of feature Rarity, it is not clear whether this is due to tacit influence of background beliefs or

whether people are consciously reasoning about the expected epistemic utility of the evidence. This ambiguity is indicative of a more general confusion (Oaksford Chater & Larkin, 1999; Klayman and Ha, 1987) in the literature on hypothesis testing where it is unknown whether people's apparent sensitivity to the probabilistic structure of their environment is the result of hard-wired heuristics, or is due to extensive on-line processing of environmental probabilities. We will now describe two experiments designed to resolve this ambiguity.

## Experiment 1

In this experiment, we used problems which were structurally identical to those used by Mynatt et al (see above). In Mynatt et al's experiment participants received a scenario containing a target object, said to possess two features, and two hypothesised categories. Next participants received a piece of evidence concerning the rate at which one of the hypothesised categories possessed one of the features. Finally, participants were asked to select one of the remaining three pieces of information to help them make a judgement about category membership.

In Experiment 1 we manipulated the relationship between the rarity of the evidential features and the explicit information presented about the rate at which features were present under either hypothesis. In the belief-compatible conditions, participants were told that a rare feature was present in only 10% of cases for the initial hypothesis (e.g. 10% of car X's do over 165 mph) or that a common feature was present in 80% of cases (e.g. 80% of car X's have radios). In the belief-incompatible conditions, participants were told that the rare feature was present in 80% of X's or that the common feature was present in 10% of X's. If a simple rarity heuristic is operating then we would expect participants still to favour rare features regardless of the explicit information given. However, if they are reasoning on-line about the probability of the evidential features then the percentage data should interact with the Rarity manipulation. Specifically, when told that the common feature is present in only 10% of X's (common feature, belief-incompatible), we might now expect diagnostic choices to go up (and focal choices to be suppressed) even though these involve the common feature. This is because people could reason that most Y's will probably have the feature and hence the choice will be diagnostic. When told that 10% of X's contain the rare feature (belief-compatible) we might also expect a drop in the usual diagnostic choice rates for rare evidence since they will expect Y to have a similar rate. Hence, the on-line processing hypothesis predicts a cross-over interaction between the two variables.

## Method

One hundred and eighty seven students from the University of Plymouth took part in this experiment which had a 2x2 between participants design. Each participant received a booklet comprising of an instructions page and four

problems. The basic structure of the problems used was identical to that used by Mynatt et al. The factors manipulated were Rarity and the strength of the initial statistic presented (we will refer to this variable as the Percentage variable). The Rarity manipulation in this experiment was between participants and was achieved by manipulating the first feature about which participants were given some evidence. These features were chosen on the basis of a pre-test and are shown in Table 1.

Table 1: Results of pre-test on materials used in Experiments 1 and 2.

Content	2 <sup>nd</sup> Feature	1 <sup>st</sup> Feature - Common	1 <sup>st</sup> Feature - Rare
House	Garage	Garden	Swimming pool
Engineer	Company car	Earns £14,000 pa	Earns £60,000 pa
Car	Has a radio	Top speed 90 mph +	Top speed 165 mph +
Holiday Villa	Built last 20 years	£150 per week	£1000 per week

The Percentage manipulation was achieved by manipulating, between participant, the strength of the initial piece of evidence. Half of the participants were told that 10% of instances of the focal category shared a feature with the target whilst the other half were told that this figure was 80%. The problem contents employed in this experiment concerned a house, an engineer, a car and a holiday villa. The order of the evidential options was counterbalanced whilst the order of the problems was randomised.

## Results and Discussion

Evidence selection patterns, when collapsed across experimental condition, are very similar for all problem contents. On the Engineer problem 41% of selections were of B (diagnostic selections), 50% were of C (further information about the focal hypothesis) and 9% were of D (information for the non-focal hypothesis concerning the second feature). The equivalent statistics for the Spanish Villa problem are 37%, 50% and 13%, 42%, 44% and 14% for the Car problem and 41%, 50% and 9% for the House

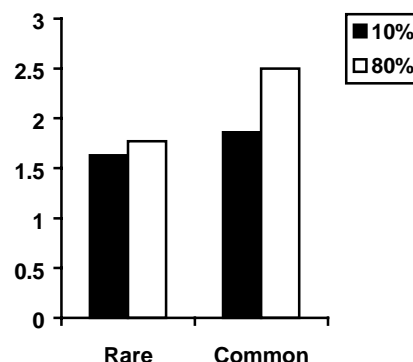
Table 2: Item choices as a percentage of total choices in each condition for Experiment 1.

	Rare		Common	
	10	80	10	80
Item B	47%	43%	41%	28%
Item C	42%	44%	46%	63%
Item D	11%	13%	13%	9%

problem. Selection frequencies for the entire experiment broken down by experimental condition are presented in Table 2. The mean number of pseudodiagnostic, or item C, choices was calculated for each participant across the four

problem contents. The mean number of item C choices, broken down by Rarity and Percentage Type, is presented in Figure 1. A 2x2 between-participants Anova was carried out on the mean number of item C choices. A significant main effect was found for Rarity ( $F(1, 183) = 4.53$ ,  $MSE = 2.40$ ,  $p < .04$ ). The mean number of item C choices made by participants in the Rare and Common conditions was 1.70 ( $S.D. = 1.57$ ) and 2.18 ( $S.D. = 1.54$ ) respectively. Neither

Figure 1: Mean Number of Focal Selections by Condition in Experiment 1.



the main effect of Percentage ( $F(1, 183) = 2.98$ ,  $MSE = 2.40$ ,  $p > .08$ ) nor the interaction between Rarity and Percentage ( $F(1,183) = 1.19$ ,  $MSE = 2.40$ ,  $p > .25$ ) were found to be statistically significant.

These results suggest the operation of a rarity heuristic in hypothesis testing. Our results contained a significant main effect of Rarity, although the apparent interaction fell short of significance. Whilst it is clear from Table 2 and Figure 1 that the percentage information has no effect on choice rates for rare information, Figure 1 does reveal a marginally significant trend ( $p < .06$ ) for common choices to be debiased in the belief-incompatible condition.

The trend for focal selections to increase when the initial evidence is that less than 50% of focal instances possess a common feature was found previously by Mynatt et al (1993) and interpreted by them as due to the initial evidence disconfirming hypothesis X and switching focus to Y. Although this trend is also consistent with an on-line processing hypothesis, that hypothesis also predicts a corresponding increase in focal choices when rare information was present at 80%. The latter trend was clearly absent. However, the trend which is to be seen in our data is consistent with the claim, made by Mynatt et al, that people select evidence relevant to the hypothesis they believe to be true.

## Experiment 2

In our previous experiment we demonstrated that people are relatively insensitive to explicit changes in the initial piece of information that they receive and that the rarity effect is



robust in the presence of such changes. In Experiment 2 we aimed to test whether people are sensitive to contextual changes which should also affect the epistemic utility of their choices.

In this experiment we attempted to reduce the implicit diagnosticity of the rare features by presenting them in a context where they would not be rare. For example, in the Car scenario both types of car were said to be sports cars and hence much more likely to have a high top speed. As all participants in the experiment were told that 80% of Xs possessed either the common or rare features used in previous experiments, the implicit change to the scenarios was the only difference between the two conditions run in this experiment and the 80% conditions of the previous experiment. Combining both sets of conditions gives us a 2x2 design with rarity of initial feature and relationship between alternatives the between subject variables.

### Method

One hundred and four new participants from the University of Plymouth were recruited for this experiment each of whom received a booklet comprising of a set of instructions and four problems. The instructions for this experiment were identical to those used in Experiment 1 whilst the problem contents used were almost identical to those used in the previous experiments. The rarity manipulation in this experiment was achieved with the same features as used in previous experiments and all participants were told that 80% of instances of the focal category shared a feature with the target object.

The difference between the new conditions in this experiment and the 80% conditions of the previous experiment is the implicit diagnosticity of the rare feature. Thus, we will refer to the second between participants factor in this experiment as Implicit Diagnosticity.

### Results and Discussion

Once again, selection patterns, when collapsed across our new experimental conditions, are very similar for all four problem contents. On the Engineer problem 37% of

Table 3: Item choices as a percentage of total choices in each condition for Experiment 2.

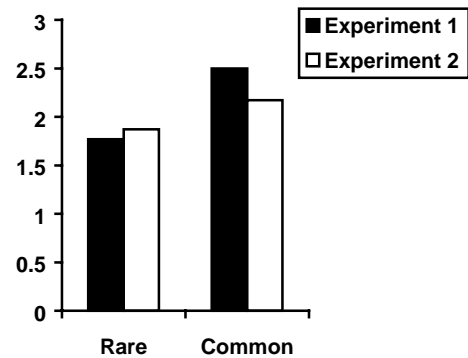
	Low Implicit Diagnosticity (Exp. 2)		High Implicit Diagnosticity (Exp. 1)	
	Rare	Com	Rare	Com
Item B	42%	34%	43%	28%
Item C	47%	54%	44%	63%
Item D	11%	12%	13%	9%

selections were of B, 50% were of C and 13%% were of D. The equivalent statistics for the Villa problem are 35.5%, 52% and 12.5%, 38%, 51% and 11% for the Car problem

and 40%, 49% and 11% for the House problem. Selection frequencies for the entire experiment broken down by experimental condition are presented in Table 3.

The mean number of item C choices was calculated for each participant across the four problem contents. The mean number of C choices, broken down by Rarity and Implicit Diagnosticity, are presented in Figure 2. In order to examine the effect of our Implicit Diagnosticity manipulation a 2x2 between participants Anova (with Rarity as the second factor) was carried out on the mean number of item C choices made by participants in this experiment and participants in the 80% conditions of Experiment 1. Once again, a significant main effect was found for Rarity ( $F(1, 199) = 5.745$ ,  $MSE = 2.365$ ,  $p < .02$ ). The mean number of item C choices made by participants in the Rare and Common conditions was 1.81 (S.D. = 1.57 and 2.33 (S.D. = 1.55) respectively. Neither the main effect of Implicit Diagnosticity ( $F(1, 199) = 0.235$ ,  $MSE = 2.365$ ,  $p > .6$ ) nor the interaction between Rarity and Implicit Diagnosticity ( $F(1,199) = 0.95$ ,  $MSE = 2.248$ ,  $p > .33$ ) were found to be statistically significant.

Figure 2. Mean number of focal selections in Experiment 2 as a function of condition.



These results demonstrate the persistence of people's tendency to make fewer focal choices when the initial piece of information concerns a rare feature, even when the implicit diagnosticity of that rare feature has been contextually reduced. This provides further evidence of a robust rarity heuristic that is relatively insensitive to contextual variations.

### General Discussion

The first conclusion to be drawn from the results of the experiments described in this paper is that they support the findings of Feeney, Evans and Clibbens (1997; in press). People are sensitive to the probabilities of the evidential items about which they reason on the PD task. More important is our failure to moderate the effects of feature rarity using either an explicit statistical manipulation or an

implicit contextual manipulation. The failures of these manipulations suggest that the effect of feature rarity is mediated via a hard-wired heuristic rather than any sophisticated on-line processing of probabilities. Whilst this heuristic is sensitive to rare features of objects, it is insensitive to changes in explicit statistical and implicit contextual information which affect the diagnosticity of those features. Accordingly, although we agree with Oaksford, Chater, and Larkin (1999) who argue that the very existence of probabilistic effects in hypothesis testing tasks indicates that people perform some on-line processing of probabilities, we feel that our results strongly suggest that the extent of such processing is severely limited. The effects of feature rarity on the PD task seem instead to be due to the operation of a relatively inflexible rarity heuristic.

### Functional and Dysfunctional Aspects of a Rarity Heuristic

As with any heuristic in judgement or hypothesis testing, a rarity heuristic conveys both advantages and disadvantages. Most obviously, given the results of our experiments, an inflexible rarity heuristic renders the information processor insufficiently sensitive to changes in the diagnosticity of rare experimental features. However, as we have claimed elsewhere (Feeney et al, in press), sensitivity to feature rarity allows us to use our background beliefs about the probability of the evidence to evaluate hypotheses even in the light of normatively incomplete evidence. For example, imagine you have been asked to decide whether your sister's car, which possessed a top speed of over 165 mph and a radio, is a model X or a model Y. Given your background knowledge about the features, you can be more confident that the car is an X when told that 95% of Xs have a top speed of over 165 mph than when told that 95% of Xs have a radio. Thus information about feature rarity may be used to make a decision even when normatively complete evidence is missing.

As well as supporting inference with incomplete information, we believe that another candidate function for a rarity heuristic in hypothesis testing might be checking the limitations of hypotheses. Defining the scope of hypotheses in this way has recently become a topic of interest for cognitive psychologists. For example, Lopez (1995) has found that the majority of participants presented with a premise such as:

Dogs have a merocrine gland 1

and asked if they would prefer to find out whether wolves or bulls had a merocrine gland in order to check the more general premise that

All mammals have a merocrine gland 2

preferred to check bulls rather than wolves. This preference is viewed as being normatively correct as it obeys the notion that the more diverse is the evidence in favour of a hypothesis the stronger the support for that hypothesis is (see Carnap, 1951; Popper, 1959). Osherson, Smith, Wilkie,

Lopez and Shafir (1990) have proposed a model of category based induction which captures the diversity principle. This model accounts for people's preference for diverse premises by supposing that the strength of a categorical argument depends on the degree to which the premise categories are similar to both the conclusion category and instances of the lowest level category which includes both premise and conclusion categories.

There are situations in which it may be impossible to make the similarity calculations upon which Osherson et al's model relies. For example, it is common in the literature on category-based induction to use premises with blank predicates i.e. premises about which the participant is unlikely to have any a priori beliefs. This is done to minimize the effects of the predicate on participants' judgements. It is also possible to use blank premise categories in these experiments where the participant had no knowledge about the premise and conclusion categories except their size. In this case although the information required for a similarity calculation is unavailable one can check whether a general hypothesis also applies to a rare or unusual event. Thus we can greatly increase our confidence in the hypothesis (when it can account for the rare event) or limit the hypothesis (when it cannot).

The importance of such a limiting function may be seen when one considers that several lines of theoretical and experimental work suggest that it is the interaction between a heuristic or strategy and the environment in which it is used which determines the success or failure of that heuristic (e.g. Evans, Handley, Harper and Johnson-Laird, 1999; Gigerenzer et al, 1999). This argument was most explicitly made by Klayman and Ha (1987) who defined the probabilistic structure of environments where their positive test strategy would not be successful. The consequences of a mismatch between the environment and the positive test strategy is most dramatically illustrated by the failure of participants on Wason's 2 4 6 task to limit their initial rule thereby leading to a failure to discover the experimenter's more general rule.

The experiments described in this paper demonstrate the use of a heuristic which counteracts the effects of a positive test strategy. The standardly obtained finding on the pseudodiagnosticity task is that subjects tend to search for more information about the hypothesis supported by the existing evidence. In most cases this leads to pseudodiagnostic responding. In our experiments we have demonstrated that the tendency to select information about rare features produces diagnostic responding. In a similar fashion, one can imagine a scientist who is committed to a hypothesis that is too narrow but, because of the use of a positive test heuristic, is unable to find disconfirmation. The attempt to apply this hypothesis to a rare event or phenomena may provide the evidence required for the scientist to broaden the hypothesis.

Whilst we see the diversity principle (Osherson et al, 1990; Lopez, 1995; Spellman, Lopez and Smith, 1999) and a rarity heuristic as being complementary, there is an

important distinction to be drawn between them. Lopez (1995) has argued that the diversity principle does not breach the positive test heuristic. One of the adaptive functions of the rarity heuristic, on the other hand, is that - as in the experiments reported in this paper - it does violate a strategy based on positive testing. Its existence and use in everyday hypothesis testing is likely to be one reason why we are not surrounded by the calamitous results of a reliance on positive testing.

## Conclusion

In conclusion, we have demonstrated that the effects of rarity on hypothesis testing, although present, are insensitive to statistical and contextual manipulation. We have argued that these results support the existence of a Rarity heuristic in hypothesis testing. Finally, we claim that if such a heuristic does exist, it is likely, in many cases, to complement the operation of other heuristics known to operate when people select information to help them find out about the world.

## References

- Anderson, J.R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
- Carnap, R. (1951). *Logical foundations of probability*.
- Doherty, M.E., Mynatt, C.R., Tweney, R.D., & Schiavo, M.D. (1979). Pseudodiagnosticity. *Acta Psychologica*, 49, 111-121.
- Evans, J.St.B.T. (1989). *Bias in human reasoning: Causes and consequences*. Hove, UK: Lawrence Erlbaum Associates Ltd.
- Evans, J.St.B.T. & Over, D.E. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.
- Evans, J.St.B.T., Handley, S.J., Harper, C.N.J. & Johnson-Laird, P.N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25, 1495-1513.
- Feeney, A., Evans, J.St. B.T. & Clibbens, J. (1997). Probabilities, utilities and hypothesis testing. *Proceedings of the 19<sup>th</sup> Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.
- Feeney, A., Evans, J.St.B.T. & Clibbens, J. (in press). Background beliefs and evidence interpretation. *Thinking and Reasoning*.
- Feeney, A., Evans, J.St.B.T. & Venn, S. (2000) The effects of beliefs about the evidence on hypothesis testing. Unpublished manuscript, Department of Psychology, University of Durham.
- Gigerenzer, G., Todd, P. & the ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.
- Klayman, J. (1995). Varieties of confirmation bias. In J. Busemeyer, R. Hastie & Medin, D.L. (Eds.), *The psychology of learning and motivation*, Vol. 32: Decision making from a cognitive perspective (pp 385-419). San Diego: Academic Press.
- Klayman, J. & Ha, Y-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94, 211-228.
- Lopez, A. (1995). The diversity principle in the testing of arguments. *Memory and Cognition*, 23, 374-382.
- Mynatt, C.R., Doherty, M.E. & Dragan, W. (1993). Information relevance, working memory and the consideration of alternative. *Quarterly Journal of Experimental Psychology*, 46A, 759-778.
- Oaksford, M. & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608-631.
- Oaksford, M., Chater, N. & Larkin, B. (1999). Probabilistic effects in data selection. *Thinking and Reasoning*, 5, 193-243.
- Osherson, D.N., Smith, E.E., Wilkie, O., Lopez, A. & Shafir, E. (1990). Category-based induction. *Psychological Review*, 97, 185-200.
- Popper, K.R. (1959). *The logic of scientific discovery*. Hutchinson: London.
- Spellman, B.A., Lopez, A. & Smith, E.E. (1999). Hypothesis testing: Strategy selection for generalising versus limiting hypotheses. *Thinking and Reasoning*, 5, 67-91.
- Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Psychology*, 12, 129-140.
- Wason, P.C. (1966). Reasoning. In B.M. Foss (Ed.), *New horizons in psychology* (Vol. 1). Harmondsworth, UK: Penguin Books.

## Acknowledgements

This research was funded by grant R000222426 from the Economic and Social Research Council of the United Kingdom.

# Modeling Orientation Effects in Symmetry Detection: The Role of Visual Structure

Ronald W. Ferguson (r-ferguson@northwestern.edu)

Department of Computer Science, Northwestern University  
1890 Maple Avenue; Evanston, IL 60201, USA

## Abstract

Symmetry detection is a key part of human perception. One incompletely understood aspect of symmetry detection concerns *orientation effects*. The best-known orientation effect is the *preference for vertical symmetry*, where symmetry around a vertical axis is detected more quickly and accurately than symmetry at other orientations. Current symmetry detection models have difficulty explaining this effect. Using MAGI (Ferguson, 1994), we show how orientation effects may be caused by interactions between the perceived visual relations and the current reference frame. As evidence for this explanation, we simulate several orientation characteristics, including the preference for vertical symmetry and Wiser's (1981) theory of "intrinsic axes". Finally, we successfully simulate the results of a classic study by Palmer and Hemenway (1978) which explores the relationship between the preference for vertical symmetry, multiple symmetries, and inexact symmetry. Collectively, these results show that orientation effects may be due to characteristics of detected visual relations rather than either exact point-to-point equivalencies or the bilateral symmetry of the visual system.

## Introduction

Symmetry detection is a core mechanism in perception, shape recognition, and perceptual organization. Yet the processes underlying symmetry detection are only partially understood. Studies of symmetry detection have revealed psychological characteristics more complex than previously assumed even a few decades ago.

One such set of characteristics are *orientation effects*: interactions between symmetry detection and the visual reference frame. Orientation effects are interesting because they separate human performance in judging symmetry from symmetry's geometric definition. In geometric terms, symmetry is orientation-invariant, yet human symmetry detection depends critically on a figure's orientation. In addition, under certain circumstances symmetric figures also

influence the visual reference frame.

Orientation effects can be placed into three broad categories: the preference for vertical symmetry, the preference for multiple symmetries, and the effect of symmetry on a figure's object-centered reference frame.

**Preference for vertical symmetry.** Bilateral symmetry is more quickly and accurately detected when the symmetry axis is vertical (Attneave & Olson, 1967; Bornstein & Krinsky, 1985; Chipman & Mendelsohn, 1979; Corballis & Roldan, 1975; Goldmeier, 1936/1972; Julesz, 1971; Mach, 1893/1986; Palmer & Hemenway, 1978). In most cases, vertical symmetry is easier than horizontal symmetry, which in turn is easier than diagonal symmetry.

A longstanding explanation for the preference for vertical symmetry is that it depends on the human visual system's own vertically bilateral structure. In this framework, originally suggested by Mach (1893/1986), human vision provides better and faster results for symmetries aligned with its own symmetric structure. Several visual subsystems have been proposed as this effect's locus, from eye placement (Mach, 1893/1986) to the corpus collosum (Braitenberg, 1984; Herbert & Humphrey, 1996). However, most of these explanations focus on the retina and structures just beyond it (Corballis & Roldan, 1975; Jenkins, 1982; Julesz, 1971). Thus, these explanations are known as *retinocentric models*.

Retinocentric models, while theoretically elegant, fail to explain a key result: vertical symmetry is still preferred when the retina is *misaligned* with the symmetry axis. Rock and Leaman (1963) showed that the preference for vertical symmetry is still present when a figure is vertical with respect to the gravitational reference frame, but the subject's head is tilted 45° away from vertical.

**Symmetry in figures with intrinsic axes.** The preference for vertical symmetry disappears or is greatly attenuated for

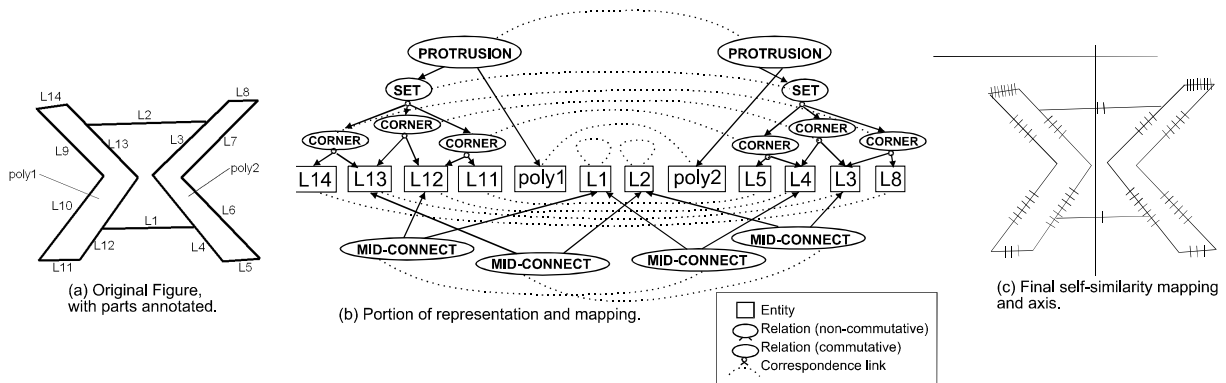


Figure 1: MAGI detects symmetry by aligning visual relations. Figure (a) shows a line drawing given to MAGI as a vector graphics file, with its vector elements labeled. Figure (b) shows a subset of the figure's visual relations (12 of 18 entities, 14 of 118 spatial relations) generated for those visual elements. Dotted lines indicate mapping links produced by MAGI. Note that two line segments, L1 and L2, map to themselves. Figure (c) indicates the full set of entity correspondences (using hash marks) and the axis produced by MAGI.

some kinds of figures. Figures with a good "intrinsic axis" (Palmer, 1983; Wiser, 1981) apparently impose their own reference frame, allowing recognition at any orientation.

**Preference for multiple symmetries.** Symmetry is also judged more quickly and accurately when a figure contains multiple symmetries (Royer, 1981; Wagemans, Van Gool, & d'Ydewalle, 1991). The preference for multiple symmetries is separate from the preference for vertical symmetry, and can produce additive results (Humphrey & Humphrey, 1989; Palmer & Hemenway, 1978).

Orientation effects pose significant challenges for cognitive models of symmetry detection, which have difficulty modeling interactions between symmetry detection and the visual reference frame. Some symmetry detection models, such as the so-called "brushfire" models (Blum & Nagel, 1978; Brady, 1983), do not use the reference frame at all. Other models use the reference frame in a limited sense – for example, utilizing it to find horizontally-aligned dots to link in symmetric dot patterns (Jenkins, 1983; Wagemans et al., 1991). These latter models can partially explain the preference for vertical symmetry by positing that some fixed set of orientations are tried until symmetry detection succeeds. At the same time, these models apply only to dot patterns, and cannot easily be extended to orientation effects found in more complex stimuli, such as polygons. More problematic, however, is that these models cannot explain how figures with good intrinsic axes eliminate the preference for vertical symmetry, nor why the order of preferences is first vertical, then horizontal, then diagonal (instead, it is typically assumed that this set of orientations results from either natural selection or perceptual learning in a world rich in vertically-symmetric objects). Finally, because these models assume a fixed orientation for each symmetry-detection attempt, and require exact symmetry, they have difficulty detecting even minor deviations from the assumed set of orientations (e.g., symmetry at a 38° angle).

A clue to resolving this quandary may be found in recent evidence that perceptual relations, such as connectivity relations and boundary characteristics, play a role in symmetry detection. Baylis and Driver (1994) provide evidence that symmetry detection in polygons may depend in part on curvature minima along figure boundaries. Ferguson, Aminoff & Gentner (1996) showed that specific qualitative differences, such as concavity or number-of-vertices mismatches, contributes to the speed and accuracy of symmetry judgments. Wagemans' bootstrap model (Wagemans et al., 1991) uses sets of conjoined "virtual quadrilaterals" to add higher-order structure to symmetric dot patterns, allowing the model to detect skewed symmetry.

If perceptual relations play a role in symmetry detection, they may be linked to orientation effects. Some have suggested (Goldmeier, 1936/1972; Rock, 1983) that the preference for vertical symmetry may be rooted in the phenomenological reversibility (or commutativity) of left-right spatial relations, which is not true of above-below relations. In other words, the preference for vertical symmetry is a product of how spatial relations, rather than symmetry-detection processes, depend on visual orientation. For our purposes, we term this the *horizontal commutativity conjecture*.

In this paper, we use MAGI (Ferguson, 1994; in preparation), our model of symmetry detection, to show why and how the horizontal commutativity conjecture may be true. The resulting explanation avoids at least three problematic assumptions of previous models: 1) that the symmetry detection process must use a set of fixed orientations; 2) that symmetry must be exact; or 3) that symmetry-detection is retinocentric.

This paper is arranged as follows. First, we briefly describe the MAGI model. Then, MAGI is used to explain the preference for vertical symmetry and the effect for intrinsic axes. We then perform an in-depth simulation of a classic study of the orientation effects for multiple and near symmetries (Palmer & Hemenway, 1978). We conclude by discussing the implications of these results, the model's limitations, and possible future research.

### The MAGI model of symmetry detection

The basis of the MAGI model (Figure 1) is that *symmetry is like analogy*. Specifically, symmetry may use the same cognitive processes found within other analogical reasoning such as analogy, similarity and memory access. As a result, symmetry may share the flexibility and domain-generalizability found in these other kinds of analogical reasoning.

MAGI models symmetry detection within the framework of structure mapping. MAGI creates a within-description mapping using the constraints of Structure Mapping Theory (Gentner, 1983) to align similar sets of relational structure. In other forms of analogical reasoning, such as similarity comparison and analogy, the mapping process aligns relations in base and target descriptions. In MAGI's symmetry detection, mapping is performed over a single relational description. MAGI also uses additional mapping constraints to maximize the self-similarity of the mapped portions.

For visual figures<sup>1</sup>, MAGI works directly from a vector-based line drawing. To obtain a description of the visual relations in the drawing, MAGI uses GeoRep (Ferguson & Forbus, 2000), a spatial representation engine. GeoRep represents visual relations detected early in perception, including element connectivity (such as corners and intersections), parallel elements, horizontally- and vertically-oriented structure, and protrusions and indentations in the figure boundary. MAGI then performs a self-similarity mapping over this relational description (Figure 1 shows an example of GeoRep's representation and MAGI's mapping).

MAGI's algorithm (see Ferguson, 1994, in preparation) is very similar to the Structure Mapping Engine (SME; Falkenhainer, Forbus, & Gentner, 1989; Forbus, Ferguson & Gentner, 1994). MAGI's self-similarity mappings are created using a local-to-global mapping process that enforces a set of six mapping constraints. Four of these constraints are adopted from SME: 1) the *tiered identity* constraint, which allows only expressions with identical predicates to align; 2) the *one-to-one mapping* constraint; 3) the *parallel*

---

<sup>1</sup> MAGI can also be used on non-visual stimuli, such as story narratives (Ferguson, 1994) or diagrams containing conceptual as well as visual regularity (Ferguson & Forbus, 1998). However, here we concentrate on visual symmetry alone.

*connectivity constraint*, which mandates that any aligned expression must also align its arguments; and 4) the *systematicity constraint*, which prefers large interconnected mappings with deep relational structure to smaller or unconnected mappings.

MAGI's final two constraints are specific to symmetry detection. The *limited self-matching* constraint states that an expression or entity may map to itself (i.e., self-match) only when it is the argument of an expression that is not a self-match. In Figure 1, this allows entity L1 to map to itself, because two separate *mid-connect* expressions involving L1 are aligned. The *maximal individuation* constraint encourages mappings that maximize the interconnectivity of each of the two mapped parts, and minimize the interconnectivity of the mapped parts with one another. In Figure 1, this constraint distributes the mapped *mid-connect* expressions to provide maximum entity overlap with other mapped expressions, such as the mapped *protrusion* expressions.

These constraints, as enforced by MAGI, produce one or more symmetry mappings. Each mapping contains a set of aligned entities and expressions and a systematicity score.

In MAGI, as in SME, systematicity is measured using a "trickle-down" structural evaluation mechanism (Forbus & Gentner, 1989). This mechanism gives higher scores to deeper expression matches and to matched entities with many matched superexpressions. For MAGI, this score is an approximate measure of "how symmetric" an object seems. For example, visualize a square and the X-shaped figure from Figure 1. Both figures have perfect geometric symmetry, but to MAGI, the X-shaped figure will have higher systematicity than the square because mapped expressions in the former are deeper and more interconnected than in the latter. Similar effects could be found even if we controlled for equivalent figure size and the number of segments.

A mapping also produces candidate inferences (as in SME) by carrying over unmapped structure that intersects mapped structure. Candidate inferences often indicate qualitative differences between the sides of the figure.

Once MAGI has found a self-similarity mapping, it uses the set of aligned entities to determine the axis. Using a Hough transform voting algorithm (Duda & Hart, 1987), MAGI produces either an axis or an object-centered reference frame for the mapping.

The nature of analogical mapping provides MAGI with a number of useful characteristics not found in other symmetry models. MAGI's symmetry detection is extremely

robust in the face of minor asymmetries and distracters. Symmetry mappings can also indicate qualitative differences between otherwise symmetric figures by producing candidate inferences. Finally, MAGI can link perceptual and conceptual symmetries in diagrams (Ferguson & Forbus, 1998), showing how self-similarity is utilized in perceptual reasoning tasks.

## Modeling the preference for vertical symmetry and intrinsic axes

Using the MAGI model, it is possible to test the horizontal commutativity conjecture. We begin by assuming that some visual relations are *orientation-dependent* (such as the *above* relations highlighted in Figure 2-A). Along with having orientation-dependent relations, we also can assume that vertically-oriented visual relations are directed, while horizontally-oriented relations are commutative. There is substantial evidence of just this dichotomy in human visual processing (Rock 1983). Humans often confuse left and right, but seldom confuse up and down.

Now we can see how mapping relational structure affects the produced mapping. Given (A), MAGI produces a vertical symmetry mapping. The vertical mapping is due to the alignment of many orientation-dependent visual relations, including the *above* relations. When the figure is rotated 45° (B) and then remapped, the set of orientation-dependent relations changes with it, and this affects the elements that MAGI aligns. Even though all the visual elements have moved relative to (A), MAGI's mapping of (B) is also vertical due to this new set of orientation-dependent relations. In other words, MAGI exhibits a preference for vertical symmetry.

Note that orientation-dependent visual relations do not dictate the mapping MAGI produces. Orientation-dependent relations are only part of the set of visual relations for any given figure, and for that reason, figures with sufficient visual structure can be mapped at many different orientations.

This explains why some figures may have good intrinsic axes that eliminate the preference for vertical symmetry. Figure 2-C shows MAGI's mapping of one of Wiser's (1981) example figures. Because the visual structure of this figure is distinctive enough to produce a symmetry mapping without orientation-dependent relations, this figure produces an axis at almost any orientation.

## How symmetry can adjust the frame of reference

This demonstration, however, only partially answers questions about the nature of orientation effects. If this model is correct, then how does the visual system detect symmetry in figures that neither have a good intrinsic axis nor are oriented vertically? Does the system have to try many different orientations, either serially or in parallel?

No, it doesn't. Instead, MAGI can use the initial partial mapping of a figure to find a potential new reference frame, and then shift the frame of reference to obtain a new representation of the figure. With this new representation, it can then reconstruct the symmetry of the figure as if it was presented in a vertical orientation.

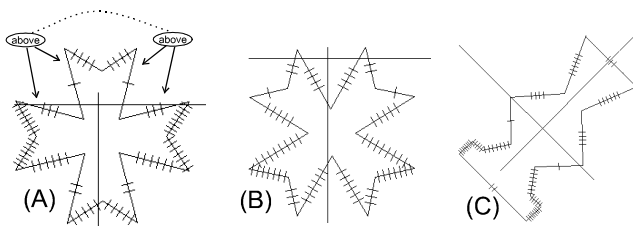


Figure 2: How orientation-dependent relations affect MAGI's symmetry mapping. Vertically-oriented relations in A and B enforce different mappings, even though the figures are identical. The preference for vertical symmetry can be overcome if there is sufficient structure when orientation-dependent relations are absent, as in (C), redrawn from Wiser (1981).

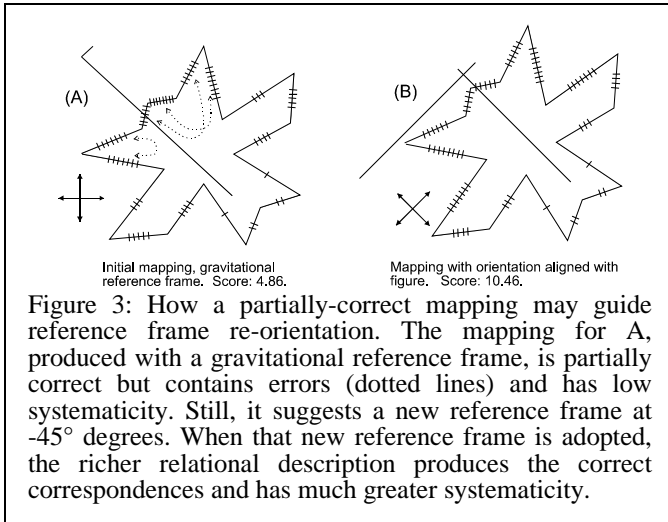


Figure 3: How a partially-correct mapping may guide reference frame re-orientation. The mapping for A, produced with a gravitational reference frame, is partially correct but contains errors (dotted lines) and has low systematicity. Still, it suggests a new reference frame at  $-45^\circ$  degrees. When that new reference frame is adopted, the richer relational description produces the correct correspondences and has much greater systematicity.

Figure 3 shows how this may occur. In the original figure (A), the mapping created by MAGI is only partial, and the resulting mapping has low systematicity and some incorrect correspondences. This is because the figure has insufficient visual structure to produce the correct mapping at this orientation (i.e., it does not have a good intrinsic axis). However, this partial mapping is sufficient to produce a potential new orientation for the figure, based on the parts of the mapping that do correspond. When the reference frame for the figure is set at this new orientation (B), the figure can be mapped as if it were at the vertical orientation, producing a richer set of orientation-dependent relations, and an axis is produced. In other words, the partial symmetry mapping tells the system to "tilt its head," and when it does so, it is rewarded by a set of visual relations that lead to a much richer symmetry mapping.

Although we do not yet have a theory of what mapping characteristics lead the viewer to re-orient the visual reference frame given a partial mapping (it may depend on several factors, including the task demands), clearly it is possible for the viewer to shift the reference frame using these clues. As a result, it is possible to see symmetry at an angle without presuming that the symmetry detection process must choose a set of orientations beforehand. One possible characteristic allowing a reference frame shift might be the systematicity of the initial mapping, a factor we return to in the next section.

### A Simulation in Depth

We now show the results of a simulation of an experiment (Palmer and Hemenway, 1978) testing both the preference for vertical symmetry and the effect of multiple symmetries.

Palmer and Hemenway's study used a set of 30 stimuli (Figure 4). The figures are 16-gons, containing five different symmetry types: single, double, and quadruple symmetry, rotational symmetry, and near symmetry. These figures were displayed at four different orientations: tilted left ( $-45^\circ$ ), vertical ( $0^\circ$ ), tilted right ( $+45^\circ$ ), and horizontal ( $+90^\circ$ ). In the first experiment, subjects had to judge whether the stimulus was mirror symmetric (requiring negative responses for rotational and near symmetry). Response latency and accuracy were measured.

	Quadruple	Double	Single	Near	Rotational
A					
B					
C					
D					
E					
F					

Figure 4: The stimuli used in Simulation 1 (redrawn from Palmer and Hemenway, 1978) arranged by symmetry type.

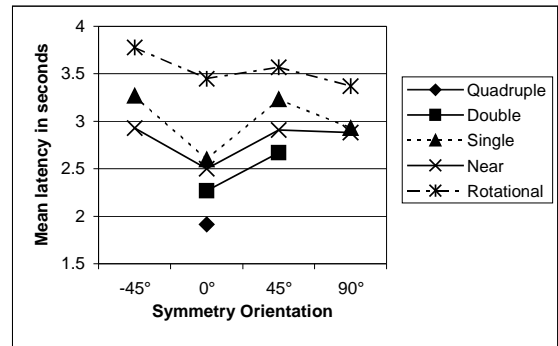


Figure 5: Palmer & Hemenway Experiment 1 results. Graph shows response time latency at four symmetry orientations. Redrawn from Palmer & Hemenway.

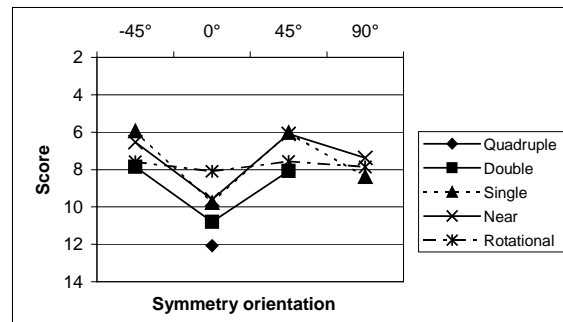


Figure 6: Results of simulating Experiment 1 using MAGI. Graph shows average systematicity score of each figure's best mapping (the Y-axis is inverted for easier comparison to Figure 5). Aside from the rotational symmetry results, MAGI duplicates the experimental results, with consistently higher systematicity scores for figures more quickly detected by human subjects.

Palmer and Hemenway's results (Figure 5) show a clear preference for vertical symmetry, with vertical better than horizontal, and horizontal better than diagonal (Figure 5 shows response latencies – accuracy results were similar). An effect was also found for multiple symmetries, with quadruple better than double, and double better than single symmetry.

For our simulation, the study's 30 stimulus figures were given to MAGI as line drawings. Each figure was presented at up to four orientations, as in the original study. We then used the systematicity score of MAGI's top mapping as a measure of the strength of the relational symmetry.

The results from MAGI are shown in Figure 6. With the exception of rotational symmetry, the results closely mirror those of Palmer and Hemenway, with vertical symmetry having the highest systematicity score, followed by horizontal symmetry and diagonal symmetry. Notably, these effects are reproduced separately for double, single, and near symmetries, as in the original study. MAGI's results also reproduce the effect for multiple symmetries, with quadruple symmetry producing the highest systematicity scores, followed by double symmetry, and then single and near symmetries. These latter two symmetry types produce roughly equal results, as in the original experiment.

The one difference between the two graphs are the results for rotational symmetry. For both MAGI and humans, rotational symmetry results varied only slightly with respect to orientation. However, while rotational figures showed the worst latencies for humans, the systematicity scores MAGI produced are average relative to the other symmetry types. One explanation for this difference, as noted in Palmer and Hemenway's analysis, is that in the original experiments subjects were to accept only mirror-symmetric figures, and thus had to reject rotationally symmetric figures. This means that the high latencies in the original experiment may not be due to a low sense of the figures' symmetry, but because subjects' needed to avoid that sense to produce a negative response. MAGI was not constrained to judge only mirror symmetry, and so frequently found rotational mappings.

We briefly note a second result. In a second experiment, Palmer and Hemenway showed subjects the same 30 figures solely in the vertical orientation, meaning that subjects no longer had to look for symmetry at multiple orientations. This had the effect of greatly decreasing the average response latencies (from a mean of 2626 ms. to 1111 ms.). While accuracy and response time results for quadruple, double, and single symmetry maintained their previous ordering, the error rate for near symmetry shot up from 1.4% to 16.7% from the first to the second experiment, an error rate more than twice the rate for any other symmetry type, while the error rate for rotational symmetry decreased.

The MAGI model suggests a possible explanation. Because the experiment's demand characteristics reduced response time, and because only vertical symmetry was used, it would no longer be necessary to consider partial mappings as indicators of alternative symmetry orientations. Simpler factors, such as the lack of candidate inferences (indicating qualitative asymmetry) might suffice. This strategy is not problematic for quadruple, double, or single symmetries, since exact symmetries do not produce

candidate inferences. Nor is it a problem for rotational symmetries, which always produce candidate inferences. However, near-symmetric figures produce few or no candidate inferences in MAGI. When MAGI was run on the near-symmetric figures, each figure only produced a few candidate inferences and one (in Figure 4's row E) produced none. The relative scarcity of candidates inferences may have made asymmetry detection difficult for near-symmetric figures and lead to subjects' high error rate.

## Conclusion

These results demonstrate that a structure-mapping model of symmetry detection can concisely explain orientation effects using a few simple assumptions: 1) that visual structure is at least partially orientation-dependent; 2) symmetry detection is performed by mapping visual structure; and 3) partial mappings are used to find potential mappings and suggest alternate frames of reference. Using this simple model, we simulated the preference for vertical symmetry, showing that the preference for vertical over horizontal symmetry, and for both over diagonal symmetry, was not the result of a pre-established list of potential orientations, but the natural result of a visual system where vertically-oriented relations are phenomenological different than horizontally-oriented relations (the horizontal commutativity conjecture). Similarly, we showed that the preference for multiple symmetries could be modeled with the same assumptions. We showed the correctness of this model by running it on the stimuli of Palmer & Hemenway (1978), which tested both of these effects, and MAGI reproduced the same general pattern of results. Finally, we showed why some figures with good "intrinsic axes" (Palmer, 1983; Wisner, 1981) do not show the same preference for vertical symmetry (an explanation currently beyond the capabilities of other models of symmetry detection). This defined conditions when the sense of symmetry is strong enough to overcome effects of orientation. These collective results suggest that a structure-mapping model of symmetry detection, such as MAGI, could provide a better analysis of a wide variety of symmetry-related phenomena.

There are several limitations with the current model, however. Because the relational mapping depends on the visual relations found in the figure, representation assumptions can drastically change MAGI's results. In the current study, we have attempted to minimize this effect by using GeoRep's default representation engine, which builds a set of relations based on the visual relations assumed to be built by Ullman's universal visual routines (Ullman, 1984). However, further research is needed to test the reliability of these assumptions. MAGI's dependence on spatial relations leaves open the question of exactly when quantitative differences (such as small differences in the angles of corresponding corners) are detected. When such differences exist, but these differences are not qualitative, MAGI does not detect them. Other limitations of GeoRep and MAGI precluded other possible simulations. Because GeoRep does not have a model of grouping, it was not possible to model orientation effects based on grouped items (Palmer, 1983).



This research also creates interesting new questions. The effect for multiple symmetries bears closer analysis. Initial results suggest that the effect is a result of the greater number of visual relations found in figures with multiple symmetries, as well as the greater systematicity of systems with many similar subparts. However, this result should be tested in another domain.

### Acknowledgments

This research was supported by the Cognitive Science and Computer Science programs of the Office of Naval Research, by DARPA's High-Performance Knowledge Bases initiative, and by the National Science Foundation under the Learning and Intelligent Systems program. Sincere thanks go to Dedre Gentner, Ken Forbus, Laura Allender, Alex Aminoff, Steve Palmer and an anonymous reviewer for useful feedback.

### References

- Atneave, F., & Olson, R. K. (1967). Discriminability of stimuli varying in physical and retinal orientation. *Journal of Experimental Psychology*, 74(2), 149-157.
- Baylis, G. C., & Driver, J. (1994). Parallel computation of symmetry but not repetition within single visual shapes. *Visual Cognition*, 1, 377-400.
- Blum, H., & Nagel, R. N. (1978). Shape description using weighted symmetric axis features. *Pattern Recognition*, 10.
- Bornstein, M. H., & Krinsky, S. J. (1985). Perception of symmetry in infancy: The salience of vertical symmetry and the perception of pattern wholes. *Journal of Experimental Child Psychology*, 39, 1-19.
- Brady, M. (1983). Criteria for representation of shape. In J. Beck, B. Hope, & A. Rosenfeld (Eds.), *Human and Machine Vision* (pp. 39-84). New York: Academic Press.
- Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA: MIT Press.
- Chipman, S. F., & Mendelsohn, M. J. (1979). Influence of six types of visual structure on complexity judgments in children and adults. *JEP:HPP*, 5, 365-378.
- Corballis, M.C., & Roldan, C.E. (1975). Detection of symmetry as a function of angular orientation. *JEP:HPP*, 1, 221-230.
- Duda, R. O., & Hart, P. E. (1987). Use of the Hough Transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1), 11-15.
- Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The Structure-Mapping Engine: Algorithm and examples. *Artificial Intelligence*, 41, 1-63.
- Ferguson, R. W. (1994). MAGI: Analogy-based encoding using symmetry and regularity. *Proceedings of 16th Cognitive Science Conference*. Atlanta, GA.
- Ferguson, R. W. (in preparation). MAGI: A model of symmetry and repetition detection. .
- Ferguson, R. W., Aminoff, A., & Gentner, D. (1996). Modeling qualitative differences in symmetry judgments, *Proceedings of 18th Cognitive Science Conference*.
- Ferguson, R. W., & Forbus, K. D. (1998). Telling juxtapositions: Using repetition and alignable difference in diagram understanding. In K. Holyoak, D. Gentner, & B. Kokinov (Eds.), *Advances in Analogy Research*. Sofia, Bulgaria: New Bulgarian University.
- Ferguson, R. W., & Forbus, K. D. (2000). GeoRep: A flexible tool for spatial representation of line drawings, *Proceedings of the 18th National Conference on Artificial Intelligence*. Austin, Texas: AAAI Press.
- Incremental structure mapping. In *Proceedings of 16th Cognitive Science Conference*. Atlanta, GA.
- Forbus, K. D., & Gentner, D. (1989). Structural evaluation of analogies: What counts? In *Proceedings of 11th Cognitive Science Conference*.
- Gentner, D. (1983). Structure-Mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.
- Goldmeier, E. (1936/1972). Similarity in visually perceived forms. *Psychological Issues*, 8(1), 14-133.
- Herbert, A. M., & Humphrey, K. G. (1996). Bilateral symmetry detection: Testing a 'callosal' hypothesis. *Perception*, 25, 463-480.
- Humphrey, G. K., & Humphrey, D. E. (1989). The role of structure in infant visual pattern perception. *Canadian Journal of Psychology*, 43(2), 165-182.
- Jenkins, B. (1982). Redundancy in the perception of bilateral symmetry in dot textures. *Perception & Psychophysics*, 32(2), 171-177.
- Jenkins, B. (1983). Component processes in the perception of bilaterally symmetric dot textures. *Perception & Psychophysics*, 34(5), 171-177.
- Julesz, B. (1971). *Foundations of Cyclopean Perception*. Chicago, IL: University of Chicago Press.
- Mach, E. (1893/1986). On symmetry, In *Popular Scientific Lectures*. LaSalle: Open Court Publishing.
- Palmer, S. E. (1983). The psychology of perceptual organization. In J. Beck, B. Hope, & A. Rosenfeld (Eds.), *Human and Machine Vision*. New York: Academic Press.
- Palmer, S. E., & Hemenway, K. (1978). Orientation and symmetry: Effects of multiple, rotational, and near symmetries. *JEP:HPP*, 4, 691-702.
- Quinn, P. C. (1994). The categorization of above and below spatial relations by young infants. *Child Development*, 65, 58-69.
- Rock, I. (1983). *The Logic of Perception*. Cambridge, MA: The MIT Press.
- Rock, I., & Leaman, R. (1963). An experimental analysis of visual symmetry. *Acta Psychologica*, 21, 171-183.
- Royer, F. L. (1981). Detection of symmetry. *JEP:HPP*, 7(6), 1186-1210.
- Ullman, S. (1984). Visual routines. *Cognition*, 18(1-3), 97-159.
- Wagemans, J., Van Gool, L., & d'Ydewalle, G. (1991). Detection of symmetry in tachistoscopically presented dot patterns. *Perception & Psychophysics*, 50, 413-427.
- Wiser, M. (1981). The role of intrinsic axes in shape recognition, *Proceedings of 3rd Cognitive Science Conference* (pp. 184-186). Berkeley, CA.

## Visual Learning for a Mid Level Pattern Discrimination Task

**I. Fine** (fine@salk.edu)

Department of Psychology; University of California, San Diego, 9500 Gilman Drive  
La Jolla, CA 92093-0109, USA

**Robert A. Jacobs** (robbie@bcs.rochester.edu)

Department of Brain and Cognitive Sciences, University of Rochester  
Rochester, NY 14627, USA

### Abstract

Our goal was to examine the plasticity of the human visual system at mid to high levels of visual processing. It is well understood that early stages of visual processing contain cells tuned for spatial frequency and orientation. However images of real-world objects contain a wide range of spatial frequencies and orientations. We were interested in how different spatial frequencies and orientations are combined. We used a pattern discrimination task - observers were asked to discriminate small changes in a “wicker-like” stimulus consisting of six superimposed sinusoidal gratings. Observers were asked to discriminate a 15% spatial frequency shift in two of these sinusoidal components, which were masked by four noise components. We found large amounts of perceptual learning for this task – over eight sessions of training observers’ average percent correct increased by 31%, corresponding to their thresholds dropping to a third of their initial values. Further experiments suggest that learning was based on changes within a mid level stage of processing intermediate between low-level analyzers tuned for orientation and spatial frequency and high-level pattern matching or object tuned cells. This mid level stage seems to be “very roughly Fourier” and combines information from individual gratings using probability summation. This stage of processing is also remarkably plastic compared to earlier stages of processing.

### Introduction

A great deal is known about low level visual pattern analyzers and their role in visual perception. At early stages of processing retinal input is represented by low level analyzers tuned for spatial frequency and orientation with receptive fields of limited spatial extent - properties very similar to simple cells in V1 (see Graham, 1989 for a review). However images of real-world objects contain a wide range of Fourier components, and therefore the combination of information across these low level analyzers is necessary to reliably recognize objects. Evidence suggests that there may be mid level mechanisms selectively pooling information across low level analyzers tuned for a wide range of spatial frequencies or orientations (e.g. Georgeson, 1992; Derrington & Henning, 1989; Burr & Morrone, 1994; Graham & Sutter, 1998; Olzak & Thomas, 1999).

It has been argued that relatively early stages of the visual system (V1) change with training (e.g. Ball & Sekuler, 1987; Fahle & Edelman, 1992; Sagi & Tanne,

1994; Ahissar & Hochstein, 1995,1996; Saarinen & Levi, 1995; Fahle & Morgan, 1996; Schoups & Orban, 1995). In addition, some learning effects have been noted (Olzak, personal communication, 1995; Fiorentini & Berardi, 1981) for tasks involving compound grating discriminations thought to involve mid level mechanisms.

The following experiments provide support for the existence of mid-level mechanisms pooling over analyzers tuned for spatial frequency and orientation. These mid level mechanisms are shown to be far more adaptable as a function of experience than low level analyzers.

### Experiment 1

The purpose of Experiment 1 was to measure learning for a complex “wicker” stimulus that required observers to combine information over a wide range of spatial frequencies and orientations.

### Methods

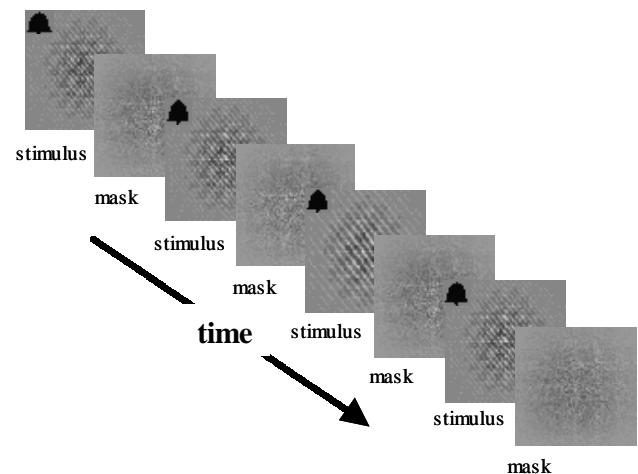


Figure 1: Diagram of the task used in the experiment.

Five observers were asked to perform a four alternative forced choice discrimination task (Figure 1). Four stimuli were presented sequentially in time. A two-dimensional white noise pattern was presented after each stimulus to reduce afterimage interference. Observers were asked to

indicate which of the four stimuli was different from the others using a key press. There are two important advantages of this four alternative forced choice procedure. First, the chance success rate was 25%, thereby providing more information per trial than a two alternative forced choice task. Second, such a task allows a same-different judgment without potential criterion effects (observers showing a bias towards responding same or different). The task was carried out using free-fixation. Observers were given auditory feedback and were self paced. Observers completed eight sessions, and completed 250 trials per session. Observers never carried out more than a single session in a day, and carried out three to five sessions a week.

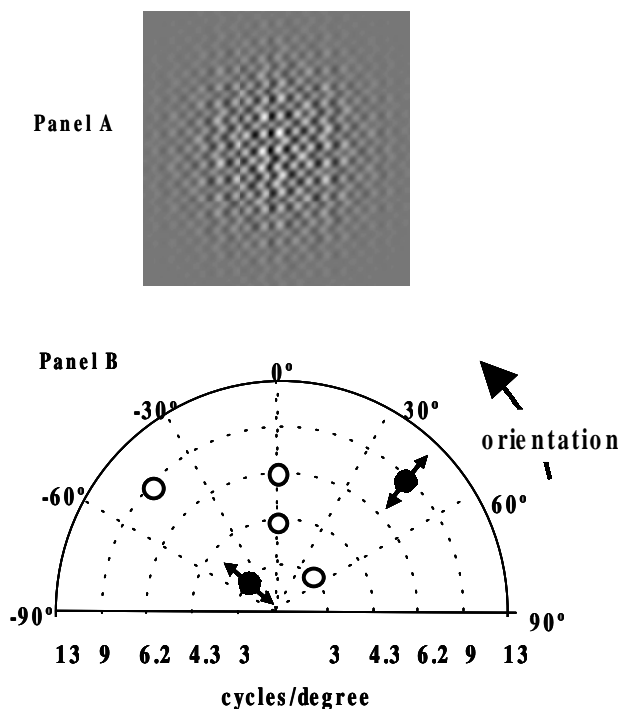


Figure 2: Panel A. Illustration of a typical stimulus. Panel B. Fourier representation of stimuli. The radius represents spatial frequency and the angle represents orientation. One signal component had a spatial frequency of 2.55 or 3.45 cycles/degree, an orientation of  $-45^\circ$  and contrasts varying between 1.6-12.8%. The other signal component had a spatial frequency of 7.65 or 10.35 cycles/degree, an orientation of  $45^\circ$  and contrasts varying between 5.5-44%. There were four sinusoidal noise components, represented by empty circles: 1) spatial frequency of 9 cycles/degree,  $-45^\circ$  orientation, 11% contrast 2) spatial frequency of 3 cycles/degree,  $45^\circ$  orientation, 3.2% contrast 3) spatial frequency of 4.3 cycles/degree,  $0^\circ$  orientation, 7.1% contrast 4) spatial frequency of 6.2 cycles/degree,  $0^\circ$  orientation, 7.1% contrast.

Figure 2 Panel A shows what a typical stimulus looked like. Each stimulus contained two signal compo-

nents and four sinusoidal noise components. Figure 2, Panel B represents the stimuli in Fourier space using polar coordinates. The radius represents spatial frequency and the angle represents orientation. The black filled circles represent the two possible signal components. These signal components were widely separated in orientation (at least  $90^\circ$  to each other) and widely separated in spatial frequency (approximately two octaves apart). One signal component was centered on 3 cycles/degree and had an orientation of  $-45^\circ$  and the other signal component was centered on 9 cycles/degree and had an orientation of  $45^\circ$ . Observers were asked to detect a 15% shift in the spatial frequency of the signal components, represented by the black arrows. The contrasts of the signal components were manipulated (based on pilot data) so each observer was presented with a range of difficulty levels. The empty circles in Figure 2 represent the four sinusoidal noise components that were added to the stimulus.

Stimuli were modulated spatially by a two dimensional Gaussian envelope with a sigma of 0.5693 degrees and temporally by a Gaussian envelope of sigma 0.237 seconds centered within a 0.67 second temporal window. The phases of the sinusoidal noise components were varied randomly across each interval of each trial. The phases of the signal components were varied randomly between each trial, and remained constant across the four intervals within each trial. Stimuli were presented using the VideoToolbox and Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997).

Observers were undergraduate or graduate students from the University of Rochester, varying in age between 19-28 years of age. Observers had normal or corrected to normal vision. Further details of this experimental procedure are described in Fine & Jacobs (2000).

## Results and Conclusions

The black squares in Figure 3 show the percent correct as a function of session averaged across observers in Experiment 1 (black squares). All five observers showed a significant improvement in their performance over eight sessions. Observers' average percent correct increased by 31%, corresponding to a two-third decrease in their thresholds.

Most perceptual learning studies have been carried out using simple stimuli (grating discrimination or Vernier tasks). Learning effects for these low level tasks tend to be small or non-existent in the fovea (e.g. Fiorentini and Berardy, 1981, Beard, Levi and Reich, 1995). In contrast, we found large learning effects in the fovea, suggesting strongly that our task is mediated by a higher stage of processing than more simple tasks, and that this stage of processing is far more plastic than earlier stages.

These improvements in performance with practice were relatively long lasting, none of the observers showed any

decline in performance when retested more than a month after training.

It is worth noting that observers showed faster improvement for easier stimuli than for more difficult stimuli, suggesting possible bootstrapping from easy to difficult stimuli (Ahissar & Hochstein, 1997).

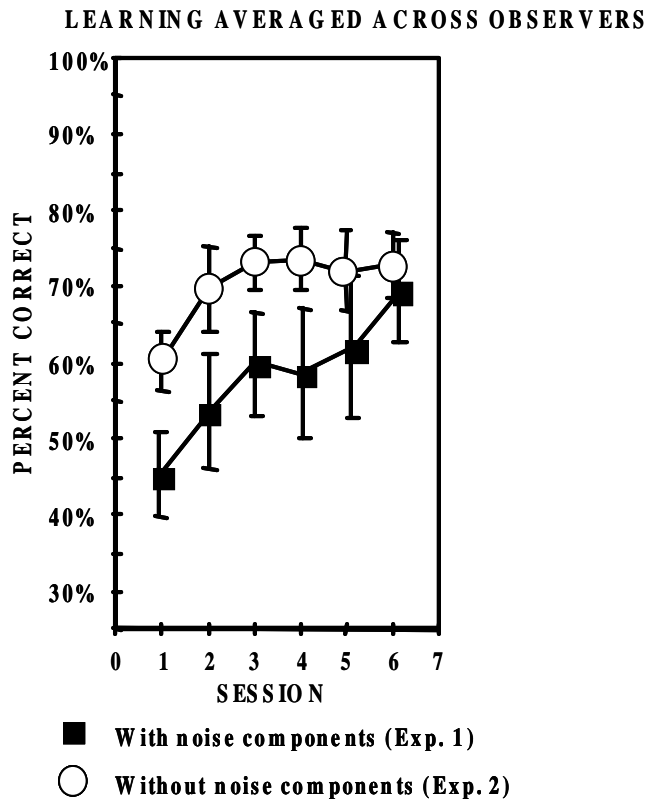


Figure 3: Percent correct as a function of session averaged across observers with the sinusoidal noise components (Experiment 1 - black squares) and without the noise components (Experiment 2 - empty circles). The x-axis shows the session and the y-axis shows the percent correct. Standard error bars are shown.

## Experiment 2

The extent of learning found in Experiment 1 suggests that performance in our task might be mediated by a mid level stage of processing rather than earlier stages. Experiment 2 was designed to exclude the possibilities that the learning found in Experiment 1 was due either to learning in low level mechanisms, or to improved non-visual cognitive strategies (such as learning the key press procedure, learning to fixate, learning the temporal structure of the task, etc.).

If the learning demonstrated in Experiment 1 was due to tuning changes within low level analyzers tuned for both spatial frequency and orientation then removing the sinusoidal noise components would not affect the amount of learning shown. The sinusoidal noise components in Experiment 1 were positioned so as to be invisible to

analyzers tuned for the spatial frequency and orientation of the signal components (see Figure 2, Panel B) - every noise component differed from the signal components by at least 45 degrees of orientation or almost two octaves of spatial frequency. Estimates of the tuning of low level analyzers by other authors predict little low level masking between sinusoidal components separated by either two octaves of spatial frequency or 45 degrees orientation (Graham, 1989).

If the learning in Experiment 1 was due to non-visual cognitive strategies then we would expect an equal amount of learning in Experiments 1 and 2 - the only difference between the two experiments was in the visual stimulus.

## Methods

Display and task were identical to those used in Experiment 1. Only the stimulus differed in Experiment 2, in that the sinusoidal noise components (the empty circles of Figure 2) were no longer present - i.e. observers were asked to discriminate changes in spatial frequency within a simple plaid pattern.

Without the noise components the task would be trivially easy for the contrast levels and spatial frequency shifts used in Experiment 1. The difficulty of the task was adjusted by reducing the spatial frequency shift to between  $\pm 2.5\%$  and  $\pm 12.5\%$  (as opposed to 15% in Experiment 1) to avoid ceiling effects.

Three observers were given six sessions of training on the task.

## Results and Conclusions

As shown by the empty circles in Figure 3, observers showed much less learning without the sinusoidal noise components. Observers showed some learning between sessions 1 and 2, but little learning after the second day. There was no significant drop in threshold across the three observers.

Differences in the amount of learning between Experiment 1 and 2 cannot be explained by ceiling effects. Initial performance was closely matched for the majority of subjects. In Experiment 1 three of the five observers in performed between 50-60% correct in the first session. In Experiment 2 two of the three observers performed between 50-60% correct in the first session. In addition, none of the observers' performance reached 90% correct by the end of training in either experiment.

There was some learning ( $\sim 7\%$ ) between the first and second day in both Experiment 1 and Experiment 2. Given that we used naïve observers we think it likely that these learning effects are mainly due to non-visual factors - learning the key press procedure etc. However an alternative possibility is that this learning between the first and second day was due to learning in low level analyzers.

In any case, most of the learning shown in Experiment 1 was *after* the second session and cannot be due either to learning in low level analyzers or to learning better non-visual cognitive strategies.

### Experiment 3

We were interested in how observers might be combining information from the two signal components. We have found that observers' performance in Experiment 1 can be well described using an independent probability summation model where observers correctly discriminate the "odd man out" if they detect a shift in either component (Fine & Jacobs, 2000). Experiment 3 was designed to further test whether observers' ability to combine information could be reasonably approximated using an independent probability summation model.

Other possible combination models include non-independent combination of information from the two signal components (as suggested by Olzak and Thomas, 1999) or some type of "pattern" or "template" matching.

The task carried out in Experiment 1 can be subdivided into two tasks, as shown in Figure 4. In the *same sign* task the "odd man out" was distinguished from the distracting stimuli by both signal components being shifted in the same direction in Fourier space. In half the trials both signal components were shifted higher in spatial frequency, as shown in Figure 4 Panel A. In the other half of the trials both signal components were shifted lower in spatial frequency.

In the *opposite sign* task the "odd man out" was distinguished from the distracting stimuli by both signal components being shifted in opposite directions in Fourier space. In half the trials the high spatial frequency component was shifted higher in spatial frequency, and the low spatial frequency component was shifted lower. In the other half of the trials, as shown in Figure 4 Panel B, the high spatial frequency component was shifted lower in spatial frequency, and the low spatial frequency component was shifted higher.

Independent probability summation implies that detecting a shift in the low spatial frequency signal component is unaffected by the direction of the shift in the high spatial frequency signal component, and vice versa. Any relationship between the directions of the spatial frequency shifts within the two signal components would be invisible to such a mechanism. Consequently, according to an independent probability summation model we would expect perfect transfer of learning from same sign to opposite sign tasks.

According to most non-independent models, including pattern matching, one would expect incomplete transfer between the two stimuli.

### Methods

Display and task were identical to those used in Experiment 1, however observers were either exclusively trained with same sign stimuli, then tested with opposite sign stimuli, or were trained with opposite sign stimuli, then tested with same sign stimuli. Four observers were tested in all, two were trained with same sign stimuli and two were trained with opposite sign stimuli. Observers were given six sessions of training before being tested with the novel stimuli.

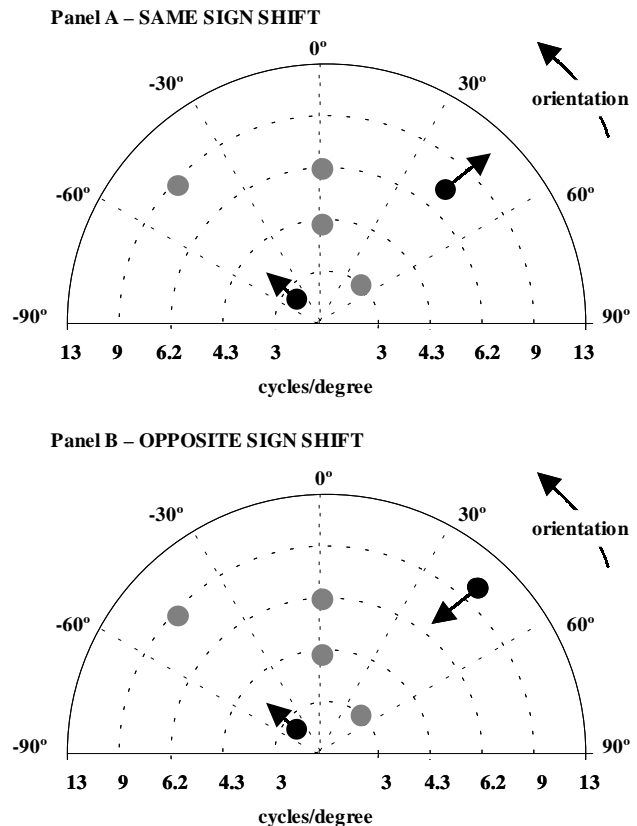


Figure 4: Polar plot of the stimuli used for the transfer of task experiment. Panel A shows the same sign shift stimulus and Panel B shows the opposite sign shift stimulus.

### Results and Conclusions

None of the four observers showed any drop in performance when tested with the novel stimulus. Interestingly only one of the four observers even noticed that the stimulus had changed. This perfect transfer of learning between same and opposite sign tasks is consistent with observers combining information independently, and is incompatible with most non-independent models (Olzak & Thomas, 1999), including pattern matching.

Interestingly, the shift in the signal components in the same sign task is compatible with a change of scale (as if both signal components moved closer or further away from the observer), while the shift in the signal components in the opposite sign task is compatible with a change in shape. The total transfer of learning between the two tasks suggests that "scale-invariance" may not yet be differentially encoded at this stage of processing.

### General Conclusions

Our data support the existence of a mid level stage of processing intermediate between low and high levels of visual processing. This level of processing seems to be

“very roughly Fourier” in that it still represents stimuli in terms of their spatial frequency and orientation. Information from low level analyzers tuned for spatial frequency and orientation seems to be combined using probability summation. This stage may be responsible for beginning to selectively process information, extracting the combinations of spatial frequency and orientation that define meaningful objects. As our knowledge of the mechanisms underlying mid level visual tasks increases it should be possible to ask increasingly refined questions about the role of these mid level mechanisms, and in particular, the role adaptability plays in allowing such mechanisms to represent an unpredictable world. Interestingly, our studies show that this mid level stage of processing seems to be far more plastic than earlier stages.

As cells become more specific in what they represent, an increasing number of cells become necessary if all possible stimuli are to be represented. This is the paradox of the “grandmother cell” – not every possible object can have its own feature detectors in the brain without a prohibitive number of cells. Despite this apparent paradox, cells in the brain have been shown to be remarkably specific (e.g. Desimone, Albright, Gross, & Bruce, 1984; Logothetis, Pauls, Poggio, 1995). Neural plasticity may be a way of alleviating the trade-off between cell specificity and limited cell numbers. By dynamically changing neural representations as a function of experience cells can be preferentially allocated to represent behaviorally important stimuli. If this is the case, then we should find an intimate relationship between plasticity and specificity - as representations become more selective, they should also become more plastic.

### Acknowledgments

This work was supported by NIH grants R29-MH54770, P30-EY01319 and EY 01711. We would like to thank R. Aslin, G. Boynton, K. Dobkins, N. Graham, D. MacLeod, W. Makous, L. Olzak and two anonymous reviewers for their advice and comments, and E. Bero, L. O’Brian, A. Pauls and M. Saran for help conducting the experiments.

### References

- Ahissar, M., & Hochstein, S. (1996). Learning pop-out detection: Specificities to stimulus characteristics. *Vision Research*, 36(21), 3487-3500.
- Ahissar, M., Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, 387, 401-6
- Ball, K., & Sekuler, R. (1987). Direction-specific improvement in motion discrimination. *Vision Research*, 27(6), 953-965.
- Beard, B.L., Levi, D.L., & Reich, L.N. (1995). Perceptual learning in parafoveal vision. *Vision Research*, 35(12), 1679-1691.
- Brainard, D. H. (1997) The Psychophysics Toolbox, *Spatial Vision*, 10:443-446.
- Burr, D.C., & Morrone, M.C. (1994). The role of features in structuring visual images. In M. Morgan (Ed.), *Higher-Order Processing in the Visual System*. New York, New York: Wiley.
- Derrington, A.M., & Henning, G.B. (1989). Some observations on the masking effects of two-dimensional stimuli. *Vision Research*, 29(2), 241-246.
- Fahle, M., & Edelman, S. (1992). Long-term learning in vernier acuity: Effects of stimulus orientation, range and of feedback. *Vision Research*, 33(3), 397-412.
- Fahle, M., & Morgan, M. (1996). No transfer of perceptual learning between similar stimuli in the same retinal position. *Current Biology*, 6(3), 292-297.
- Fine, I., & Jacobs, R.A. (2000). Perceptual learning for a pattern discrimination task. *Submitted to Vision Research*.
- Fiorentini, A., & Berardi, N. (1981). Learning in grating waveform discrimination: Specificity for orientation and spatial frequency. *Vision Research*, 21(7), 1149-1158.
- Georgeson, M.A. (1992). Human vision combines oriented filters to compute edges. *Proceedings of the Royal Society of London*, 249B, 235-245.
- Graham, N. (1989). *Visual pattern analysers*. Oxford, UK: Oxford Science Publications.
- Graham, N., & Sutter, A. (1998). Spatial summation in simple (Fourier) and complex (non-Fourier) channels in texture segregation. *Vision Research*, 38(2), 231-257.
- Karni, A., & Sagi, D. (1991). Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences of the United States of America*, 88, 4966-4970.
- Liu, Z. & Weinshall, D. (1999). Mechanisms of generalization in perceptual learning. *Advances in Neural Information Processing Systems*, MIT Press.
- Logothetis, N.K., Pauls, J. & Poggio, T. (1995) Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552-63.
- Olzak, L.A., & Thomas, J.P. (1999). Neural recoding in human pattern vision: Model and mechanisms. *Vision Research*, 39(2), 231-256.
- Pelli, D. G. (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies, *Spatial Vision* 10:437-442.
- Saarinen, J., & Levi, D.M. (1995). Perceptual learning in vernier acuity: What is learned? *Vision Research*, 35(4), 519-527.
- Sagi, D., & Tanne, D. (1994). Perceptual learning: Learning to see. *Current Opinion in Neurobiology*, 4(2), 195-199.
- Schoups, A.A., Vogels, R., & Orban, G.A. (1995). Human perceptual learning in identifying the oblique orientation: retinotopy, orientation specificity and monocularly. *Journal of Physiology*, 483, 797-810.

# Decoding Syntactic Parameters: The Superparser as Oracle

**Janet Dean Fodor (jfodor@gc.cuny.edu)**

Ph.D. Program in Linguistics; CUNY Graduate Center  
365 Fifth Avenue, New York, NY 10016 USA

**Virginia Teller (teller@cs.hunter.cuny.edu)**

Department of Computer Science; Hunter College CUNY  
695 Park Avenue, New York, NY 10021 USA

## Abstract

Syntactic parameter setting has proven extremely difficult to model. The original 'switch-setting' metaphor failed because parametrically relevant properties of a natural language sentence cannot be recognized without considerable structural analysis. The result has been a move to trial-and-error learners which attempt to guess a grammar that can analyze (parse) the current input sentence. But standard variants of grammar guessing are wasteful of the parametric information in input sentences because they use it only as feedback after a candidate grammar has been chosen. We show here that performance is significantly improved by a 'superparsing' routine which constructs a candidate grammar on-line in response to the properties of the input sentence. No sentences then need to be discarded for lack of a grammar to parse them. The gain in learning speed can be quantified in terms of the average number of sentences required for convergence. Superparsing can be achieved by the normal sentence parsing routines, applying a grammar that incorporates all possible parameter values. The superparsing learner is robust and imposes no special demands on the input.

## Natural Language Acquisition

Children exposed to a sample of sentences from a natural language acquire its grammar in a few years. There is as yet no computational model of the acquisition process that is both effective and psychologically realistic. The conception of the learner's task was greatly simplified with the advent of parameter theory (Chomsky, 1981, 1995) under which a natural language grammar consists of an innate component (Universal Grammar, UG) and a selection from among a finite set of properties by which languages can differ (the parameters). Depending on the particular linguistic theory assumed, a parameter might be a choice between grammar rules, or between the presence or absence of a rule in the grammar. But as developed in the Chomskyan framework a parameter specifies a more abstract property of grammatical derivations, such as the direction of case assignment by a verb, or the derivational level at which a universal constraint applies, or, in more recent versions, the 'strength' of a syntactic feature on a tree node. The learner's task is to select the correct setting for each relevant parameter, i.e., each parameter whose value contributes to the derivation of at least one sentence of the target language. Though the concept is simple, the task is more challenging than was originally appreciated.

The earliest idea was that some property of an input sentence would trigger the correct setting of each parametric switch. But the needed property detectors could not be devised, because the characteristic properties of sentences derived by means of some parameter value  $v$  (inter alia) are

not sufficiently uniform and superficially identifiable (Clark, 1994; Gibson & Wexler, 1994). To find the abstract properties that identify  $v$ , the derivation of the sentence must be computed, i.e., the sentence must be parsed. But parsing is work, and the computational workload of a learner must be kept within psychologically plausible limits. The problem is that when the parser lacks the correct grammar, as it does by definition in a learning event, it must apparently try out multiple grammars until it finds a successful one. Even though the number of parameters is limited, there are too many possible grammars for it to be feasible for a learner to try them all on a single sentence, either serially or in parallel.

This is where current learning models diverge. Some test one grammar per sentence (e.g. Gibson & Wexler, 1994, henceforth GW94) and are very slow to converge on the correct grammar. Others test batches of grammars at a time (e.g. Clark, 1992; Nyberg, 1992) and thereby go beyond what it is plausible to suppose a child is capable of. In this paper we discuss a novel way to use the parser to decode the parameter values that license an input sentence, without undue use of resources (Fodor, 1998a). We provide here a quantitative assessment of the substantial increase in learning speed that this model permits compared with traditional grammar guessing models.

## A Simple Model of Grammar Selection

A learning event begins with the learner receiving a novel input sentence  $s$ , a sentence of the target language not licensed by the learner's current grammar  $G_c$ .  $G_c$  may have some parameters set to correct values, but it also has one or more set to incorrect values or not set at all. The learner first attempts to parse  $s$  with  $G_c$ . If the parse were successful, no change would be made to  $G_c$ ; this is error-driven learning (Gold, 1967). Since by hypothesis  $G_c$  does not license  $s$ , this parse attempt fails, and the learning device seeks an alternative grammar. To preserve psychological plausibility we will make the strong assumption here that only one more parsing attempt may be made on this same input. Hence, the task of finding a new grammar that does license  $s$  must be achieved by means of just one parse. If it is not,  $s$  must be discarded and learning must await further input. Each such discard increases the total number of inputs needed before the target grammar is attained, and hence decreases the speed of learning and increases the effort expended. It is important, therefore, for the learner to make good use of the one parse test of a new grammar that it is able to conduct.

However, standard grammar guessing procedures extract only minimal information from their interrogation of the input. Because of the single-parse restriction, just one

grammar must be selected to undergo the parse test, and this selection must of necessity be made *prior* to parsing. The success or failure of this parse attempt with the hypothesized grammar,  $G_h$ , provides feedback on the basis of which the learner decides whether or not to adopt  $G_h$ .<sup>1</sup> If the parse is successful the learner will adopt  $G_h$ ; if the parse fails the learner is assumed to retain  $G_c$ . This policy of shifting to a new grammar only if it licenses the current input is the **Greediness** constraint of GW94 and others. If  $G_h$  happens to be the target grammar, it will be retained permanently and learning is complete. A grammar is correct for the target language if it has the target value for each parameter relevant to the language. If some parameters are irrelevant to the target language there will be an equivalence class of correct parameters. For convenience in what follows we will refer to these grammars collectively as 'the target grammar',  $G_t$ .

### Assessing Grammar Selection Efficiency

A simple random choice learning model such as this will demonstrably converge on the target grammar (Berwick & Niyogi, 1996). However, learning is slow largely because the learning component bases its actions on the mere fact that some randomly chosen  $G_h$  does, or does not, license  $s$ . We will show that learning could be substantially faster if instead the parser could reliably identify for the learner a grammar that licenses  $s$ . For the present let us suppose that this information is provided to the learner by an oracle. Later we will show how this oracle can be implemented.

Given  $G_c \neq G_t$ , what is the probability that the learner will shift to  $G_t$  as a result of an encounter with an arbitrary input sentence  $s$ ? At the point where  $G_c$  is rejected, the random choice learner (without oracle) picks an alternative from among the set of all possible grammars, of which there are  $2^n$  for  $n$  binary parameters. (For simplicity we treat  $G_c$  as a candidate grammar even though it has just failed.) Of these,  $2^i$  are correct (are in the equivalence class  $G_i$ ) where  $i$  = the number of parameters irrelevant to the target language  $e$ . Thus the probability that the learner's selected  $G_h$  is  $G_t = 2^i/2^n$ . Observe that this is not sensitive if  $s$  uniquely determines every parameter value in the target grammar, the learner has no more chance of guessing correctly than if  $s$  is fully ambiguous. This is because, as noted, this learner must make its selection *before* testing out the selected grammar on  $s$ , and so it cannot restrict its guesses to grammars which license  $s$ .

Imagine now that this learning device is equipped with an oracle which offers the learner a grammar that licenses  $s$  (any one of the grammars in the domain that do so). Then the learner could take this grammar to be  $G_h$ , and avoid wasting attention on any grammar that does not license  $s$ . Let us say that a learning device which considers only grammars that

license the current input meets the **Licensing** condition.<sup>2</sup> For a learner that satisfies Licensing, the chance of hypothesizing  $G_t$  would be  $1/A$ , where  $A$  is the degree of ambiguity of  $s$ , measured as the number of grammars in the domain that license  $s$  divided by  $2^i$  (the irrelevance factor). Clearly this is responsive to how informative the language sample is. For extremely ambiguous input ( $A$  approaching  $2^{n-i}$ ), the success rate is hardly better than without the oracle. But if  $s$  is unambiguous with respect to even one relevant parameter, the probability of a successful guess is increased.

This shows up in the speed of learning, estimated in terms of the number of inputs required, on average, to arrive at  $G_t$ . This is the reciprocal of the probability of a successful guess.<sup>3</sup> For the oracle learner this is  $A$ ; for the random guess learner, it is  $2^{n-i}$ . Table 1 shows the differences in average number of inputs consumed for various values of  $A$  and numbers of relevant parameters ( $= 2^{n-i}-A$ ). On average, performance is improved by a factor of  $(2^{n-i}-A)/2^{n-i}$ . The oracle learner, unlike the random choice learner, benefits to the extent that the input constrains the set of candidate grammars.

Table 1  
Reduction due to oracle in average inputs to convergence

Avg. A	Number of parameters relevant to $G_t$ ( $=n-i$ )				
	10	15	20	25	30
1	1023	32,767	1,048,575	33,554,431	1,073,741,823
10	1014	32,758	1,048,566	33,554,332	1,073,741,822
100	924	32,668	1,048,476	33,554,422	1,073,741,724
1000	24	31,768	1,047,576	33,553,432	1,073,740,824
1 million	-	-	48,576	32,554,432	1,072,741,824
1 billion	-	-	-	-	73,741,824

The values of  $A$  range from 1 to  $2^a$ , where  $a$  is the number of parameters relevant to  $G_t$  whose values  $s$  does not determine. For a simple example: Assume that 30 parameters are relevant to  $G_t$  and  $a = 25$ . Such a sentence might be licensed by exactly 2 grammars, with opposite values for each of those 25 parameters. Or it might be licensed by grammars with all possible combinations of values for those parameters, of which there are  $2^{25} = 33,554,432$ . The former situation we will term **sparse ambiguity**, and the latter **dense ambiguity**; clearly, all situations in between are possible also.

It seems likely that the parametric ambiguity of natural

<sup>1</sup>Licensing is related to Greediness but the difference between them is important. Licensing applies in the selection of  $G_h$ , while Greediness governs only the grammar adoption stage at the end of each learning event. A learner that respects Licensing can also respect Greediness. The simple learning model discussed above shows that it is possible to obey Greediness but not Licensing.

<sup>2</sup>Licensing is related to Greediness but the difference between them is important. Licensing applies in the selection of  $G_h$ , while Greediness governs only the grammar adoption stage at the end of each learning event. A learner that respects Licensing can also respect Greediness. The simple learning model discussed above shows that it is possible to obey Greediness but not Licensing.

<sup>3</sup>Homogeneity is assumed in these calculations; no grammar is antecedently more likely than any other to license  $s$  or to be  $G_t$ .



languages is quite sparse ( $A$  much less than  $2^3$ ). In the miniature natural language domain defined by 3 parameters presented in GW94, ambiguity is less than fully dense in every one of the 11 sentence types in which two or more parameters are ambiguously expressed. It remains to be seen how this scales up in a domain of more realistic size. But the principle is clear. With maximally sparse ambiguity, a sentence could be ambiguous with respect to every parameter and nevertheless offer the oracle learner a 50% chance of guessing the target grammar. In general, for a constant degree of parametric ambiguity in terms of  $a$ , sparse ambiguity is more informative for a learning system capable of making use of it, i.e., a learning system that has knowledge of which grammars do and do not license the current input.

In a simple random choice system this information is unobtainable. It could be established only by testing every possible grammar on  $s$ , which clearly violates the limit of one parse per sentence (plus the original parse with  $G_c$ ). Of course, this one-parse limit is just one instantiation of a practical ban on excessive processing, and the limit might be raised to two or three parses per sentence. But this will make little difference. In order to significantly reduce the amount of input needed for convergence, it would be necessary to permit testing of each sentence with as many grammars as required to find one that licenses it.<sup>4</sup>

However, there are other ways of improving the quality of grammar selection which do not presuppose an ability to sort grammars into those that do and do not license  $s$ . We review these in the next section. Their effects are less easy to quantify, but it is highly doubtful that either singly or jointly they could substitute for the usefulness of a Licensing oracle.

### Criteria for Grammar Selection

Given a particular input sentence  $s$  from the target language, which grammar in the domain is it optimal for a learner to hypothesize? The grounds for selecting a grammar may be of several kinds, differing with respect to how much information they draw from the current learning situation. The preference for one grammar over another may be (a) independent of the current situation, or it may (b) reference the current grammar, and/or it may (c) reference properties of the current input. Some criteria of this latter kind may (d) require parsing of  $s$  with more than one new grammar, and thus exceed the limit on feasible processing for a learner without an oracle, which must select candidate grammars before knowing how they relate to the input.

#### (a) Orderings on the Class of Grammars

Grammar orderings may be imposed by linguistic principles of markedness. One value of a parameter may be less marked (more favored) than its other value; e.g., local binding of anaphors may be less marked than long-distance anaphors (Manzini & Wexler, 1987). Or parameters may be ordered with respect to each other: the marked value of one parameter may be less marked than the marked value of another

<sup>4</sup>Parse-testing two grammars on  $s$  would double the chance of guessing  $G_c$ . This would be helpful as if  $s$  were unambiguous with respect to one parameter. However, testing  $m$  grammars on each sentence would increase the chance of success only linearly in  $m$ , not exponentially.

parameter (Clark, 1989). Linguists have mostly been cautious about embracing markedness theory, but many markedness-type rankings are nevertheless implicitly assumed in linguistic descriptions (Wacholder, 1995). Also under type (a): grammars could be prioritized by linguistic maturation if, as has been proposed, some aspects of UG develop later than others (Wexler, 1999).

Criteria of type (a) may be helpful in resolving parametric ambiguities. To the extent that linguistic markedness has an impact on the frequency of grammar adoption by language communities (though this is a fraught topic), type (a) criteria can reflect the antecedent likelihood that any given grammar is the target. They may also reduce effort by holding learners to simpler or linguistically more natural grammars as long as the evidence permits.

#### (b) Rankings Relating to $G_c$

One  $G_c$ -related criterion is the Single Value Constraint (SVC) of GW94, which requires the learner to select grammars that differ from  $G_c$  in the value of just one parameter. Another is the assumption of 'indelible' or 'deterministic' learning, which requires that  $G_h$  include  $G_c$ . For parameter theory this is taken to mean that once a parameter has been set, it may never be switched to its other value (Clahsen, 1990/91).

Type (b) criteria reflect what has already been gained by experience of the target language, insofar as this is compressed into the grammar  $G_c$  that the learner has been led to so far. (A learner is standardly assumed to have no memory of past inputs or past grammar hypotheses, other than their legacy in determining the current grammar.) Because of Greediness, a grammar that has been adopted by the learner may be assumed to be more likely to have some parameters correctly set than an arbitrary grammar in the domain; and a grammar similar to such a grammar may be presumed to share its virtues. The worth of these considerations has been disputed, but we need not enter the debate here; see Berwick & Niyogi (1996) and Sakas & Fodor (in press) for discussion. Clearly it is desirable for a learning device to have some way to hold onto past gains. To adopt a completely fresh hypothesis at each step, as permitted in an unconstrained guessing model, does nothing to improve the probability of success as learning proceeds.

#### (c) Rankings Based on Properties of $s$ Identifiable by Parsing $s$ with at Most One New Grammar

Type (c) criteria are sensitive to the current input but compatible with the ban on excessive processing even for a learner that first selects a grammar and then tests it. Input-sensitive criteria can deliver hard information. They constitute the learner's contact with the facts of the language and so should be a particularly helpful guide to the correct grammar.

Greediness and error-driven learning fall under type (c) as well as (b), since they refer to  $s$  as well as  $G_c$ . Greediness ranks all grammars that do not license  $s$  lower than  $G_c$ . The requirement of error-driven learning ranks  $G_c$  above all other grammars if  $G_c$  licenses  $s$ . These two input-sensitive criteria can be incorporated into a simple grammar guessing procedure without a Licensing oracle, because each can be checked with limited resources: a parsing attempt with  $G_c$

for error-driven learning, and then with one new grammar for Greediness.

By contrast, some input-sensitive selection criteria do not qualify as type (c) conditions because they require (or may do so) the checking of two or more new candidate grammars. This threatens to violate the ban on excessive processing for a learner without oracle. Licensing (as opposed to Greediness) is one casualty already noted: it imposes the tough requirement that a grammar must be known to be capable of parsing  $s$  in order to be selected for parsing  $s$ . Also not possible under type (c) is comparison of the derivations assigned to  $s$  by different candidate grammars, as would be necessary for application of a structural simplicity metric.

Even the grammar-similarity constraints of type (b) are affected by the limitation on processing. The SVC has been demonstrated by GW94 to be too stringent in that, in conjunction with Greediness, it can trap the learner at a local maximum where there is no grammar that both licenses an input sentence and differs from  $G_c$  in only one parameter value. If a range of alternative grammars could be evaluated, a more general **Closeness** criterion could be applied instead. The learner would adopt the grammar most similar to  $G_c$  among those that license  $s$  (with dead heats resolved by random choice or other criteria). The adopted grammar would differ from  $G_c$  by only one parameter value in many cases, but could differ by two or more if necessary. This would maintain the fruits of past learning while eliminating all local maxima. (Note that Closeness can be seen as a generalization of error-driven learning:  $G_c$  is to be changed only to the extent that is necessary in order to license the input.) But for this we must move up to type (d) criteria, which are not feasible for a standard grammar guessing learner.

#### **(d) Rankings Based on Properties of $s$ Identifiable Only by Evaluation of Multiple New Grammars**

Closeness is an ideal similarity metric but (unlike the less flexible SVC) it is a comparative criterion which demands knowledge of all the grammars that license  $s$ , so that the one most similar to  $G_c$  can be selected. This puts it beyond the scope of any resource-limited pre-parse grammar selection process. Also falling under type (d) would be a simplicity measure which favors grammars that assign the smallest syntactic tree, or the shortest transformational derivation, compatible with the word string. This selection criterion seems very plausible both linguistically and psychologically, but is not easy to impose. How could a resource-limited learner set about discovering which of a million or a billion grammars assigns the simplest structure to  $s$ ?

In general: Adding suitable grammar selection principles to a random choice learner can improve performance, compensating in part for inefficiency due to inability to discriminate between grammars that do and do not license  $s$  before committing resources to those that do not. However, the present analysis of grammar selection strategies makes clear that the most potent selection principles are also beyond the reach of such a system, and for much the same reason. We next show that both weaknesses can be remedied by the same means. With one change in how the parse test is conducted, the guessing learner can gain both a Licensing oracle which eliminates useless grammar guesses, and also the powerful type

(d) criteria which improve the quality of guesses in case more than one grammar meets the Licensing condition, i.e., in case of parametric ambiguity.

### **Superparsing: A Constructive Process of $G_h$ Selection**

Inefficiency results from formulating a grammar hypothesis in advance of parsing the input string. This was assumed to be unavoidable, given the patent unfeasibility of first analyzing the string with all grammars as a basis for selecting one from among them. But if an optimal grammar choice cannot be made *before* parsing  $s$ , or *after* parsing  $s$ , perhaps it can be made *in the course* of parsing  $s$ . The solution we will outline is to let the ongoing parse shape the formulation of  $G_h$ . Total parametric decoding cannot be achieved by this means, for reasons we will explain, but most of the desirable learning characteristics we have been seeking do follow. By the end of the parse, the learner will know of one grammar that licenses  $s$ . Hence there will be no wastage of input due to lack of a grammar to parse it. The grammar the parser finds will always be drawn from among the  $A$  grammars that license  $s$ , rather than from the total set of  $2^n$  grammars (or  $2^{n-i}$  relevant grammars), so the learner will be taking full advantage of parametric disambiguation provided by the input. Where disambiguation is not total, Closeness and a structural simplicity metric can be applied to choose a good candidate, as indicated below.

Selecting a grammar that licenses  $s$ , during the course of parsing  $s$ , is feasible. Fodor (1998a) suggested the following procedure. The parsing routines set about parsing  $s$  with the current grammar  $G_c$ . If  $s$  is not licensed by  $G_c$  this parse attempt will break down at some place in the sentence. When it does, the parser should not stop and merely report back its failure, as we assumed earlier. Instead, it should supplement  $G_c$  with all possible parameter values and continue processing  $s$  with this 'supergrammar'  $SG$ .  $SG$  must afford at least one parse for  $s$  (as long as the sentence contains no unknown lexical items, and does not cause a severe 'garden path' beyond the capacity of the parser to recover from; see Fodor, 1998b). Where there is a choice of analysis for  $s$ , priority is given to the parameter values in  $G_c$ ; this incorporates the error-driven learning condition. But new parameter values can be made use of as needed. Any new parameter value that is found to be necessary for parsing  $s$  is adopted by the learner. Thus, the superparser shuttles through the sentence flipping parameter settings as it goes, in response to the demands of the input sentence. Its output consists of (i) a complete parse tree, and (ii) a grammar that satisfies Licensing.

If  $s$  is fully unambiguous with respect to all the parameter values it expresses, the superparser has no choices to make (above the usual within-grammar ambiguity resolution choices of normal sentence processing). If  $s$  is ambiguous with respect to  $a$  parameters,  $SG$  assigns it up to  $2^a$  distinct parse trees. In principle the parser might identify them all. In practice it could not, since this would require massive parallel parsing which would violate the general ban against excessive processing (even though it doesn't strictly violate the one-parse-per-sentence constraint imposed above). More reasonable is to suppose that the parser employed by the

learner for superparsing is the same parser that will be used throughout life for sentence comprehension. A standard assumption is that this is a serial device which, when it hits a point of ambiguity, selects one structural analysis to pursue for the sentence.<sup>55</sup> (Parallel parsing models have been proposed, but to conserve resources their parallelism is strictly limited, and their consequences for superparsing do not differ significantly from those of serial parsing.) Thus the superparser may be faced with choices to make between alternative ways of resetting parameters to assign an analysis to *s*. It can output only one of the *A* grammars that would satisfy Licensing. The choice between them might be random, or other selection criteria must be invoked.

Markedness and conservatism criteria (types (a) and (b)) could be employed, as well as input-sensitive type (c) criteria. The more powerful type (d) criteria such as Closeness and a minimal structure constraint are also available in this system. In fact, both of the latter are more or less automatic consequences of superparsing given that the human parser is a least-effort device (Inoue & Fodor, 1995). For instance, the Minimal Attachment parsing strategy entails that superparsing will prefer simple, compact trees over more complex ones; the learning device inherits this and so favors grammars that assign simpler structures. A conservative policy of staying close to the previous grammar will result if, as is natural, the parser makes the effort of changing parameter settings in  $G_c$  only when it is forced to do so to avoid parse failure. Again, parser preference translates into learner preference. In much the same way, frequency sensitivity in parsing could lead to frequency-sensitive learning (e.g. Charniak, 1993).

The exact mix of these various criteria remains to be established (e.g., Minimal Attachment versus minimal resetting of parameters). The supergrammar model allows various policies for resolving conflicts; which of these is adopted by human learners is an empirical question. To the extent that these criteria help the learner select an optimal grammar from among those that license *s*, the fact that they can be applied by a superparsing learner means that its efficiency gain compared with pre-parse selection criteria is even greater than was calculated above (Table 1). However, exact benefits are not easily quantified. The effects of Closeness and other such rankings are complex, and are best assessed by simulation studies. This awaits future research.

### Limitations of Superparsing

Does superparsing as a means of parameter setting carry significant costs to offset these advantages, so that no net gain in efficiency results? This appears not to be so.

As noted, the parsing routines need not be unusually powerful. The mechanism can be the normal human sentence parsing device, which clearly must be present in children for comprehension of sentences already licensed by the current grammar. Thus, all that is special about the superparser is that

it applies the supergrammar, augmented with all possible parameter values. This could exact a heavy cost in on-line processing due to the massive ambiguity of sentences in relation to the supergrammar, far greater in many cases than ambiguity levels relative to a settled adult grammar. However, the added cost of ambiguity is negligible as long as no attempt is made to compute all analyses of a sentence. For a serial parser, alternative parses are evaluated only momentarily as each new word is encountered and attached into the parse tree. They are not pursued through the sentence, and are not multiplicative. As soon as one of the alternatives has been chosen, the others can be forgotten. And arguably, even the selection process is cost-free in a least-effort system, since it consists of adopting the first (simplest) attachment option that is computed (Frazier & Fodor, 1978; Lewis, 1999). The only difficult analyses will be (i) those the human sentence parser has trouble with even when the grammar is settled, e.g., center-embedded constructions; and (ii) analyses which are not complex in themselves but are systematically masked by more attractive analyses allowed by the supergrammar. Grammar guessing models without superparsing would suffer from (i), but not from (ii). The incidence of such cases is not known. They could lead to false negative reports from the parser to the learner, indicating wrongly that *s* is not licensed by the grammar being tested. For examples and discussion see Fodor (1998b).

A requirement for smooth functioning of the superparsing routine is that the parameter values defined by UG are such that they can be added into a natural language grammar without altering its basic character. The competing values of one parameter must be able to co-exist in the same grammar without internal contradiction. And the parameter values temporarily added into  $G_c$  to create the supergrammar should be no harder to access and use on-line than other elements of natural language grammars. This may preclude any kind of precompiling process by which the combination of  $G_c$  and the added parameter values is reformulated for convenience in parsing, since the computational costs of repeated compiling would be added into the workload of the superparser. For some kinds of parameters (e.g., Subjacency applies at Surface Structure or at Logical Form; Huang, 1981/82) these conditions are hard to meet. But a variety of current linguistic theories conceive of parameter values as fragments of tree structure (see Fodor, 1998c), and these 'treelets' do meet the needs of superparsing. They can be directly added into a normal grammar to create another perfectly normal grammar, only slightly more complex than the original, and yet incorporating all the structural options that UG permits.

The one limitation of superparsing that is unavoidable is that it delivers only one structural analysis for each sentence. Because of the ban on excessive processing, it is impossible for the parser to present the learning component with all analyses for *s*, to compare and evaluate in order to make the best possible guess. The process of selecting one of the licensing grammars is piecemeal and order-dependent as each ambiguity must be resolved as it arises on-line. Interestingly, this appears to do relatively little damage, because there seems to be an excellent fit between the choices made on-line by the human parser and the choices that a well-designed learning device would be expected to favor: the minimization of derivational complexity, and the minimization of grammar

---

<sup>55</sup>In a serial parse, the selected resolution of an ambiguity may prove to be incorrect, by failing on subsequent words of the sentence. In a garden path situation such as this the superparser would engage in reanalysis procedures just as the human parser normally does in the case of a garden path.

revision. Whether this is merely a coincidence is not clear, but at any rate it is fortunate for superparsing as a method for parametric decoding.

## Summary and Conclusions

A start has been made here on quantifying the efficiency advantage of a learning device which has the ability to read off a set of parameter values for licensing a sentence, in the course of parsing that sentence in the normal way for comprehension. The superparsing approach was developed originally for a different purpose. It was designed to provide a feasible ambiguity detection system, so that all parametrically ambiguous input could be discarded. This permitted development of a non-guessing learning routine, capable of error-free parameter setting based exclusively on unambiguous items in the input sample (Fodor, 1998a). Whether this is the best research goal, either for modelling human learning or for engineering applications, remains to be seen. The error-free learner has the advantage that it never has to re-set a parameter. Also, once it has set a parameter it can ignore the alternative value of that parameter thereafter, so the size of the domain to be searched for  $G_i$  shrinks as learning proceeds. It is an empirical issue whether these benefits balance the need to discard all ambiguous sentences in the input sample. It is therefore of interest, as we have shown here, that superparsing can also make a useful contribution to a grammar guessing routine.

To the extent that the input sample does carry parametric information, superparsing allows the learner to exploit it. Despite its modest consumption of resources, and despite its practical inability to list all parametric analyses of a sentence, the superparser nevertheless extracts from a sentence *all* the definitive parametric information it contains. If a sentence is compatible with only one grammar in the whole domain, the superparser will identify that grammar and the learning component will adopt it. If a sentence is less informative, e.g., is compatible with a thousand grammars, the superparser will identify one of the thousand. (Which one of the thousand it identifies depends on which ambiguity resolution criteria it applies on-line.) All parameter values expressed unambiguously by a sentence will be set correctly by the time it has been parsed. The values of the other parameters can only be guessed. They will be left unchanged from previous learning where possible; otherwise they will be changed to some combination of values which licenses  $s$ .

The superparsing learner is quite undemanding about the nature of its input. For example, it does not require a language to contain unambiguous triggers for all of its parameter values. A simple random grammar guessing learner also needs no unambiguous triggers, but that is because, as noted above, it gains hardly more from unambiguous than from ambiguous input. By contrast, a non-guessing error-free learner is very choosy; it needs an unambiguous trigger for each parameter, and moreover it needs some of these to be fully unambiguous (i.e., unambiguous with respect to all parameters they express). The superparsing learner has the dual virtues that it can use the information in fully unambiguous triggers when they are present, but it can also make progress when input sentences are ambiguous with respect to many (or even all) of the parameters they express. Thus it is robust as well as efficient.

## Acknowledgements

This research was supported in part by PSC-CUNY Grants 61595-00-30 to Fodor and 61403-00-30 to Teller.

## References

- Berwick, R. C. and Niyogi, P. (1996) Learning from triggers. *Linguistic Inquiry* 27.4, 605-622.
- Charniak, E. (1993) *Statistical Language Learning*, MIT Press, Cambridge, MA.
- Chomsky, N. (1981) *Lectures on Government and Binding*, Foris Publications, Dordrecht.
- Chomsky, N. (1995) *The Minimalist Program*, MIT Press, Cambridge, MA.
- Clahsen, H. (1990/91) Constraints on parameter setting: A grammatical analysis of some acquisition stages. *Language Acquisition* 1.4, 361-391.
- Clark, R. (1989) On the relationship between the input data and parameter setting. *NELS* 19, 48-62. GLSA, UMass.
- Clark, R. (1992) The selection of syntactic knowledge. *Language Acquisition* 2.2, 83-149.
- Clark, R. (1994) Finitude, boundedness and complexity. In B. Lust, G. Hermon and J. Kornfilt (eds.) *Syntactic Theory and First Language Acquisition: Cross-linguistic Perspectives*. Lawrence Erlbaum, Hillsdale, NJ.
- Fodor, J.D. (1998a) Unambiguous triggers. *Linguistic Inquiry* 29.1, 1-36.
- Fodor, J.D. (1998b) Parsing to Learn. *Journal of Psycholinguistic Research* 27.3, 339-374.
- Fodor, J.D. (1998c) What is a parameter? Presidential Address to the Linguistic Society of America.
- Frazier, L. and Fodor, J.D. (1978) The sausage machine: A new two-stage parsing model. *Cognition* 6, 291-325.
- Gibson, E. and Wexler, K. (1994) Triggers. *Linguistic Inquiry* 25.3, 407-454.
- Gold, E.M. (1967) Language identification in the limit. *Information and Control* 10, 447-474.
- Huang, C.-T. J. (1981/82) Move *Wh* in a language without *Wh* movement. *The Linguistic Review* 1, 369-416.
- Inoue, A. and Fodor, J.D. (1995) Information-paced parsing of Japanese. In R. Mazuka and N. Nagai (eds.) *Japanese Sentence Processing*, Lawrence Erlbaum, Hillsdale, NJ.
- Lewis, R. (1999) Attachment without competition: A computational model of race-based parsing. 12th Annual CUNY Conference on Human Sentence Processing.
- Manzini, R. and Wexler, K. (1987) Parameters, binding theory and learnability. *Linguistic Inquiry* 18.3, 413-444.
- Nyberg, E. (1992) *A Non-deterministic Success-driven Model of Parameter Setting in Language Acquisition*. Unpublished Ph.D. Dissertation, Carnegie Mellon University.
- Sakas, W. and Fodor, J.D. (in press) The structural triggers learner. To appear in S. Bertolo (ed.) *Parametric Linguistics and Learnability: A Self-contained Tutorial for Linguists*, Cambridge University Press, Cambridge, UK.
- Wacholder, N. (1995) *Acquiring Syntactic Generalizations from Positive Evidence: An HPSG Model*. Unpublished Ph.D. Dissertation, City University of New York.
- Wexler, K. (1999) Maturation and growth of grammar. In W.C. Ritchie & T.K. Bhatia (eds.) *Handbook of Language Acquisition*, Academic Press, San Diego.

# Lexical contact during speech perception: A connectionist model

**Eric Forbell** (eforbell@bowdoin.edu)

Department of Computer Science; 8650 College Station  
Brunswick, ME 04011

**Eric Chown** (echown@bowdoin.edu)

Department of Computer Science; 8650 College Station  
Brunswick, ME 04011

## Abstract

A connectionist architecture comprised of cell assemblies was developed and applied to the problem of speech perception at the phonemic and lexical levels. The problem addressed involved a disagreement amongst theorists over the possible sources of lexical priming effects. Speech was encoded in the model as the temporal activity of phoneme units that are connected to higher-level word assemblies. The lexical layer was topographically organized based upon similarity of phonemic structure. Lateral inhibition at the lexical level was shown to be both necessary and sufficient to support results from phonological priming experiments involving human participants.

## Introduction

Speech processing represents a rich source of constraints for the development of neural models of cognition. These constraints are particularly challenging since they often arise out of the temporal nature of the task, a weakness of most connectionist models. This paper presents a connectionist model that addresses a long-standing matter of controversy in the psycholinguist literature involving a task that is highly temporal in nature.

The model task involves lexical contact, which is defined as the phase of speech processing whereby the representations activated by the speech input make initial contact with the lexicon. The data comes from a series of studies to test a variety of priming conditions (Hamburger & Slowiaczek, 1996; Slowiaczek & Hamburger, 1992). The basic design of these studies involved prime words that are related to target words by sharing initial phonemes (i.e. "black" and "bleed" share two initial phonemes). Primes are presented 500 ms. before the targets and the subject's only task, called shadowing, is to repeat the target word aloud as quickly and accurately as possible. The major result is that when three initial phonemes are shared, response time actually slows by about 40 ms. whereas there is very little effect when one or two phonemes are shared (Hamburger & Slowiaczek, 1996).

The model for these data is based upon a variant of Hebb's cell assembly called TRACE (Kaplan, Sonntag, & Chown, 1991). The cell assembly is particularly suited for this type of modeling since it is well grounded biologically and was originally proposed to address issues of temporal processing in a neural framework. TRACE modeled a single cell assembly and has since been extended to model

multiple cell assemblies in a more general system called multiTRACE (Chown, 1994; Sonntag, 1991).

## Lexical and phonological priming

The data from Slowiaczek and Hamburger suggest that phonologically similar words compete at the lexical level of speech recognition (1996). Specifically, this competition was observed in a priming paradigm whereby the primes were phonologically related to target words by the number of initial phonemes. The critical number of overlapping initial phonemes was three, for primes that were presented 500 ms before the targets. Additionally, this data also elucidated the cause of a facilitatory effect (a decrease in RT) in the low-similarity case (1- to 2- phoneme overlap) that was presenting some difficulty in interpretation.

It was suggested from other studies that low-similarity facilitation only occurs if the phonological relatedness proportion (PRP) is high (50 %) (Goldinger, Luce, Pisoni, & Marcario, 1992). That is, the number of trials containing phonological overlap was manipulated across subject groups, and it was only in groups containing a majority of trials with phonological overlapping (high PRP) that displayed facilitation effects (Hamburger & Slowiaczek, 1996). Goldinger et al. explained this prelexical facilitation effect by suggesting that subjects were strategically assuming that the initial phoneme of the target would be the same as that of the prime, because the majority of previous trials had been this way (1992). When the experiment was controlled for subject expectancy, however, this facilitatory effect was virtually eliminated and only the 3-phoneme overlap interference was observed (Hamburger & Slowiaczek, 1996). Nevertheless, there remains a debate in the literature over whether either effect truly comes from competition among lexical candidates or whether it is simply an artifact of the experimental design.

A different kind of priming has been studied with a smaller interstimulus interval (the time between the end of the prime word and the beginning of the target) of 50 ms. Goldinger et al. found that for uncommon targets (low-frequency words) preceded by phonologically related primes, response time was increased (1992). This phonological inhibition result is included to provide a bigger picture of the priming literature, but was not a focus of this work.

## The multiTRACE model

Hebb developed the cell assembly construct to address questions concerning the temporal nature of neural processing. Essentially a cell assembly is a large collection of neurons which act in concert and which have temporal extent due to their recurrent connections and their corresponding ability to “reverberate.” Hebb’s theory lost favor initially in part because he omitted inhibition, a construct for which there was no evidence at the time. More recently, however, cell assemblies have undergone something of a revival as advances in neuroscience have been incorporated in the theory (Kaplan, et al., 1991) and experimental evidence for their existence has been found (Amit, 1995).

In the Kaplan model of cell assemblies, called TRACE (Tracing Recurrent Activity in Cognitive Elements), the emphasis was on simulating the internal dynamics of a population of neurons that would comprise a cell assembly. In the TRACE model various neural control mechanisms were postulated to play different functional roles in the cognitive system. For example, inhibition is useful as a selection mechanism when multiple cell assemblies are competing to become active. A major addition to cell assembly theory by the Kaplan group was to add fatigue to counterbalance the reverberation inherent in a highly recurrent system.

TRACE, which serves as the basis for multiTRACE, uses a set of difference equations that are updated at each time step to model the collective behavior of a large group of neurons. The equations model various biological functions such as activity, neural fatigue, short-term connection strength, long-term connection strength, sensitivity to firing, and network or external input (Table 1).

Kaplan et al. argued that units built with these basic properties have a number of advantages over the simple units used in many traditional connectionist models (1991). Different levels of activity in a cell assembly, for example, can serve different cognitive purposes, such as coding for conscious versus unconscious processing. The major questions left open by the original work on TRACE was how the notion of a single cell assembly could be extended to the cognitive system as a whole.

Table 1: The basic multiTRACE equations

Update Equations	Delta Equations
$A(t+1) = A(t) + \Delta A$	$\Delta A = (A + \bar{A}I)\bar{A}V - A^{\theta_L} + A\bar{A}^{\theta_C}\bar{V}$
$F(t+1) = F(t) + \Delta F$	$\Delta F = \phi_g A\bar{F} - \phi_d F$
$S(t+1) = S(t) + \Delta S$	$\Delta S = \sigma_g AS - \sigma_d S$
$L(t+1) = L(t) + \Delta L$	$\Delta L = 0.0$
	$V = \frac{1}{v}(S + L)F$
	$I = I^{exc} - I^{inh}$ (expanded in text)

$\theta_l$  : unit loss  
 $\theta_c$  : inh. competition  
 $v$  : normal factor  
 $\phi_g$  : fatigue growth  
 $\phi_d$  : fatigue decline  
 $\sigma_g$  : STCS growth  
 $\sigma_d$  : STCS decline

A: activity  
F: neural fatigue  
S: short-term connection strength  
L: long-term connection strength  
I: network input

\*  $\bar{X}$  denotes quantity  $(1 - X)$

The multiTRACE model extends the cell assembly idea by building models with collections of cell assemblies. Sonntag originally created multiTRACE to study sequence learning in the context of cell assemblies (1992). Chown later extended the model to deal with other forms of learning, for example modeling the effects of the arousal system on learning (1994). The development of multiTRACE has been increasingly less abstract, starting from the very general problem of modeling sequences, to the current work which addresses a very specific body of data.

### Applicability to lexical priming

The multiTRACE framework provides a natural way to model the lexical priming data presented in the previous section. Each phoneme and lexical unit is represented by a cell assembly as part of a hierarchical structure (Fig. 1). Phonemes which are part of a word are strongly linked to the lexical units at the higher level (e.g. the phoneme “b” will be strongly linked to the cell assembly representing “black” but not the one representing “flack”). The activation of the lexical units at the higher level corresponds to perception and therefore the ability of the subject to repeat the word. The theory is that competition between these units accounts for the differences in timing.

This sort of perceptual competition forms the basis of a number of connectionist models and stems from evidence that similar concepts tend to interfere with each other more than dissimilar ones as part of what Kinsbourne called “the functional cerebral distance principle” (Kinsbourne, 1982). This interference comes in the form of lateral inhibition between cells near each other in the cortex. The idea is simple; words that are similar (e.g. “black” and “blast”) will be stored in nearby locations in cortex, meaning that they will greatly inhibit each other. As one becomes highly active during perception it will naturally inhibit the other, making perception a kind of winner-take-all proposition. In terms of the brain, the cell assemblies underlying these representations will be close to each other. A given cell assembly will have a kind of inhibitory surround which will typically prevent its close neighbors from being simultaneously active (Fig. 1).

In the context of a cell assembly model the interference seen when the target word shares three phonemes with the prime must come from competition, and the competition must come from the prime word itself. Since the prime word and the target word share three initial phonemes in common they will be represented very near each other in the brain, and therefore they will have a great deal of lateral

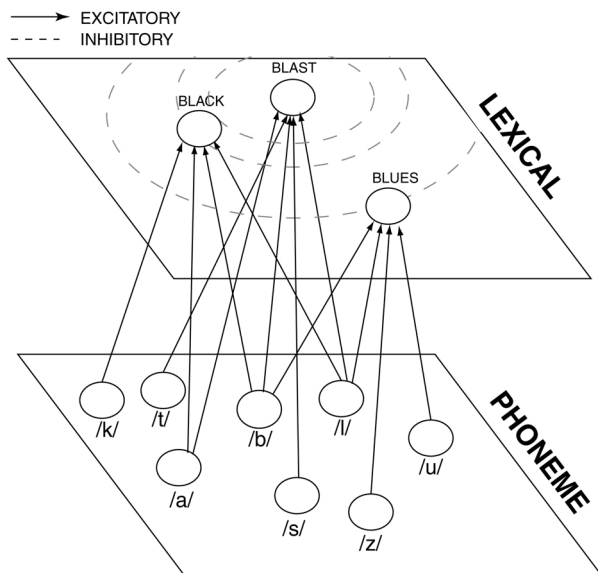


Figure 1: The two-tiered connectionist architecture. (Inhibitory output shown only for BLAST unit)

inhibitory connections. These data are very useful for extending cell assembly theory because they provide a number of useful constraints on the temporal dynamics of cell assembly activation, as well as information on the general layout of cell assemblies in the cortex.

### The implementation

In the original implementations of multiTRACE it was not necessary to explicitly model lateral inhibition as the model was not applied to highly similar concepts as is the case here. In updating the model we devised a scheme, based upon Kinsbourne’s functional distance principle, whereby the lateral inhibitory connections between cell assemblies was determined by the shared number of phonemes. This rule also reflects the spatial layout of inhibitory processes in association cortex, as well as areas like the early visual system, where lateral inhibition is widely known to exist and has important functional implications, such as enhancing contrast sensitivity.

The competition generated by lateral inhibition results from an active cell assembly. Since the priming effects vary according to the test conditions it is important to understand the factors that affect the time course of activity in a cell assembly. In Hebb’s original formulation the only was a cell assembly’s capacity for reverberation (in the multiTRACE model this comes in the form of internal long term connection strength or LTCS). Later through simulation it was determined that inhibition was necessary as a control mechanism (Rochester, et al., 1956). More recently, Kaplan et al. (1991) did a series of simulations showing that with the addition of two more biologically and theoretically motivated mechanisms – fatigue and short term connection strength (STCS) – that it was possible to model different time courses of activity corresponding to different parts of the cognitive hierarchy. For example, cell assemblies near

the perceptual interface would be expected to have a high refresh rate in order to be ready for the next input. On the other hand cell assemblies that participate in long term planning would be expected to stay active for longer periods of time. In our model we conceptually represent these differences as different hierarchical layers.

Although it is possible that hierarchy can emerge naturally in a flat network structure, we felt that such a network design would quickly become confusing and therefore limiting. Additionally, our network structure reflects the layered organization of human cerebral cortex. Groups of multiTRACE units were combined into layers in the current system, and static properties of individual units were inherited from their “parent” layers. It is our conjecture that nearby layers will have similar parametric properties (e.g. in the perceptual layer cell assemblies will all tend to have high refresh rates). The lexical priming data provides an excellent test of such conditions and the potential usefulness of the cell assembly construct, since it can be applied to model widely different types of cognitive functioning.

The basic structure of the simulation was based upon a two-tiered network with each tier representing a different level of the cognitive hierarchy. In this case, because each layer is part of the perceptual interface, they have virtually identical parameter settings (Table 2). The primary layer contained units responsive to phonological stimuli theoretically produced by the primary and secondary auditory cortices. The secondary layer comprised of lexical units that respond to the phonological structure of a spoken word, not its meaning. For example, a lexical unit for “blast” received equal vertical connections from the phonemic /b/, /l/, /a/, /s/, and /t/ units. Because the data from Hamburger and Slowiaczek was not concerned with the typicality effect, we built in the assumption that all of the cell assemblies had the same internal connection strength (1996). Differences in typicality could easily be modeled by introducing variability in connection strength within cell assembly units.

Table 2: Layer parameters and timings

Parameter	Phonemic layer	Lexical layer
Fatigue growth	0.15	0.15
Fatigue decline	0.04	0.04
STCS growth	1.0	1.0
STCS decline	0.2	0.2
Activity duration	300 ms	700 ms
Fatigue recovery	500 ms	950 ms

\* STCS: Short-term connection strength

Despite both layers of units being similar in their static properties, their differentiation in cognitive speed emerged as a function of the hierarchical structure. The units in the word layer received input from several phonemic units over the course of time, depending on the length of the spoken word. Since the mapping between layers was not one-to-one, average unit durations were 300 ms and 700 ms for the phonological and lexical layers, respectively (Table 2). Fatigue recovery times, being dependent on activity, were similarly proportioned.

The topological organization of the lexical layer was also an important component of the current model. Interference, as suggested by the Hamburger and Slowiaczek data (1996), results from competition at the lexical level and is a function of word-form similarity. Therefore, the lexical map was built using the amount of initial phoneme overlap between lexical units as a distance approximation in cognitive space. The amount of inhibition imposed on some target unit  $k$  is a function of the distance to a neighboring unit  $j$ , as well as this neighbor’s activity and fatigue level:

$$I_{jk}^{inh} = \frac{A_j(1 - F_j)}{D_{jk}} \quad (1)$$

j: source unit  
k: target unit  
D: distance  
t: Time

The net inhibitory input for a lexical unit then becomes the combination of local inhibition and regional inhibition imposed on the layer. This regional inhibition is a positive feedback mechanism that controls the spread of activity in a given region, in this case a layer, and is based upon the total activity in that layer:

$$I_k^{inh} = \frac{1}{L} \left( \sum_{j=1}^n I_{jk}^{inh} \right) + G \left( \sum_{i=1}^n A_i \right) \quad (2)$$

n: number of units in a layer  
G: global inhibition factor (0.5)  
L: lateral inhibition factor (2.0)

The excitatory input to a cell assembly in multiTRACE is computed in a conventional connectionist manner. However, the sum of a unit’s long-term connection strength and short-term connection strength to another unit serves as the weight value ( $w_{ij}$ ) typically seen in most connectionist models:

$$I_{jk}^{exc} = (LTCS_{jk} + STCS_{jk}) A_j \quad (3)$$

$$I_k^{exc} = \sum_{i=1}^n I_{jk}^{exc} \quad (4)$$

j: source unit  
k: target unit  
n: number of incoming connections for unit  $k$

### Simulation design and procedure

As in the original experiment, four prime conditions were created: *no relation* and three degrees of *phoneme overlap* (1-3). The representative words for each condition are presented in Table 3; the actual words were chosen arbitrarily for demonstrative purposes.

Table 3: Simulated experiment design

Condition	Prime	Target
No relation	“dream”	“black”
1-phoneme overlap	“bind”	“black”
2-phoneme overlap	“blues”	“black”
3-phoneme overlap	“blast”	“black”

In order to simulate an incoming stream of speech, the phoneme units comprising the prime and target words were activated in a serial manner, separated by an interval ranging from 20-40 ms, with a greater spacing reserved for vowel sounds. Using this approximation method, the experiment was easy to simulate. The set of phonemes corresponding to the prime word were activated first, followed by the target phoneme string 500 ms after the prime sequence had concluded.

The simulations were expected to show that there is a fundamental difference in processing between the high similarity (3-phoneme overlap) and low similarity conditions (no relation and 1-phoneme overlap). That is, the response time of the target word unit represented at the lexical level should be increased due to the earlier presentation of the prime word stimulus.

### Results

Our initial experiments show the relevant trends in the data (Table 4); interference resulting from lexical competition was observed in the high-similarity conditions (2- and 3-phoneme overlap) and not in the low-similarity conditions (0- and 1-phoneme overlap). To date we have not replicated the exact time-course for this interference that was found in the behavioral evidence, but we have found that the general trends are simple to generate in the model. Since the model presented here is considerably simpler than that of its human counterpart, and does not take into account the effects of semantic top-down influence, for example, which may also affect timing, we do not wish to spend too much of our effort trying for an exact match at this stage.

Table 4: RT differences (experimental – control) in simulated and actual experiment

Condition	Simulation (ms)	Actual (ms)
No relation	-	-
1-phoneme overlap	0	-4
2-phoneme overlap	40	-8
3-phoneme overlap	190	36

In the 3-phoneme overlap condition, the competition between “black” and “blast” is striking (Fig. 2). The time course of the prime word’s activity is sufficiently slowed in this condition as well as in that of the target’s. That is, be



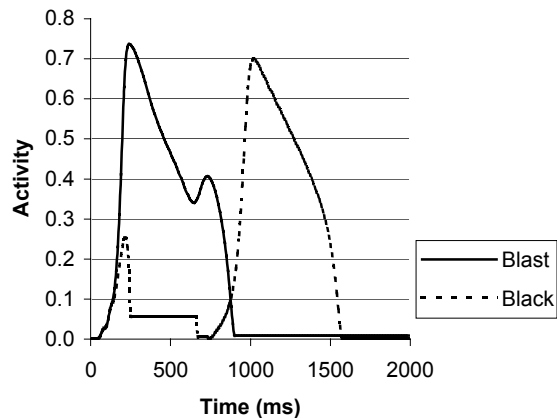


Figure 2: Competition between the “black” and “blast” lexical units

cause “blast” and “black” share three initial phonemes, each unit receives considerable input from the incoming speech stream. However, when the fourth segment of the speech stream (/s/) is presented to the phoneme layer, the net input for the “blast” unit begins to dominate over that of the “black” unit (Fig. 3). This initial competitive advantage is then reinforced by the increasing lateral inhibition “blast” is imposing on “black,” effectively increasing the competitive gap. However, the presentation of the target word 500 ms later in addition to the fatigue of the “blast” unit will allow “black” to win the second competition, albeit more slowly than in the control condition.

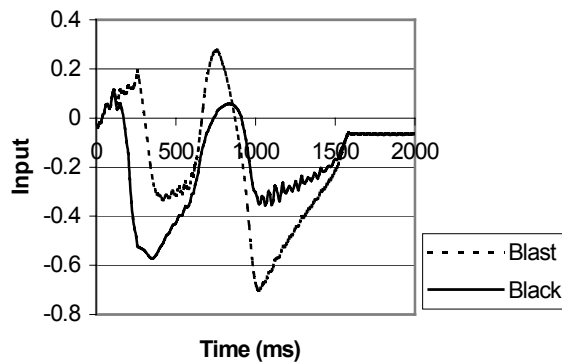


Figure 3: Net input for “blast” and “black” units

## Conclusion

This work serves two purposes. First, we have presented a biologically grounded model that addresses a key controversy in the psycholinguistics literature. Our results support Hamburger and Slowiaczek’s theory that the lexical priming results can be explained in terms of competition between phonologically similar words. Second, their data provides crucial constraints in exploring the temporal dynamics of neural processing in cognition.

For better or worse the dominant connectionist modeling paradigm has long been back-propagation. In recent years, however, interest has grown in recurrent models such as the one presented in this paper. The development of such models will be predicated upon their ability to account for psychological data with a temporal component. This work represents an important step in that direction. The data being modeled required incremental changes in an existing model. As the goal of our continuing research is not to alter previous work in order to support future data, this current work was successful in that components were identified—lateral inhibition and layering—that when inserted into the existing model were able to support the new data. Also, while these additions extend the modeling capabilities of current cell assembly implementations, they do not accomplish this at the cost of simplicity. That is, lateral inhibition and hierarchy fit very naturally into the multiTRACE model and are well supported theoretically.

With regards to the specific modeling task discussed here, future work will involve observing the competition dynamics as the scale of the system is increased. The high-similarity interference phenomenon was observed in a system constructed from roughly ten words, but now that the architecture is in place it will be possible to see how robust the effect will be as the number of words is increased. Another goal of this continued effort will be to reduce the amount of manual network design, because that has resulted in a more discrete representation of the lexical space under study than may be desired to obtain truly generalized conclusions. At the time of publication, however, a sampling of systems of roughly 160 to 200 words generated from a normal distribution have produced statistically significant effects similar to those discussed here (including low-similarity facilitation), and it has been shown that the amount of inhibition at the lexical level is crucial to mimicking the behavioral evidence.

In addition to the artificial nature of the network construction, this discreteness in representation is also a by-product of the necessary simplicity of the current multiTRACE model, in that individual network units represent populations of neurons, thereby limiting our knowledge of how the simulated assemblies can relate to one another in a neurobiological sense. For example, it is not clear to what degree cortical representations for words are distinct or if they overlap. However, this simplicity in the model does not damage its biological credibility. That is, because the internal representations of these simulated cell assemblies remains unspecified, theoretically this allows neurons to be redundantly represented across several units in the model.

## Acknowledgments

The authors would like to thank Louisa Slowiaczek for her invaluable assistance in this project. The first author was partially funded by a Hughes grant during the summer.

## References

- Amit, D.J. (1995). The Hebbian paradigm reintegrated: Local reverberations as internal representations. *Behavioral and Brain Sciences*, 18(4): 617-657.
- Chown, E. (1994). Consolidation and learning: A connectionist model of human credit assignment. Doctoral dissertation. The University of Michigan.
- Hamburger, M. & Slowiaczek, L.M. (1996). Phonological priming reflects lexical competition. *Psychonomic Bulletin & Review*, 3(4): 520-525.
- Hebb, D.O. (1949). *The Organization of Behavior*. John Wiley.
- Goldinger, S. D., Luce, P.A., Pisoni, D.B., & Marcario, J.K. (1992). Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 18, 1211-1238.
- Kaplan, S., Sonntag, M. & Chown, E. (1991). Tracing recurrent activity in cognitive elements (TRACE): A model of temporal dynamics in a cell assembly. *Connection Science*, 3, 179-206.
- Kinsbourne, M. (1982). Hemispheric specialization and the growth of human understanding. *American Psychologist*, 37(4), 411-420.
- Rochester, N., Holland, J.H., Haibt, L.H., & Duda, W.L. (1956). Tests on a cell assembly theory of the action of the brain, using a large digital computer. *IRE Transactions on Information Theory*, IT-2:80—93.
- Slowiaczek, L.M. & Hamburger, M. (1992). Prelexical facilitation and lexical interference in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18(6): 1239-1350.
- Sonntag, M.. (1991). Learning sequences in an associative network: A step towards cognitive structure. Doctoral Dissertation. The University of Michigan.

# The Resemblance of One-year-old Infants to Their Fathers: Refuting Christenfeld & Hill (1995)

Robert M. French, Serge Brédart, Johanne Huart, Christophe Labiouse  
Department of Psychology  
University of Liège, 4000 Liège, Belgium  
(rfrench; Serge.Bredart; jhuart; clabiouse @ulg.ac.be)

## Abstract

In 1995 Christenfeld and Hill published a paper that purported to show at one year of age, infants resemble their fathers more than their mothers. Evolution, they argued, would have produced this result since it would ensure male parental resources, since the paternity of the infant would no longer be in doubt. We believe this result is false. We present the results of two experiments (and mention a third) which are very far from replicating Christenfeld and Hill's data. In addition, we provide an evolutionary explanation as to why evolution would *not* have favored the result reported by Christenfeld and Hill.

## Introduction

Science overwhelmingly favors positive results. To appreciate this, one need look no further than the almost exclusive emphasis on Type I error detection in the social sciences and the quasi-religious status of the inequality " $p < 0.05$ ." Since everyone knows that the null hypothesis cannot be proved, only rejected, it follows that the only results worth pursuing are those that involve the rejection of a null hypothesis. This is why it is so much harder to establish (and publish) negative results. But sometimes null hypotheses are rejected incorrectly and subsequently setting the record straight proves to be very difficult indeed. For every published rejection of a null hypothesis, a far greater number of failures-to-replicate are usually necessary to convincingly establish that the published result was most probably in error.

Although examples of this problem abound, it is instructive to briefly recall the well-known experiments on the chemical transfer of memory done by McConnell (1962) and others. Planarian worms were trained to respond in a certain "correct" way to a light source. These worms were then killed and their RNA fed to a new set of worms, who, by dint of having ingested the previous worms' RNA, would supposedly respond correctly to the light source more often than worms in a control group. Once this result, buttressed by theoretical arguments about the role of RNA in memory, was published it became very hard to unseat it, in spite of numerous failures-to-replicate (e.g., Bennett & Calvin, 1964; Byrne et al., 1966). Thus, even though numerous failure-to-replicate papers had begun to appear as early as 1964, many courses on memory into the 1970's still included McConnell's results on the chemical transfer of memory (see, for

example, Munn, Fernald, & Fernald, 1969; Hilgard, Atkinson, & Atkinson, 1971).

In this paper we will present two negative results that we hope will help serve to establish the falsehood of a published result — namely, the claim of greater resemblance between one-year-old infants and their fathers than their mothers (Christenfeld and Hill, 1995). This result received very wide international attention when it was published in 1995. The result is now cited often but we believe, both for theoretical and empirical reasons, that it is wrong. In the present paper, we will present our own failures-to-replicate the original results and will give a theoretical justification for our results. We hope that this will lead other researchers to also critically examine the originally published results of Christenfeld and Hill before they become firmly, and in our opinion wrongly, entrenched as fact.

The remainder of this paper is organized as follows. We begin by briefly presenting the claim of Christenfeld and Hill (1995). This will be followed by the results of two independent experiments (i.e., different subjects, different sets of stimuli, etc.) that fail to replicate their results. We will then give a theoretical justification for why evolution would most likely have produced our results and not those of Christenfeld and Hill.

## Christenfeld & Hill (1995): One-year old Infants Resemble Their Fathers

Christenfeld and Hill (1995) reported a result in 1995 that appeared in *Nature* and received considerable attention throughout the world, both in the scientific and the popular press. They claimed to have found greater facial resemblance between one-year-old children and their fathers than between one-year-old children and their mothers. Their result had wide appeal, in particular, because it seemed to agree with a prediction of evolutionary psychology (Gaulin and Schegel, 1980) — namely, that "It could then be to a baby's advantage to look like the father, to encourage paternal investment [on the part of the male parent]" (Christenfeld & Hill, 1995) since a mother can be quite sure that the baby is hers but the father cannot.

According to Christenfeld and Hill, greater father-child resemblance would be to the baby's advantage because it would encourage the father's investment in its survival, since he would be able to clearly identify

the child as his own. This would tend to produce a differential survival rate among children who, at age one (when they were most in need of resources from the father for their survival), looked like their fathers and those who did not.

### Overview of the Two Experiments

In an initial experiment (not reported here) involving 200 subjects done soon after Christenfeld and Hill's paper first appeared, we were unable to reproduce their results. We thought that perhaps there might be some problem with the photographic stimuli we were using. (Christenfeld and Hill declined our request to make their original stimuli available.) We therefore created a second set of stimuli, careful to make sure that there the photos displayed no beards, glasses, hats, or other features that might distract from the identification task. However, once again, we failed to replicate Christenfeld and Hill's results. These results are reported in Experiment 1 (see Brédart & French, 1999).

We then created another, entirely new set of stimuli and designed the experiment to record participants' reaction times during identification. As before, we found virtually no difference in the level of correct identification of children and their real mother compared to children and their real father. Further, in addition to repeating the results of the first experiment, there was no significant difference between correct child-mother and child-father identification times.

We have now attempted to reproduce Christenfeld and Hill's result with three different sets of stimuli with three different groups of participants using two different measures (% of correct identification and reaction time). In no case did we find any significant differences in father-infant and mother-infant identification. In other words, we have what we believe to be good empirical evidence that belies the originally reported findings of Christenfeld and Hill.

### Experiment 1<sup>1</sup>

#### *Subjects*

One hundred and eighty undergraduate students at the University of Liège participated in the experiment. Thirty subjects (15 female and 15 male) were randomly assigned to each condition. Their ages ranged from 18 to 30 years (mean age = 21.84).

#### *Stimuli and materials*

Twenty-eight Caucasian families provided five photographs: three photographs of the same child at one year, three years, and five years, as well as one photograph of the mother and one photograph of the father taken when the child was approximately one year old. For fourteen families, the child was a girl, for the

other fourteen families the child was a boy. The stimuli presented to subjects were scanned versions of these photographs (size = 5x4 cm) of faces. None of the faces had glasses, beards or moustaches.

#### *Procedure*

On each trial, participants were presented with the face of a child and, according to the condition, the faces of three women or three men. Their task was to identify the child's parent among the three presented adult faces. There were 28 trials (14 different girls and 14 different boys). The photographs were displayed in the same way as in the Christenfeld and Hill study: the child's face was presented in an upper position and the three adults' faces were placed beneath the child's face. The presentation positions of the adult photos were appropriately randomized. Participants were tested individually. Each were each presented with the 28 sets including one child and three possible parents in a different random order.

#### *Results*

The design of the experiment was as follows. The age of the child (one-year-old, three-year-old and five-year-old) and the sex of the parent were between-subjects factors while the sex of the child was a within-subjects factor. A 3 (age of the child) X 2 (sex of the parent) X 2 (sex of the child) ANOVA with repeated measures on the last factor revealed a significant main effect of the age of the child ( $F(2,174) = 6.614, p < .01$ ), no main effect of the sex of the parent ( $F(1,174) < 1$ ) and no main effect of the sex of the child ( $F(1,174) < 1$ ). The analysis revealed no significant interaction between the first two factors ( $F(2, 174) < 1$ ), no significant interaction between the second and the third factor ( $F(1,174) < 1$ ) and no three-way interaction ( $F(2,174) < 1$ ). The main effect of the age of the child was qualified by a significant interaction between this factor and the sex of the child ( $F(2,174) = 5.988, p < .01$ ), but the magnitude of this interaction effect is low ( $\eta^2 = 0.06$ ). This interaction was analyzed using Tukey HSD post-hoc tests. These tests showed that, while the level of parent identification from pictures of girls did not change across the three ages, it did for pictures of boys. Parent identification was better for five-year-old boys than from one-year-old boys ( $p < .0001$ ) and than for three-year-old boys ( $p < .01$ ). No significant difference appeared between one-year-old and three-year-old boys ( $p = 0.56$ ). Post-hoc tests indicated no significant effect of the sex of the child on parent identification at age one, three or five (all  $p$ 's  $> .10$ ).

A control analysis taking the items as the random factor was also carried out. This analysis did not reveal any significant main effect of the sex of the child ( $F(1,26) < 1$ ), of the sex of the parent ( $F(1, 26) < 1$ ) and of the age of the child ( $F(2, 52) = 1.982, p = 0.15$ ). Nor did it reveal any interaction effect (all  $p$ 's  $> .20$ ). The results of this control analysis confirmed that the

---

<sup>1</sup> Originally appeared in *Evolution and Human Behavior*. Reprinted with permission.

significant interaction effect obtained in the preceding analysis was not a strong effect.

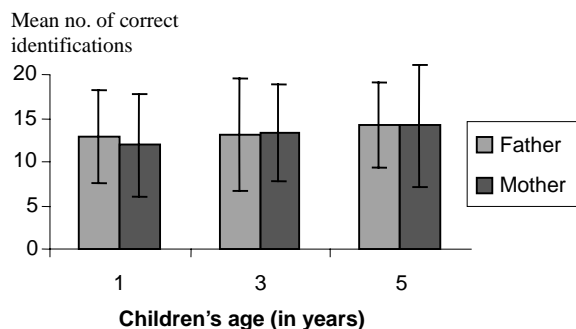


Figure 1. Mean number of correct identifications (out of 30) of parents of children at various ages (one-SD error bars). There is no significant difference in the level of correct identification of mothers versus fathers based on children's faces.

To reiterate, our analyses showed *no significant difference between the level of correct identification of mothers and the level of correct identification of fathers from children's faces* (Figure 1 and Table 1). Christenfeld and Hill (1995) did not perform a direct comparison between levels of identification of mothers and fathers. They simply compared the level of identification of mothers and fathers to the chance level of 33.3 percent by means of student t-tests, the items being the random factor. We also carried out this analysis for our data by comparing the mean number of identifications to chance ( $1/3 \times 30$  subjects = 10). At all ages tested, our results indicate that, while correct identification of mothers and fathers was significantly, although not overwhelmingly, higher than chance, there is no significant difference between the degree of father-identification and mother-identification.

It is particularly important to note that while the degree of correct association of parents with children is anywhere between 7 and 14% higher than chance, it remains surprisingly poor. In all cases, *non-identification exceeds 50%*.

## Discussion

Present results do not replicate those of Christenfeld & Hill's (1995) study. Young children aged 1, 3 and 5 do not appear to resemble their fathers significantly more than they resemble their mothers.

It could be objected that the sample of faces used in this experiment is not a representative one. In fact, there is no clear reason why our sample of items would not be representative of the larger Caucasian population in general, and, crucially would be less representative than Christenfeld and Hill's original sample. Indeed, we used photographs from 28 families, whereas Christenfeld and Hill's stimuli were drawn from 24 families. Our stimuli were collected in the same way as those in the Christenfeld and Hill study, i.e. by asking

friends, colleagues and acquaintances for photographs. We do not see any a priori reason why such a procedure would lead to the construction of an unrepresentative set of faces.

Age	Parent	Mean no. of identifications (SD in parentheses)	Student <i>t</i>	<i>p</i>
1	Father	12.893 (5.363)	2.854	<.01
	Mother	11.929 (5.937)	1.719	<.05
3	Father	13.178 (6.464)	2.602	<.01
	Mother	13.321 (5.644)	3.114	<.01
5	Father	14.143 (4.859)	4.512	<.001
	Mother	14.143 (6.996)	3.134	<.01

Table 1. Mean number of correct identifications (out of 30) as a function of the children's age and the parent's sex. Standard deviations are in parentheses. Note the absence of any significant difference in levels of correct identification of fathers and mothers based on a child's facial appearance.

Is our failure to replicate Christenfeld & Hill possibly attributable to an inappropriate sample of pictures that allowed *no* null hypothesis to be rejected? This would be very unlikely, because in *all six cases* of mothers and fathers for 1, 3, and 5 year old children, we found that the resemblance of parent to child is, as one would expect, significantly better than chance. In short, our sample *did* demonstrate a significant resemblance between parents and children, but *not that there was a significantly greater* resemblance between fathers and their children compared to mothers and their children. This means that our failure to find a significant difference in the resemblance of fathers-to-children versus mothers-to-children was not simply due to an insufficient amount of detail to be able to make resemblance assessments of any kind.

## Experiment 2

### Subjects

Forty-four undergraduate volunteers (22 females, 22 males) participated in the experiment.

### Stimuli and materials

Thirty-two Caucasian families provided three photographs: one photograph of a child at one year, one photograph of the mother and one photograph of the father taken when the child was approximately one year old. For sixteen families, the child was a girl, for the other sixteen families the child was a boy. The stimuli presented to subjects were scanned versions of these photographs (size = 5.5 x 4.5 cm) of faces. None of the faces had glasses, beards or moustaches. Stimuli were presented using E-prime® on a PC.

### Procedure

On each trial, participants were presented with the face of a child and, according to the condition, the faces of three women or three men. All photographs were displayed on the computer screen. Their task was to identify as quickly and as accurately as possible the child's parent among the three presented adult faces. Participants responded by pressing a key on the numeric keypad of the computer keyboard (1 = left photo choice, 2 = middle photo choice and 3 = right photo choice). There were 32 trials (16 different girls and 16 different boys). The position of the real parent among the three adult photos was appropriately randomized. Each participant was presented with the 32 sets consisting one child and three possible parents in a different random order. The experiment was preceded by a short practice session using four trials that were not employed later in the experiment.

### Results

The experiment had a repeated measures design with two factors : the gender of the parent and the gender of the child. One item was removed because the proportion of correct mother-infant identification was below the 2SD cutoff.

The first dependent measure was the proportion of correct identification of the parent. A 2 (gender of the parent) X 2 (gender of the child) ANOVA with repeated measures on both factors revealed no main effect of the gender of the parent ( $F(1,43) < 1$ ), no main effect of the gender of the child ( $F(1,43) < 1$ ), and no interaction effect ( $F(1,43) < 1$ ). See Table 2.

Parent	Infant gender	
	Girl	Boy
Mother	0.41 (0.16) 2.72 secs (0.85)	0.39 (0.16) 2.72 secs (0.88)
Father	0.38 (0.19) 2.82 secs (1.0)	0.38 (0.19) 2.73 secs (1.2)

Table 2. Mean proportions of correct identifications of parents, and mean correct RTs (in seconds) as a function of the gender of the child and the parent. Standard deviations are given in parentheses.

A control analysis taking the items as the random factor was also carried out and revealed the same pattern of results: none of the main and interaction effects were significant (all  $F_s < 1$ ).

The level of identification of mothers and fathers was also compared to the chance level of 33.3 percent by means of student t-tests, the subjects being the random factor. The overall mean level of correct identification of both mothers ( $m = 0.397$ ;  $t(43) = 3.696$ ;  $p < .001$ ) and fathers ( $m = 0.381$ ;  $t(43) = 2.436$ ;  $p < .05$ ) was significantly higher than chance.

The second dependent measure was the response latency (RTs). Mean correct recognition RTs to the

mother and the father were computed for each subjects and submitted to a 2 (gender of the parent) X 2 (gender of the child) ANOVA. Five subjects were excluded from this analysis: two subjects whose RTs were particularly slow (RTs > 2 SD from the sample average) and three subjects who did not provide any correct recognition in one subcategory of items. This analysis revealed no main effect of the gender of the parent ( $F(1,38) < 1$ ,  $p = 0.49$ ), no main effect of the gender of the child ( $F(1,38) < 1$ ;  $p = 0.53$ ), and no interaction effect ( $F(1,38) < 1$ ;  $p = 0.71$ ). See Table 2.

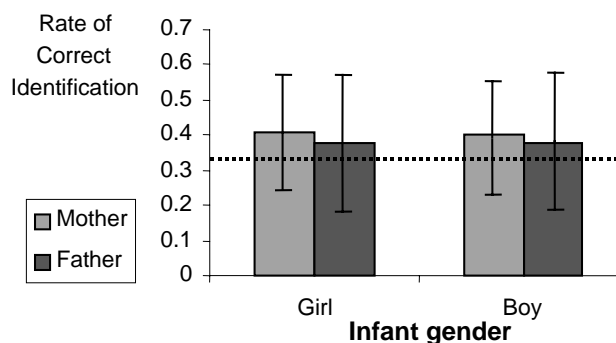


Figure 2. As in the first experiment, Exp. 2 shows no significant difference between child-father and child-mother rates of correct identification (1 SD error bars). The dotted line indicates chance level of identification.

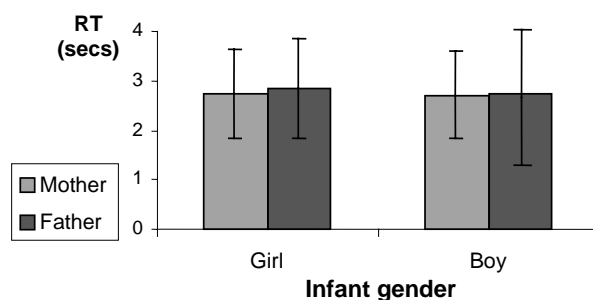


Figure 3. There is no significant difference between child-father and child-mother reaction times for correction identifications (1 SD error bars).

A control analysis taking the items as the random factor was also carried out and revealed the same pattern of results: none of the main and interaction effects were significant (all  $F_s < 1$ ).

### Discussion

This second experiment was, above all, designed to repeat and refine the results obtained in Experiment 1. Entirely new stimuli (i.e., black-and-white photographs of adults and infants) were used. And, unlike the first experiment in which children of ages 1, 3 and 5 were used, here we focused exclusively on one-year old infants. (This was because the claim of Christenfeld and Hill bears specifically on one-year old infants: it is at that age that infants supposedly resemble their fathers

more closely than their mothers.) The stimuli in this experiment were presented on a computer monitor instead of using the actual photographs, as in the first experiment. And, of course, all of the participants were different from the first experiment.

All of the results of the first experiment were reproduced in this second experiment. As in the first experiment, we found that the level of correct identification of infant-father pairs was not significantly higher than that of infant-mother pairs. We also found that, as in the first experiment, that both the levels of infant-father and infant-mother identification are significantly above chance, as one might expect. Finally, we found no significant difference in the reaction times for correct responses for both infant-father and infant-mother pairs. In other words, there is no significant difference in the speed with which people can correctly identify an infant's mother or its father.

As in Experiment 1, these results are in clear contradiction with Christenfeld & Hill (1995).

## General Discussion

The evolutionary analysis of Christenfeld and Hill is based on the supposed advantage to one-year old babies of looking more like their father than their mother in order to encourage greater resource investment on the part of the father, thereby improving their chances of survival. This theory is certainly appealing, but we believe it is undermined by a number of considerations that we will review below.

We must begin by returning to the fundamental postulate of Darwinian evolution, namely that the ultimate winners in the game of evolutionary competition are those individuals who succeed in passing on the greatest amount of their genetic material to subsequent generations. Now, there would be little obvious evolutionary pressure for a child to resemble its mother, since the maternity of a child is never in doubt. This allows us to take the degree to which a child resembles its mother as a baseline of parent-child resemblance.

The essence of the argument against greater resemblance between fathers and their infants as opposed to infants and their mothers is based on the following simple observation: If father-child resemblance was strong enough to enable a father to be certain when a child was his, it would presumably also permit a father to identify that a child *was not his* (Brédart & French, 1999). Now, in the event that a child was not his, the chances of his withholding resources from the child (or very possibly killing the child outright) would be high. Even today, step-children are far more likely to be killed by step-parents than by natural parents. In the U.S. in 1976, for example, Daly and Wilson (1988) reported that children living with one or more substitute parents were *sixty-five times* as likely to be fatally abused as children living with their

biological parents. Other studies report similar patterns of child mistreatment (for a recent short review see Daly and Wilson, 1996). Animal research has also shown the prevalence of infanticide by male rodents, carnivores and, in particular, primates (Hdry, 1979).

For much of the two-million year pre-agricultural course of human existence, three important conditions prevailed: male parental investment (Trivers, 1972) was necessary to ensure the survival of offspring, males were unable to completely control all possible sexual contact of their mates, and, finally, few individual males were able to provide resources for many females (Symons, 1987). Under these conditions, if babies had unambiguously resembled their fathers, a highly monogamous society would likely have emerged because few females would have risked the possibility of fathering another male's child, given that the bastard child would have been recognized as not belonging to her "official" (investing) mate (see also comments by R. Dawkins and other discussants following a paper by Wilson and Daly, 1997) and would thus have risked maltreatment and, quite possibly, death. In short, few females would have engaged in extra-pair copulation (EPC). However, in reality, this is contradicted by the fact that occasional EPCs by both sexes seems to be a universal feature of monogamous species (Mock and Fujioka, 1990), including humans. For example, rates of human misassigned paternity (based on blood typing tests) of 6-30% have been reported in studies done in southern England (Edwards, 1957; Philipp, 1973), 9% among the Venezuelan Yanomanö (Neel and Weiss, 1975; Smith, 1984), and 10% in rural Michigan, (Smith, 1984). Baker and Bellis (1995) have estimated a cross-cultural median EPC figure of 9%, with a range from 1.4-30%. Further, in a survey of 2078 English women, Bellis and Baker (1990) found that extra-pair copulations are significantly more likely to be timed just before ovulation than in-pair copulations. From his model of parent-infant resemblance, Pagel (1997) recently concluded that "even small amounts of paternity uncertainty are sufficient to select against parent-infant resemblance" (p.973).

Moreover, if relatively high father-child resemblance were the norm, evolution would tend to produce progressively greater degrees of father-child resemblance because any degree of resemblance significantly below that norm would engender suspicions on the part of the resource-providing male concerning the child's paternity. This would likely lead to a higher degree of resource-withholding than if the child had unambiguously resembled the father, which would ultimately translate into a lower rate of survival among those children who did not closely resemble their fathers. In other words, once evolution had established a trend of father-child resemblance in excess of baseline resemblance, there would be evolutionary pressure towards ever greater resemblance. One would therefore expect, after three

million years of selection, that there would now be a very *strong* tendency of father-child resemblance with respect to mother-child resemblance. However, our results — as well as those by Christenfeld and Hill — demonstrate that this is not the case. Indeed, in Christenfeld and Hill's data correct identification of fathers from infant faces occurred only in 49.2 percent of cases. In the two experiments reported in the present paper, the mean rate of correct identification for the father's of one-year-old children was only 10% higher than chance in the first experiment and 5% higher than chance in the second.

For these reasons, we believe that the original results reported by Christenfeld and Hill (1995) of greater father-child than mother-child resemblance in young children are most likely incorrect.

### Conclusion

We believe that the experimental results presented by Christenfeld & Hill (1995) are most likely in error. We have attempted on three separate occasions to reproduce their results, each time with new photographic stimuli and new participants. We have used two separate measures (percentage of correct identifications and reaction times for correct identifications). In all cases, we have seen no evidence whatsoever of the results reported in their paper. In this paper we report two of our experiments. In addition, we provide a theoretical justification of the outcome of our experiments. We believe that the evidence presented in this paper casts serious doubt on the originally published study by Christenfeld and Hill.

### Acknowledgments

The present paper was supported in part by Belgian Federal research grant IUAP P4/19 and research grant ARC 99/04-246 from the government of the French-speaking community of Belgium.

### References

Baker, R. & Bellis, M. (1995). *Human Sperm Competition*. London: Chapman and Hall, 142-144.

Bellis, M. & Baker, R. (1990). Do Females Promote Sperm Competition? Data for Humans. *Animal Behaviour*, 40, 997-999.

Bennett, E. & Calvin, M. (1964). Failure to train Planarians reliably. *Neurosciences Research Program Bulletin*, July-August, 3-24.

Bredart, S. & French, R.M. (1999). Do babies resemble their fathers more than their mothers? A failure to replicate Christenfeld & Hill (1995). *Evolution and Human Behavior*, 20(3), 129-135.

Byrne et al., (1966). Memory transfer. *Science*, 153, 658-9.

Christenfeld, N. & Hill, E. (1995). Whose baby are you? *Nature* 378: 669.

Daly, M. & Wilson, M. (1988). *Homicide*. Hawthorne, NY: Aldine de Gruyter.

Daly, M. & Wilson, M. (1996). Violence against stepchildren. *Current Directions in Psychological Science*, 5, 77-81.

Edwards, J. H. (1957). A critical examination of the reputed primary influence of ABO phenotype on fertility and sex ratio. *British Journal of Preventive and Social Medicine*, 11, 79-89.

Gaulin, S. & Schegel, A. (1980). Paternal confidence and paternal investment: a cross-cultural test of a sociobiological hypothesis. *Ethology and Sociobiology*, 1, 301-309.

Hdry, S. (1979). Infanticide among Animals: A Review, Classification and Examination of the Implications for the Reproductive Strategies of Females. *Ethology and Sociobiology*, 1, 13-40.

Hilgard, E., Atkinson, R. & Atkinson, R. (1971). *Intro. to Psychology*. Harcourt, Brace Jovan. 232-233.

McConnell, J. (1962). Memory transfer through cannibalism in Planarians. *J. Neurophys*, 3, 42-8.

Mock, D. & Fujioka, M. (1990). Monogamy and long-term pair bonding in vertebrates. *Trends in Ecology and Evolution*, 5, 39-43.

Munn, N., Fernald, L. & Fernald, P. (1969). *Intro. to Psychology*. Houghton Mifflin. 268-269.

Neel, J.V. & Weiss, M. (1975). The genetic structure of a tribal population, the Yanomama Indians. XIII. Biodemographic studies. *American Journal of Physical Anthropology*, 42, 25-51.

Pagel, M. (1997). Desperately concealing fathers: a theory of parent-infant resemblance. *Animal Behaviour*, 53: 973-981.

Philipp, E. (1973). Discussion: moral, social and ethical issues. In *Law and Ethics of A.I.D. and Embryo Transfer*. Ciba Foundation Symposium (Vol. 17), G.E.W. Wostenholme and D.W. Fitzsimons (Eds.). Amsterdam: Elsevier, Excerpta Medica, North-Holland, 63-66.

Smith, R. L. (1984). Human sperm competition. In *Sperm Competition and the Evolution of Animal Mating Systems*, R.L. Smith (Ed.). London: Academic Press, 601-660.

Symons, D. (1987). An evolutionary approach: can Darwin's view of life shed light on human sexuality? In *Theories of Human Sexuality*, J. H. Geer & W. O'Donohue (Eds.) NY: Plenum, 91-125.

Trivers, R. (1972). Parental Investment and Sexual Selection. In *Sexual Selection and the Descent of Man*, B. Campbell (Ed.). Chicago: Aldine de Gruyter, 136-179.

Wilson, M. & Daly, M. (1997). Relationship-specific social psychological adaptations. In *Ciba Foundation Symposium 208: Characterizing Human Psychological Adaptations*, G.R. Bock and G. Cardew (Eds.). Chichester, UK: John Wiley & Sons, 253-268.



# Memory versus Perceptual-Motor Tradeoffs in a Blocks World Task

Wai-Tat Fu (wfu@gmu.edu)  
Wayne D. Gray (gray@gmu.edu)  
Human Factors & Applied Cognition  
George Mason University  
Fairfax, VA 22030 USA

## Abstract

Using information *in-the-world* as *external memory* may be a low-cost alternative to internal memory: storage is free, and retrieval is often quick (involving a saccade) and reliable. However, when the cost of accessing external information increases, *in-the-head* storage and retrieval may become the least-cost solution. We employ the rational analysis framework (Anderson, 1990) to study the effect of varying the cost of information access on interactive behavior. Increasing the cost of information access induced a switch from information in-the-world (the perceptual-motor strategy) to information in-the-head (the memory strategy). Given the effort and unreliability of internal storage, the threshold for switching from an in-the-world to an in-the-head strategy is surprisingly low.

## Introduction

Information stored *in-the-world* can be considered as *external memory* (O'Regan, 1992). Information is retrieved from external memory via visual perception as rendered by the appropriate saccades and fixations. Recent research has suggested that when information in-the-world is readily accessible, internal storage is not needed (Ballard, Hayhoe, & Pelz, 1995); perceptual-motor strategies will be deployed to reacquire information as needed. However, when the cost of information access was increased from a simple saccade to a head movement, the perceptual-motor strategy was replaced with a strategy that placed task-relevant information into working memory (Ballard et. al., 1995). This suggests that the decision to store information in-the-head versus in-the-world is sensitive to least-cost considerations.

## The rational analysis framework

One explanation for this kind of trade-off was given by Anderson (1990). Anderson casts human memory as an optimization process. In his rational analysis framework, the goal of human memory is to retrieve knowledge that would allow us to perform the task we are currently facing. The optimization process maximizes the *expected utility* of the memory system by balancing the cost of memory search against an assumed constant expected gain<sup>1</sup> of retrieving a

desired memory item for the current task. A clear cost of memory search is time (and possibly a metabolic cost associated with time). Under Anderson's rational analysis framework, the human memory system would search a memory structure until the probability of getting the desired memory item (the expected gain) is lower than the cost of further search (i.e., when the expected utility becomes negative).

If information in the external environment can be considered as an external memory store, the cost in searching for the relevant information in the external environment can be taken as the "memory" search cost as in Anderson's rational analysis framework. In most tasks, the information stored in the external environment is continuously available (high expected gain and fixed expected cost).

If the only cost associated with internal memory were a search cost, then we would expect that in most situations internal search would be faster than external search. However, for internal memory a significant additional cost is internal storage (encoding). Storage costs would seem to be particularly problematic in the type of real-time, dynamic tasks studied by Ballard and associates. For example, in a task that required frequent memory updates, Altmann and Gray (2000) estimated that approximately 10 cycles of encoding with a duration of approximately 100 msec per cycle are needed to encode an item so that it can be retrieved 5,000 msec later. In contrast, the time for a saccade and dwell is typically estimated as 230 msec (Card, Moran, & Newell, 1983, pp. 25-28).

Compared to a memory strategy that includes encoding plus retrieval, a saccadic eye movement to a known location has a much lower time cost. Therefore, under many conditions, the expected utility of using the external environment as external memory is much higher than that of the internal human memory system. However, when the cost of information access from the external environment is high enough, the expected utility of external memory would be lower than that of internal memory. In this case, the rational analysis framework would predict a shift from external memory to internal memory. In other words, people would be more likely to adopt a memory strategy than a perceptual-motor strategy.

Unlike retrieving an item from a known external location with a saccade and dwell, retrieving an item from memory is

---

<sup>1</sup> The expected gain is defined as the product of P and G, where P is the estimated probability that the target memory item can be found, and G is the gain associated with retrieving the target

---

memory item. If C is the memory search cost, then expected utility  $E = PG - C$ .

subject to interference from previously encoded as well as other currently encoded items. If we make the additional assumption that the strength of an encoded trace fluctuates as a function of noise (Altmann & Gray, 1999; Anderson & Lebière, 1998), then retrieval from memory may take longer and is more error prone than the corresponding retrieval from the external environment. The rational analysis framework suggests that searching for an item should stop as soon as the expected gain from finding the item is less than the cost of searching. Therefore given an assumed constant expected gain, the higher the search cost of a memory item, the fewer items would be searched for and inspected before the memory system would stop searching. Since the more items the memory system considers, the more likely that the target item can be found (thus improving accuracy), increasing the search cost implies a decrease in accuracy; that is, an increase in errors. Therefore the rational analysis framework not only predicts that increasing the cost of information access in the external environment would induce a shift from external memory to internal memory, but also an increase in errors.

In this paper, we employ the rational analysis framework to study the effect of varying the cost of information access on interactive behavior. Specifically we test two predictions that we have derived from the rational analysis framework: that an increase in the perceptual-motor cost of information access will induce a shift from an external to an internal memory strategy, and that this switch will occur even though the internal search is difficult and error prone.

## Experiment

The blocks world task is based on the paradigm used by Ballard et al., (1995). The task is to copy a pattern of colored blocks shown in the *target* window to the *workspace* window, using the colored blocks in the *resource* window (for our version see Figure 1).

To do the task, participants have to remember three pieces of information: (a) the color of the block to be copied, (b) the position of the block to be copied, and (c) which blocks have or have not been copied. The first two pieces have to be obtained from the target window whereas the third piece has to be obtained by comparing the target window with the workspace window.

Ballard reported a point-of-gaze (POG) sequence of target window, resource window, target window, workspace window (TRTW). The implication of this sequence is that during the first POG to the target window (T) subjects encoded the color of the block and then picked up a block from the resource window (R). On the next POG to the target window (T) subjects encoded the block's location in the pattern. They then moved to the workspace window (W) and placed the block in the appropriate location.

As the effort needed to acquire information from a window increased from a POG to a head movement the sequence tended to change to TRW. In this case, the implication is that subjects encode both the color and the position during the first (and only) POG to the target window (T). The encoded trace persists as the subject acquires the block from the resource window (R) and places it in the workspace window (W).

Unlike Ballard et al., in our Block World task all three windows were covered by gray boxes. Throughout the task only one of the windows could be uncovered at a time. The resource and workspace windows were uncovered by moving the mouse cursor into the window. They were covered again when the mouse cursor left the window. The effort required to uncover the target window varied between each of our three conditions.

To access the information in the target window participants could adopt either a predominately perceptual-motor or a predominately memory strategy. As per Ballard et al.'s TRTW strategy, the predominately perceptual-motor strategy would entail one uncovering at the target window to obtain color information and another to obtain position information. In contrast, a predominately memory strategy (TRW) would entail one uncovering at the target window to obtain both color and position. Deciding which blocks remained to be copied would entail a second set of strategies. On these strategies Ballard is silent. However, a predominately perceptual-motor strategy might entail multiple quick uncoverings between the target and workspace window. A predominately memory strategy might entail encoding the color and position of multiple blocks at one glance.



Figure 1. The blocks world task. In the actual task all windows are covered by gray boxes and at any time only one window can be uncovered. The window at the top left is the target window, at the bottom the resource window, and at the top right the workspace window.

## Method

### Participants

Forty-eight George Mason University undergraduates participated in the study for course credit and were randomly assigned to one of the three experimental conditions.

### Equipment and software

The experiment was written in Macintosh Common Lisp and was conducted with a Macintosh PowerPC connected to

an extended keyboard, a mouse, and a 17-inch monitor. All mouse movements and keypresses were recorded and saved to a log file with 16.67 msec accuracy.

The blocks (48 x 48 pixels) that constitute each pattern were randomly chosen with the constraint that no color was used in one pattern more than twice. The blocks were placed at random in the target window's 4 x 4 grid. The workspace window was the same size as the target window and contained the same, non-visible, 4 x 4 grid.

### Design and Procedure

The three conditions were designed to vary the cost of uncovering the target window. In the *low-cost* condition, participants had to press and hold down a function key. (Participants were asked to use different hands for the keyboard and the mouse.) The target window remained uncovered until they released the key, or until the mouse cursor entered either the workspace or resource window. Once the target window was covered, to uncover it again participants had to release the key and press it again (this is to avoid the strategy of holding the key down throughout the task). In the *control* condition, the conditions for uncovering the target window were the same as for the workspace and resource windows. The target window was uncovered when the mouse cursor entered the window. The *high-cost* condition was similar to the control condition, except that participants had to move the mouse cursor inside the target window and endure a one second lockout before the target window was uncovered.

To select a block, participants moved the mouse cursor to the resource window and mouse clicked on the desired colored block. The mouse cursor then changed to a small version (16 x 16 pixels) of the selected block (eliminating the need to remember its color). To place a block in the workspace window, the cursor was moved to that window (which was then uncovered), moved to the desired position, and then clicked. When the participants believed that the pattern had been copied to the workspace window, they press the "Stop-Trial" button. A feedback window indicated whether the copied pattern matched the target pattern. If the pattern was different, participants were required to go back to finish the task before they could move on to the next pattern.

At the beginning of the experiment, instructions were given and participants were led through one demonstration trial by the experimenter. Participants then completed 40 trials. The whole experiment lasted about 45 minutes.

## Results

### Trial Time

The first ten trials of the experiment were considered practice and were excluded from further analyses. Analysis of variance (ANOVA) of time for condition by trial showed significant main effects of condition ( $F(2, 45) = 9.11, p = 0.0005, MSE = 726$ ), and of trial ( $F(29, 1305) = 131.6, p < .0001, MSE = 53.2$ ). There was no interaction between trial and condition. To determine whether the main effect of conditions was solely due to the one second delay in the

high-cost condition, the per trial time in this condition was adjusted by subtracting the amount of delay for each time the target window was uncovered. After the adjustment, the main effect of condition was not significant ( $F(2, 45) = .969, p = .39, MSE = 564$ ). However, the main effect of trial remained significant ( $F(29, 1305) = 120.1, p < .001, MSE = 46.0$ ). Orthogonal linear contrasts showed a significant linear downward trend of time across trials for the low-cost condition ( $p = .0001$ ), control condition ( $p = .0001$ ), and high-cost condition ( $p = .0001$ ), suggesting speed-up across trials. No other higher order trends were significant. The interaction between trials and conditions was not significant.

### Use of the Target Window

The trial time results seem to suggest no difference between conditions. However, detailed analyses revealed the effects of the cost of information access. An ANOVA on the number of times the target was uncovered showed a significant main effect of condition ( $F(2, 45) = 10.17, p = .0002, MSE = 159$ ). Planned comparisons revealed a significant difference between the high-cost and control ( $p = .0045$ ), as well as high-cost and low-cost conditions ( $p < .0001$ ) (See Figure 2). Subjects in the high-cost condition uncovered the target window significantly fewer times than the other two conditions. However, there was no significant difference between the low-cost and control condition.

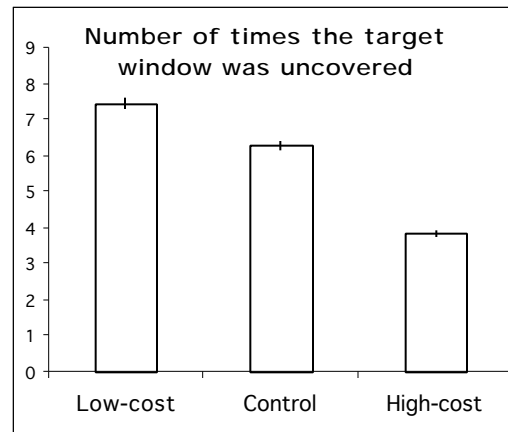


Figure 2. Mean number of times per trial that subjects uncovered the target window.

ANOVA on the time subjects spent looking at the model showed that there were significant main effects of conditions ( $F(2, 45) = 20.6, p < .0001, MSE = 300$ ), with the high-cost condition significantly spending more time than the low-cost condition ( $p < .0001$ ), or the control condition ( $p < .0001$ ). The difference between the low-cost and control condition was not significant. Overall, there was a significant downward linear trend of time spent on the target window across trials ( $p = .0002$ ). However, orthogonal linear contrasts showed that the downward linear trends for the high-cost ( $p = .02$ ) and control condition ( $p = .001$ ) were significant, but that the trend for low-cost condition ( $p = .10$ ) was not (see Figure 3).

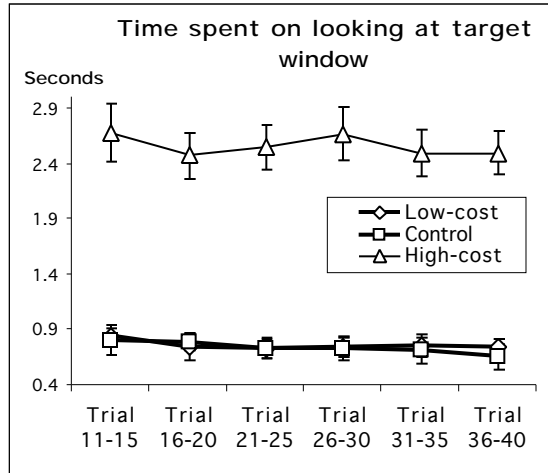


Figure 3. Time spent on looking at the target window.

For each trial, we looked at how many colored blocks the subjects copied following each of their first four accesses of the target window. We conducted a 3 x (4 x 30) ANOVA on conditions, the nth (1 to 4) uncovering of the target window, and trial (11-40). There were significant main effects of conditions ( $F(2, 45) = 19.5, p < .0001, MSE = 7.75$ ), and the nth uncovering of the window ( $F(3, 135) = 39.5, p < .0001, MSE = 13.3$ ). The interaction between conditions and the nth window uncovering was significant ( $F(6, 135) = 7.8, p < .0001, MSE = 13.2$ ) (see Figure 4). The main effect of trials was not significant. No other interactions were significant.

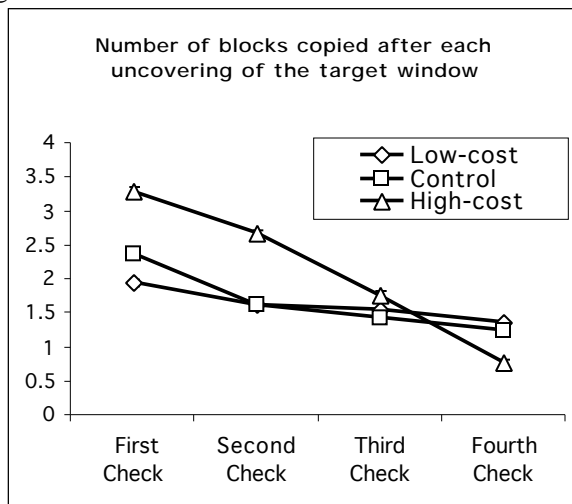


Figure 4. The number of colored blocks copied after each uncovering of the target window.

**Summary.** The analysis of how the target window was used suggest that although there was no significant difference in trial times between conditions, the strategies used by the subjects in the high-cost condition were very different from the other two conditions. Subjects in the high-cost condition uncovered the target window a fewer number of times and copied more colored blocks after each uncovering at the target window. Figure 4 illustrates the interaction between

condition and the number of blocks copied in the first four uncoverings of the target window. Clearly, the significant decrease in the number of blocks copied within a trial is at least partially due to the decreasing number that remained to be copied. However, this potential artifact does not explain the significant interaction. Subjects in the high-cost condition tended to copy more blocks in the first few accesses at the target window. In contrast, in the low-cost condition, subjects tended to copy the same number of blocks after each access. This suggests that subjects in the high-cost condition tended to adopt a memory strategy - memorizing the information of the colored blocks to reduce their reliance on the external display. On the other hand, as the cost of information access was low subjects in the low-cost condition relied more on a perceptual-motor strategy - getting the information from the display when needed.

With practice, subjects in the high-cost condition spent less time looking at the target window, but copied more colored blocks after each uncovering. This increase in efficiency (measured in terms of the time required to memorize the information of a fixed number of colored blocks) partly explains the speed-up across trials in the high-cost condition. However, the same trend was not observed in the low-cost condition. It seems that practice had no significant effect on the use of the target window in the low-cost condition (in terms of number of times they uncovered at the target window, number of colored blocks copied after each uncovering, and time spent on the target window per uncovering).

### Strategies

A finer-grained analysis is needed to understand the actual differences in strategies used. For each copied block, we extracted the sequence in which subjects uncovered windows. In the notation below, these sequences are abbreviated using the first initial of the window name. Wb indicates that they uncovered the workspace window and placed a block, Wu indicates that they uncovered the workspace window but did not place a block. For example, TRWb refers to a target-source-workspace windows sequence that ends with the placement of a block.

Table 1. Strategies used by subjects. T = uncover target window, R = uncover and pick a colored block from resource window, Wb = put selected colored block to workspace window, Wu = uncover workspace window. For example, TWuRWb represents the strategy in which the subject uncovered the target window, uncovered the workspace window, went to the resource window, picked up a colored block, and put the colored block in the workspace window.

	Low-cost	Control	High-cost
Strategy	Strategy	Strategy	Strategy
TRWb	53%	TRWb 49%	TRWb 38%
RWb	33%	RWb 38%	RWb 58%
TRTWb	5%	TRTWb 7%	TWuRWb 1%
Total	5537	5496	5752

Table 1 shows the three most common sequences used by the subjects in the three conditions (as well as the percentage of the total that these sequences represent). We

can see that the top 2 sequences (TRWb and RWb) constituted almost 90% of all the sequences used<sup>2</sup>.

Although the two dominant strategies were the same, the effect of information access cost on strategy used was clearly seen. With increasing cost, the use of the TRWb strategy decreased, while the use of RWb increased. This change of strategy nicely indicates the shift of reliance from external to internal memory. This is consistent with the results on the use of target window described above. With increasing cost of information access, subjects tended to uncover the target window less, spent more time per uncovering (time that we presume was spent encoding more information into internal memory), and used the pure memory strategy (RWb) more.

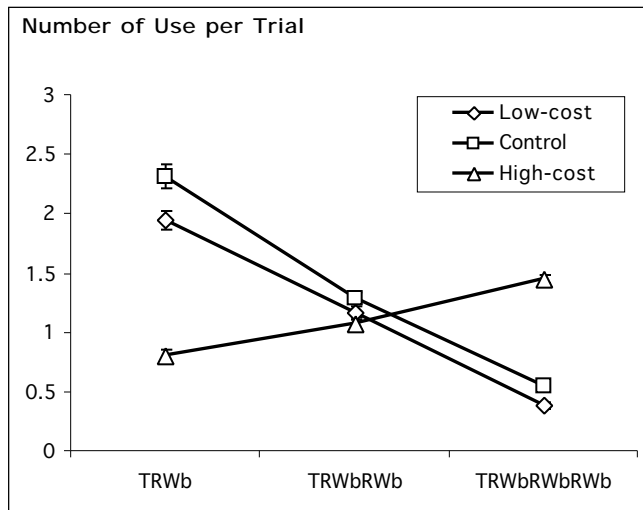


Figure 5. Number of use of strategies per trial. T = uncovering the target window, R = uncovering the resource window and picked a colored block, Wb = putting the colored block in the workspace window.

To further understand the strategy change across trials, we looked at three combinations of the two dominant strategies. We extracted the number of times subjects used either the TRWb, TRWbRWb, or TRWbRWbRWb strategy<sup>3</sup> in each of the trials. This analysis is similar to that shown in Figure 4. It captures how often subjects placed 1, 2, or 3 blocks following a single glance at the target window. A 3 x (3 x 30) ANOVA on conditions, strategies, and trials found no significant main effect of condition ( $F(2, 45) = 1.92, p = .16, MSE = 16.0$ ), nor trials ( $F(29, 1305) = .835, p < .72, MSE = 24.2$ ). However, the main effect of sequence was significant ( $F(2, 90) = 14.6, p < .0001, MSE = 19.8$ ), as was the interaction between sequence and condition ( $F(4, 90) =$

<sup>2</sup> Interestingly, TRTW, the dominant strategy described in Ballard et al (1995), was not one of the dominant strategy in this task. The difference might be that all windows in this task were covered by gray boxes, and the lower-cost saccadic strategy described in Ballard et al (1995) was not supported.

<sup>3</sup> In our categorization, these categories were mutually exclusive. Therefore, the run TRWRWRW was categorized as an instance of the TRWRWRW strategy, but was not included in the count for TRWRW or TRW.

10.9,  $p < .0001$ ) (see Figure 5). No other interactions were significant. Planned comparisons showed that all differences between the three sequences were significant ( $p < .05$ ).

The analysis further confirms the shift from the use of external to internal memory with increasing information access cost. The TRWb strategy, in contrast to the TRWbRWb and TRWbRWbRWb strategies, allowed subjects to acquire only the information necessary to copy the next colored block from the target window (external memory). Figure 5 shows that in the high-cost condition, subjects used the TRWbRWbRWb significantly more than the TRWb strategy. It suggests that subjects tended to transfer more information from the target window (external memory) to internal memory, and performed the task based on the information retained in internal memory.

### Errors and Comparisons

To test the predictions of errors from our rational analysis, we looked at the way in which different strategies affected the patterns of errors made by subjects. An error could involve placing a block with the wrong color, placing a block in the wrong position, or both. Only errors made before the subjects clicked on the "Stop-Trial" button were counted. A 3 x 30 ANOVA of the number of errors made per trial on conditions and trials was conducted. There was a significant main effect for conditions ( $F(2, 45) = 11.6, p < .0001, MSE = 1.39$ ), but not for trials ( $F(29, 1305) = 1.1, p = .33, MSE = .75$ ). The interaction between conditions and trials was not significant ( $F(58, 1305) = 1.2, p = .10$ ). The row labeled "Errors" in Table 2 shows the mean number of errors made by the subjects. The low error rates were not surprising given the simplicity of the task. However, there were significant differences between conditions. Planned comparisons showed that subjects in the high-cost condition made significantly more errors than the other two conditions ( $p < .001$  for low versus high,  $p < .0001$  for control versus high). This result supports our second prediction: that increase in information access cost increases errors.

Table 2. Mean number of errors, errors uncorrected, and comparison episodes per trial before the subjects thought they were done (when they pressed the "Stop-Trial" button.)

Dependent Variable	Conditions (cell means)		
	Low-cost	Control	High-cost
Errors	.41	.33	.68
Uncorrected	.10	.08	.14
Comparisons	.57	.31	.17

To find out whether there were differences in the subjects' ability to detect errors, we conducted a 3 x 30 ANOVA on the number of uncorrected errors after the subjects clicked on the "Stop-Trial" button by condition and trial. As suggested by the middle row of Table 2, there was no significant main effect on conditions ( $F(2, 45) = 1.26, p = .29, MSE = .365$ ). However, there was a significant main effects of trial ( $F(29, 1305) = 1.54, p = .03, MSE = .173$ ). There was a significant downward linear trend on the number of uncorrected errors across trials ( $p = .006$ ).

The decrease of the number of uncorrected errors (without any significant increase in the number of errors made) suggested that with practice, subjects in the high-cost condition became better at detecting and correcting their errors. We therefore turned our focus on how often the subjects compared the pattern in the workspace with that in the target window. The comparisons between the two windows not only served the function of error detection, but could also let the subjects keep track of what blocks had or had not been copied. As described before, this information was another critical piece of information that had to be remembered to do the task.

The number of comparison episodes was extracted. A comparison episode started when the participant went from the workspace to the target window (or vice versa) without having a block selected. Any consecutive uncoverings of the workspace and target window were counted as part of the same comparison episode. An ANOVA of the number of comparison episodes showed a significant main effect on conditions ( $F(2, 45) = 6.3, p = .004, MSE = 3.17$ ), with the low-cost condition having significantly more comparison episodes than the other two conditions (see the bottom row of Table 2). The result again confirmed our prediction: when the cost of information access is low, people prefer to adopt a perceptual-motor strategy to memory strategy.

To summarize, the memory strategy adopted by subjects in the high-cost condition seemed to be more error-prone than the perceptual strategy in the low-cost condition. However, in all conditions most errors were detected and corrected. There was therefore no significant difference in the number of uncorrected errors between conditions. The differences in number of comparison episodes revealed another aspect of the strategy difference. In the low-cost condition, subjects made many more comparisons of the workspace and the target window, further supporting the hypothesis that they did not keep track of which blocks had been copied. In contrast, the number of comparison episodes in the high-cost condition was much lower, suggesting that subjects stored the information in working memory, reducing the reliance on the external environment. A second function served by these comparisons might have been to detect and correct errors. By relying on the external display, errors could be corrected without much memorization. In contrast, subjects in the high-cost condition relied more on their memory to keep track of their task as well as to detect and correct errors. It was also shown that over practice, subjects in the high-cost condition did manage to reduce the number of uncorrected errors.

## Conclusions and Discussions

The results support two predictions derived from Anderson's (1990) rational analysis framework. Given an assumed constant expected gain, when the cost of accessing information in-the-world increased, the cost of a perceptual-motor strategy becomes greater than the cost of a memory strategy. Under such conditions, the *optimal* strategy is to

encode task-relevant information from the external environment into working memory, thereby reducing reliance on the external environment. The results indicate that when the cost of accessing external information is high, people invest more time storing information in their internal environment and rely less on the external environment to do the task. In contrast, when the cost of accessing external information is low, people spent less time encoding and rely more heavily on the external environment.

Our second prediction was upheld as well. In the high-cost condition this switch to the memory strategy was made despite its higher error rate. Indeed, the decrease in the number of uncorrected errors indicates that with practice our subjects became better at detecting and correcting errors. This finding suggests subjects were *optimizing* the strategy to reduce the overall effort required to do the task.

Under the rational analysis framework, cognition tends to optimize performance by balancing costs and benefits in a given information processing task. Our results show that the cost of information access could induce a switch from reliance on information in-the-world (perceptual-motor strategy) to in-the-head (the memory strategy). We found that although memory is a limited resource, there are conditions under which people can use it to *optimize* performance.

## Acknowledgments

The work reported was supported by a grant from the Air Force Office of Scientific Research AFOSR#F49620-97-1-0353 as well as by the National Science Foundation (IRI-9618833).

## References

- Altmann, E. M., & Gray, W. D. (1999). Preparing to forget: Memory and functional decay in serial attention. *Manuscript submitted for publication.*
- Altmann, E. M., & Gray, W. D. (2000). Managing attention by preparing to forget. Proceedings of the *Human Factors and Ergonomics Society 44th Annual Meeting* (pp. ). Santa Monica, CA: Human Factors and Ergonomics Society.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., & Lebière, C. (Eds.). (1998). *Atomic components of thought*. Hillsdale, NJ: Erlbaum.
- Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 7(1), 66-80.
- Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology*, 46(3), 461-488.

# Babies, Variables, and Relational Correlations

Michael Gasser (GASSER@CS.INDIANA.EDU)

Eliana Colunga (ECOLUNGA@CS.INDIANA.EDU)

Computer Science Department, Cognitive Science Program  
Indiana University  
Bloomington, IN 47405

## Abstract

Recent studies have shown that infants have access to highly useful language acquisition skills. On the one hand, they can segment a stream of unmarked syllables into words, based only on the statistical regularities present in it. On the other, they can abstract beyond these input-specific regularities and generalize to rules. It has been argued that these are two separate learning mechanisms, that the former is simply associationist whereas the latter requires variables. In this paper we present a correlational approach to the learning of sequential regularities, and its implementation in a connectionist model, which accommodates both types of learning. We show that when a network is made out of the right stuff, specifically, when it has the ability to represent sameness and the ability to represent relations, a simple correlational learning mechanism suffices to perform both of these tasks. Crucially the model makes different predictions than the variable-based account.

## Background

Two recent papers in *Science* have demonstrated the remarkable language learning abilities that are possessed by infants. In both cases the infants were presented with sequences of syllables embodying some sort of regularity and later tested with sequences that agreed or disagreed in certain ways with the training set. In the experiments of Saffran, Aslin, and Newport (1996), eight-month-olds heard strings of syllables consisting of randomly concatenated three-syllable “words,” sequences which never varied internally. Thus the transition probabilities within words were higher than between words. Later the infants were able to differentiate between these words and non-word three-syllable sequences which they had either heard with less frequency than the words or not heard at all. This is taken as evidence that they had picked up the statistics in the training set. Marcus, Vijayan, Bandi Rao, and Vishton (1999) presented seven-month-olds with series of three-syllable sequences separated by gaps. Each sequence consisted of two different syllables arranged in a fixed pattern, AAB, ABB, or ABA. For example, in the ABB condition, the presented patterns included sequences such as *le di di* and *ji je je*. Later the infants responded differently to novel sequences of three syllables which matched the pattern they had been trained on than to novel sequences which did not. This is taken as evidence that they had in some sense picked up the rule implicit in the training patterns.

Marcus et al. (1999) and Pinker (1999) argue that the two studies, taken together, point to at least two distinct learning mechanisms which are behind language learning. One of these, revealed in the experiments of Saffran et al. (1996),

can learn relationships such as the tendency for *ti* to immediately follow *ga*. It is sensitive to the content of the items, not caring about the similarity among different items. For Pinker (1999), this is just the *associationism* proposed in the eighteenth century by Hume and still proposed as the fundamental mechanism of the mind by modern connectionists and others. The other mechanism, revealed in the experiments of Marcus et al. (1999), can learn relationships such as the fact that the first syllable in a sequence is the same as the second but different from the third. This mechanism ignores specific content, caring only about sameness or difference. In this sense the second mechanism seems to require *variables*, placeholders which are ignorant of their specific content. For Pinker (1999), this mechanism is an instantiation of what was proposed by the early rationalists and what we think of today as “symbolic.” Thus Marcus et al. (1999) and Pinker (1999) now believe that the mind, specifically the portion of it used in language learning, is both associationist and symbolic.

The question, as Marcus et al. (1999) make clear, is not whether connectionist networks can learn to solve both kinds of tasks, but what sorts of mechanisms are required and whether these differ for the two tasks. In this paper, we present a model of the learning of regularities in patterns which accommodates both kinds of patterns in terms of **correlations**. We argue that a correlational account, to deal with the tasks in Marcus et al.’s experiment, needs two mechanisms in addition to those usually found in such accounts, neither of which amounts to explicit variables. We show how a connectionist network implementing this theory (the PLAYPEN architecture) can learn aspects of the Saffran et al. task, as well as the Marcus et al. task. What is crucial about this account is not that it handles variable-like behavior within a correlational framework but that it makes predictions that differ from the variable-based account.

## Pattern Regularity Learning

Saffran et al.’s and Marcus et al.’s experiments are not directly comparable. In Saffran et al.’s experiments, the boundaries between the patterns must be extracted, while these are provided in Marcus et al.’s task. However, both are learning tasks in which the learner is presented repeatedly with patterns consisting of sequences of syllables and extracts some sort of regularity from the sequences.

We agree with Marcus et al. and Pinker that there are other differences in what is going on in these two tasks, but we believe that both are fundamentally statistical, based on the extraction of **correlations** from input patterns. The main dif-

ference, we argue, lies in what sort of correlations: whether they are content-specific, as in Saffran et al.'s experiments, or relational and based on similarity among the elements within the sequences, as in Marcus et al.'s experiments.

We will consider tasks that are more general than those in the two original sets of infant experiments, what will refer to as **pattern regularity learning**. A learning trial for such a task consists of a pattern (not necessarily auditory) composed of elements arranged in a particular way (either sequentially or spatially), and the regularity consists of tendencies for patterns to resemble each other in particular ways. Resemblances between patterns make reference to the **position** of elements within their patterns, where position may be defined spatially or temporally. Regularity could be concerned only with a single pattern position and not with intra-pattern relationships; for example, *all patterns begin with ba*. But we will only be concerned with regularities that make reference to intra-pattern relationships, as was the case in both sets of infant experiments.

### Content-Specific Regularities

In Saffran et al.'s experiments, the resemblances between patterns concern the **specific content** of the patterns. That is, it is particular syllables which are involved in the regularities; certain combinations of syllables tend to recur. The simplest content-specific regularities (other than those that make reference to only a single pattern element) are those involving **pairwise** co-occurrences of specific elements or element features. Examples of such regularities are the following: *ba tends to be followed by gu*; *syllables beginning with b tend to be followed by syllables beginning with g*.

But the regularities in Saffran et al.'s experiments are more complex than these. Rather than simple pairwise regularities, the regularities concern co-occurrences of pairwise co-occurrences. Examples of such **higher-order** regularities are: *when gu is preceded by ba, it tends to be followed by li*; *when a syllable beginning with g is preceded by a syllable beginning with b, it tends to be followed by a syllable beginning with l*.

Not surprisingly, these statistical, content-specific regularities can be handled in a straightforward fashion in connectionist networks. Weights in most connectionist networks represent **correlations** between elements, and the regularities we have been describing are just that. However, correlations between correlations, as in the higher-order regularities, require "handle" units responsible for pairs of particular elements. These handle units can then be joined by connections whose weights encode the higher-order correlations. Figure 1 shows a network of this type. The network is of the attractor (generalized Hopfield) type, and weights are adjusted using the Contrastive Hebbian Learning algorithm (Hopfield, 1984; Movellan, 1990). For simplicity's sake, we assume separate units for the different pattern positions, ignoring the (non-trivial) problem of how element representations are shared across different positions, and we consider only the case of patterns consisting of three elements. Pairwise regularities are represented by strong weights joining pairs of PATTERN units to single CORRELATION units. Higher-order regularities are represented by strong weights on connections joining CORRELATION units. Note that this approach assumes

that higher-order regularities presuppose the pairwise regularities which they are built on. Note also that when there are multiple higher-order regularities, as in Saffran et al.'s experiments, for example, the CORRELATION layer permits these different regularities to be kept separate: one set of units and connections might represent the *ba gu mi* pattern, another the *vi ja lo* pattern.

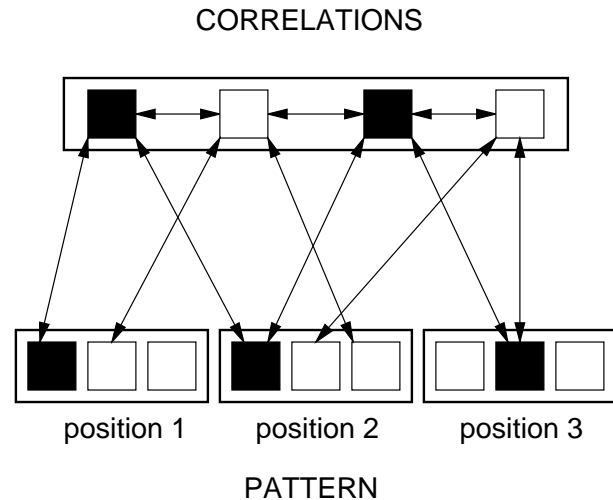


Figure 1: Network for learning content-specific regularities. Only some units and connections are shown.

Just what gets learned by such a network and how it generalizes depend on how the pattern elements are represented. We assume multiple levels of representations differing in coarseness. That is, at the least coarse level, the elements are represented in terms of the largest number of classes; at the most coarse level, they are grouped in terms of a small number of classes. Representations in connectionist networks also differ in the extent to which they are distributed vs. local. Assuming local representations for the sake of simplicity, syllables might be represented at multiple levels of coarseness as shown in Figure 2. Thus the syllable *bis* turns on a unit specific to that syllable, a unit responding to all syllables beginning with *b* and a unit responding to all consonant-vowel-consonant syllables.

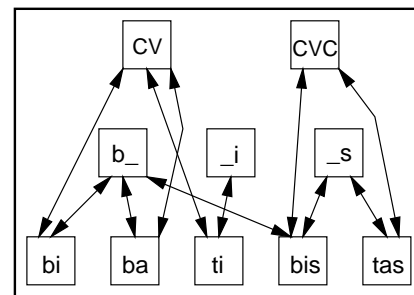


Figure 2: Representation of syllables at multiple levels of coarseness. Only a few units are shown. Arrows represent excitatory connections joining units at different levels of coarseness. Not shown are inhibitory connections forcing winner-take-all at a given level.



## Relational Regularities

Alternately, regularity within a set of patterns may be in terms of the similarity of elements within patterns; that is, the regularity may be **relational** rather than content-specific. Again the regularities may be pairwise or higher-order. Examples of pairwise relational regularities are the following: *the first element is the same as the second element; the first element tends to begin with the same consonant as the second element; the first element is different from the second.* Examples of higher-order relational regularities are the following: *the first element tends to be the same as the second element and different from the third; when the first element begins with the same consonant as the second element, the second element has the same vowel as the third.*

In these terms, then, Marcus et al.'s experiments involved both pairwise and higher-order relational regularities, as well as pairwise and higher-order content-specific regularities, though only the relational regularities are reflected in the test items.

In what follows, we discuss how relational regularities, as well as content-specific regularities, are handled within the PLAYPEN architecture.

## Accommodating Relational Regularities in a Connectionist Network

Our claim is that relational regularities, like content-specific regularities, are correlations, that is, that they involve statistical patterns of co-occurrence. Further we show how relational correlations can be learned in a connectionist network that differs from more conventional networks in that it has an explicit means of representing and learning about similarity/difference. This requires two augmentations to conventional networks: (1) a second dimension (the "binding" dimension), in addition to activation, along which units vary, and (2) "handle" units which respond to either sameness or difference on the binding dimension.

We view the task presented to the learner in Marcus et al.'s experiments as one of **grouping**, a fundamental aspect of all perceptual processing, both by humans and machines. Presented with a visual or auditory scene, people attempt both to segment it into distinct regions and to group regions together. They segment and group by making use of featural similarity, proximity, and common fate, as well as top-down knowledge of the domain. For segmentation, proximity obviously plays a large role, but for grouping, featural similarity may override proximity. Thus in rhythm perception, where grouping has been studied extensively (Handel, 1989), two elements that are separated by another may be grouped together because of their similarity to each other on some dimension. While segmentation and grouping are in some sense opposing processes, both amount to the **binding** together of regions that would otherwise not be associated with one another.

Thus any cognitive architecture that handles segmentation or grouping must offer a solution to the "binding problem," the problem of how to represent the short-term situation in which distinct cognitive units are treated as part of the "same thing." This problem has been discussed extensively in recent connectionist literature, and a family of related connectionist solutions has been proposed (Shastri & Ajjanagadde, 1993). All of these involve the augmentation of conventional archi-

tectures and algorithms with a further dimension in addition to activation along which processing units can vary. We will refer to this as the "binding dimension." Binding two units then corresponds to coincidence of those two unit's values on the binding dimension. Most often the binding dimension involves the firing of units, and binding itself is synchronization of firing (Hummel & Biederman, 1992; Mozer, Zemel, Behrmann, & Williams, 1992; Shastri & Ajjanagadde, 1993; Sporns, Gally, Reeke, & Edelman, 1989). In PLAYPEN we make use of a simpler approach: alongside its activation, each unit is characterized by an **angle**, ranging from 0 to  $2\pi$  radians. The particular value taken by a unit's angle is not what is relevant; it is its value relative to that of other units in the network. Units with similar angles are temporarily "bound" together, treated as "the same thing"; units with very different angles (differences close to  $\pi$  radians) are treated as "different things."

To permit the representation and learning of relational correlations, we need one further augmentation. Rather than taking the form of simple connections between units, relational correlations are implemented via "handle" units called **relation units**. These are of two types, **sameness units**, which tend to be activated if their input units are activated and have similar angles, and **difference units**, which tend to be activated if their input units are activated and have different angles. Each of these units represents a pairwise relational correlation of one type or the other, and the connections joining these units represent higher-order relational correlations. Thus the architecture we proposed for learning content-specific correlations (Figure 1) becomes that shown in Figure 3 for relational regularities. Again the network is of the attractor type. We have modified the standard input and activation functions and the Contrastive Hebbian Learning algorithm (Movellan, 1990) to accommodate angles and relation units. For details, see Gasser and Colunga (1998).

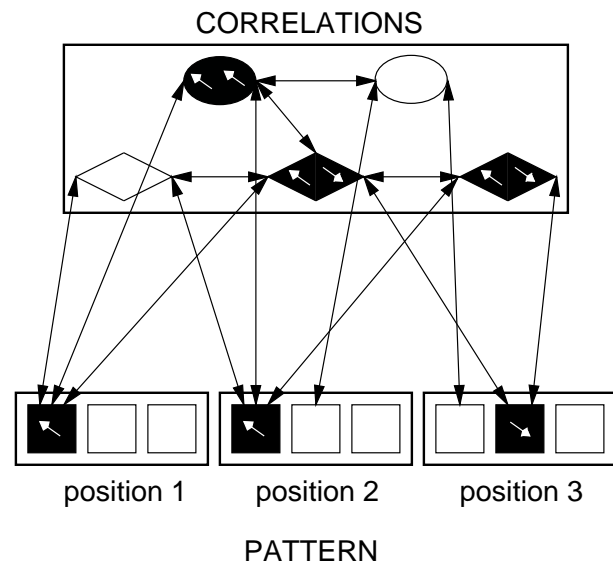


Figure 3: PLAYPEN network for learning relational regularities. Only a few units are shown. Difference relation units appear as diamonds, sameness relation units as ovals. Unit angles are indicated by arrows. A single unit within each pattern position has been activated, leading to the activation of some relation units.

Note that each unit in this network (as in the network in Figure 1) has specific content, but in addition, at any point in time, through its angle, each unit also represents a hypothesis about how the elements in the pattern are to be grouped.

### Simulation of Marcus et al.'s Experiment

Now consider again the task of Marcus et al.'s experiment. First, we agree with Seidenberg and Elman (1999) that knowledge about syllable similarity would have been learned prior to the experiment so should already be in place in the architecture. For the PLAYPEN model, this knowledge takes the form of connections (via sameness and difference units) representing the similarity or difference between syllables or syllable features. When the units representing pairs of syllables are clamped in the PATTERN layer, that is, when their activations are fixed at some positive value but their angles are still allowed to vary, these connections cause similar syllables to have the same angle and different syllables to have different angles.

We again assume a range of degrees of coarseness in syllable encodings and, for simplicity, local encodings. The presentation of a pattern, say, *le le di*, takes the form of the clamping of PATTERN units corresponding to these syllables in the relevant sequential positions. Syllable units at greater degrees of coarseness are activated (inhibitory connections between incompatible syllable units prevent all syllable units from being activated as a result of feedback from the coarse units). Further because of the built-in (or previously learned) relational connections implementing similarity, the angles of the syllables take on a pattern representing the *grouping* of the pattern elements: the first two elements make up one group, the third element another. The activated PATTERN units cause particular CORRELATION units to be activated. For example, the difference unit representing *le* in second position and *di* in third position and the difference unit representing some CV syllable in second position and some CV syllable in third position are both activated. Contrastive Hebbian Learning results in the strengthening of connections both into and between the activated CORRELATION units, as well as possibly the weakening of other connections that are not joined by activated units. Figure 4 shows some of the units and connections that are involved.

We simulated Marcus et al.'s task by training networks of this type on one of the three grammatical rules: AAB, ABA, or ABB. In each case, the set of training patterns consisted of four different syllable sequences, each formed by randomly combining syllables following the appropriate grammatical rule. Each network was trained on 50 repetitions of the training set.

The networks were then tested on 12 sequences, four each of the three kinds of grammatical rules, by clamping the units corresponding to each sequence. Each of the test sequences was novel; that is, it was formed by combinations of syllables that had never been seen before.

Since training the network leads to the strengthening and weakening of connections into and within the CORRELATIONS layer, test patterns should result in more activation on the CORRELATIONS layer if they are consistent with the training set. Thus familiarity with a test pattern was measured in the network as activation of the CORRELATIONS

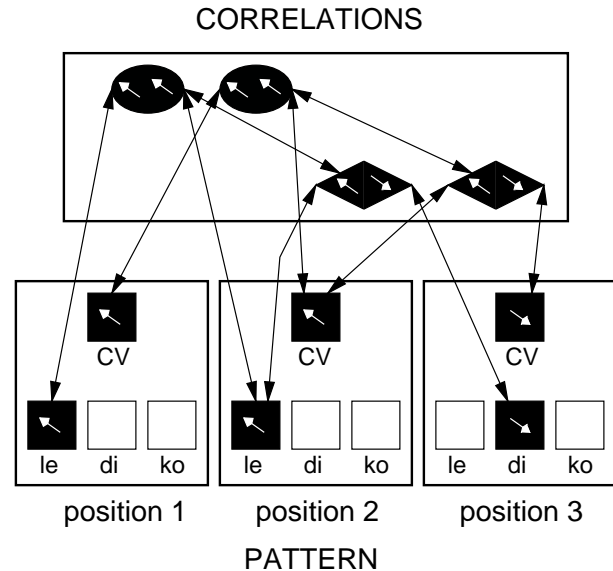


Figure 4: PLAYPEN Network implementing Marcus et al. Only a few units are shown. Connections implementing similarity between PATTERN element units and inhibitory connections between incompatible element units are not shown. The activated (black) units in the PATTERN layer are those that would be active following the presentation of the pattern *le le di*. Four of the relation units that would be activated as a result of this are shown, and ten connections that would be strengthened during the resulting learning. Two of these connections, those joining the units in the CORRELATIONS layer, represent higher-order relational correlations.

units. Because the PATTERN units include very general ones (for example, one that is activated for any CV syllable in second position), the CORRELATIONS layer should be activated relatively highly even by specific syllable sequences it has not been trained on, as long as they are consistent with the training rule.

The average results from 10 networks trained on each grammatical pattern are shown in Figure 5. The total activation of the CORRELATIONS layer was averaged over four trials of each of the test words. The expected interaction between training rule and testing rule is highly significant ( $p < .001$ ). As shown in Figure 5, the CORRELATIONS layer is more activated for novel sequences that follow the grammatical rule the network was trained on than for novel sequences that follow either of the other two rules.

There are several points to note about the way the network learns the tasks.

1. Each unit in the network encodes content information as well as relational information. Thus an activated CORRELATION unit represents at the same time the co-occurrence of particular syllables (or syllable types if it is connected to relatively coarse PATTERN units) and the co-occurrence of syllables bearing a particular similarity relation to one another.
2. Though it cannot perform the segmentation that is a part of Saffran et al.'s task, this network can learn the content-specific correlations in the three-syllable patterns in the task. Since each of the patterns consists of three different syllables, the PATTERN units would take on three differ-

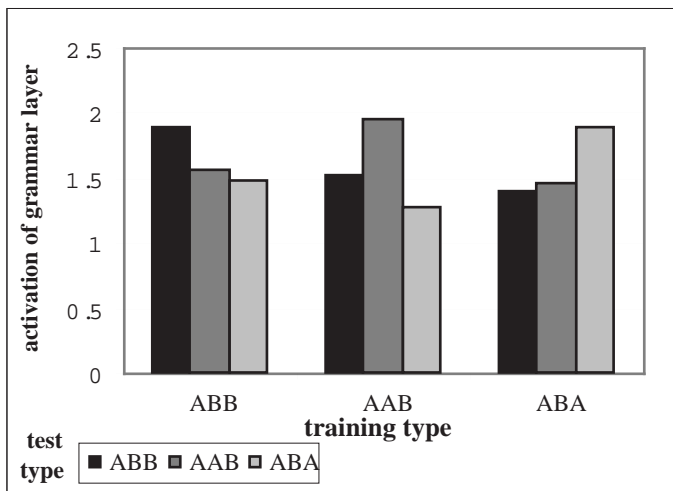


Figure 5: Networks that have been trained on sequences following a certain grammatical pattern respond with more activation to *novel* sequences obeying that same pattern than to novel sequences obeying other patterns.

ent angles for each pattern, activating difference units in the CORRELATIONS layer and resulting in learning on the connections between these units (representing higher-order content-specific regularities).

- While this was not true of Marcus et al.’s task, a set of patterns may embody more than one higher-order relational regularity. For example, in a set of four-element patterns, some patterns might be consistent with the rule AABB and others consistent with the rule ABBA. While we are unaware of experiments testing the ability of subjects to extract such rules, we assume that the ability to learn multiple rules is necessary for language acquisition. A network like that in Figure 4 (but with four positions) could learn both regularities, each as patterns of connections between the six pairwise relational regularities.

## Contrasting Two Accounts of Relational Pattern Learning

### A Rule-Based Account

A number of models have been proposed to handle the results of Marcus et al.’s experiments (Christiansen & Curtin, 1999; Seidenberg & Elman, 1999). Here we contrast only ours and the rule-based account proposed by Marcus (forthcoming). Marcus argues that tasks such as this one, in fact higher cognition and language generally, rely on the learning and manipulation of **explicit rules** containing **abstract variables**, placeholders that apply to any member of a given class.

Having been trained on a pattern learning task of the type in Marcus et al.’s experiments, the learner extracts an explicit rule of the form AAB, where A and B are now abstract variables in Marcus’s sense, and the variables are all associated with some class, say the class of CV syllables (the experiments demonstrate only that infants generalize to other members of this class).

Now consider what patterns will be recognized as familiar after training. Obviously patterns that are identical to those

appearing during training are familiar; if the learner heard the sequence *le le di* during training, that sequence will be recognized later on because it matches the AAB rule. Likewise any pattern consisting of three members of the relevant class for the variables in which the first two elements are identical also matches. So if the relevant class is CV syllables, even if the syllables *ko* and *bi* did not appear during training, the pattern *ko ko bi* will be treated as familiar, apparently just as familiar as *le le di* since all members of the class match the variables equally well. Furthermore, the sequences *le le le* and *ko ko ko* are also familiar since, assuming these variables behave like those in first-order predicate calculus, the rule does not force the third element to be different from the first and second.<sup>1</sup>

Now consider what patterns would fail to be treated as familiar. Since identity is all-or-none, patterns in which the first two elements are only similar, such as *le le di* (where  $\varepsilon$  is the vowel in *bed*) would be treated as unfamiliar. Likewise patterns in which the elements are outside the class over which the variables are defined would not be recognized. Thus, again assuming that CV syllables are the relevant class, *les les dis* would not be seen as familiar.

### The Relational Correlation Account

The relational correlation account that we have presented in this paper differs from the rule-based account in that content still matters. This is because, even when what is learned are relational, rather than content-specific, correlations, the correlations apply only to a certain range of elements. The extent of this range depends on the encoding coarseness of the PATTERN units in question, but given a range of degrees of coarseness, we can expect some relatively content-specific relational correlations to be learned, along with some more general relational correlations.

The implication is that the network’s response will depend on the degree of similarity between the training and test patterns, as well as on whether the training rule is followed. Patterns that are identical to the training patterns should result in the greatest familiarity. Those that are similar should be treated as less familiar. Those that are quite different, as in Marcus et al.’s experiments, should be still more surprising (though still less so than novel patterns that do not follow the rule).

For the network, the notion of the class over which a variable is defined does not exist. Because CVC syllables share some features with CV syllables, we can expect some generalization to CVC patterns that follow the rule, especially if they share segments with the training syllables.

Further, sameness and difference have equal status in the network, so trained on AAB patterns, the network cannot help but learn that the third element is different from the first and second, as it learns that first and the second are the same. This contrasts with the rule-based approach which requires the learning of an extra predicate to encode the distinctness of the third element.

Finally, difficulty of pattern learning should depend on the number of distinct syllables in the word. When a pattern has

<sup>1</sup>Of course, the learner could also extract in addition the explicit constraint that the third element differs from the first and second, but this would seem to be learning “more” than just the rule, so harder or less likely.

three distinct elements, the built-in connections implementing inter-element similarity and difference cause the activated PATTERN units to repel each other's angles, resulting in three different angles. However, depending on the magnitude of the weights connecting the units, there is also an attractor in the network at which there are only two different phase angles. At the same time, relation units can represent only binary relations, and strong associations between relation units can only develop for different relational regularities involving the same two objects (as in Marcus et al.'s experiments). Thus PLAYPEN has a strong preference for *two*, and in a four-syllable version of Marcus et al.'s experiment, we would expect that sequences such as ABCC would be confused with AABB and ABBB. In symbolic models, on the other hand, there is no built-in preference for a particular number of variables.

### Conclusions and Future Work

In this paper we have shown how a connectionist network with a mechanism for grouping together activated units (angles) and a mechanism for representing primitive relational knowledge explicitly (relation units) can learn the task of Marcus et al.'s experiments. While a PLAYPEN network is perhaps not a conventional neural network, we do not believe it has variables hidden in it. But whether it does or not, the key issue should be whether this model makes different predictions from alternate models, specifically from rule-based models. We have argued in the last section that this is the case. Most of these predictions are testable, and we are currently performing an experiment using visual patterns and adult subjects to test the role of similarity to training patterns in the learning of relational regularities. Preliminary results indicate that subjects are more accurate and faster at judging the familiarity of patterns following the training rule when their content is similar to that in the training patterns, as predicted by our model.

Another potential contribution of our model is the placing of "rule" learning in the context of segmentation and grouping. If we are right, then for auditory patterns such as those in the two sets of infant experiments discussed here, the considerable research on rhythm processing (Handel, 1989) is relevant and should lead to a range of predictions. For example, we might expect the relative timing or loudness of the syllables in patterns to play a role in what is learned.

Relations obviously play a fundamental role in human cognition, and we have argued elsewhere that the relational correlation framework embodied in PLAYPEN accommodates relations without sacrificing the distributed representations and simple Hebbian learning that characterize connectionist networks. Indeed the original motivation for PLAYPEN was the learning of spatial relation terms in language rather than the learning of sequences of syllables. We believe the importance of Marcus et al.'s experiments is not to demonstrate that infants can make use of variables but to show that they are good learners of relational correlations, a capacity that will be crucial as they are exposed to language in all its complexity.

### References

Christiansen, M. H. & Curtin, S. L. (1999). The power of statistical learning: no need for algebraic rules. *Proceed-*

*ings of the Annual Conference of the Cognitive Science Society, 21*, 114–119.

- Gasser, M. & Colunga, E. (1998). Where do relations come from?. Tech. rep. 221, Indiana University, Cognitive Science Program, Bloomington, IN.
- Handel, S. (1989). *Listening: An Introduction to the Perception of Auditory Events*. MIT Press, Cambridge, MA.
- Hopfield, J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences, 81*, 3088–3092.
- Hummel, J. E. & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*, 480–517.
- Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283*, 77–80.
- Marcus, G. F. (forthcoming). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT Press, Cambridge, MA.
- Movellan, J. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In Touretzky, D., Elman, J., Sejnowski, T., & Hinton, G. (Eds.), *Proceedings of the 1990 Connectionist Models Summer School*, pp. 10–17. Morgan Kaufmann, San Mateo, CA.
- Mozer, M. C., Zemel, R. S., Behrmann, M., & Williams, C. K. I. (1992). Learning to segment images using dynamic feature binding. *Neural Computation, 4*, 650–665.
- Pinker, S. (1999). Out of the minds of babes. *Science, 283*, 40–41.
- Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by eight-month-old infants. *Science, 274*, 1926–1928.
- Seidenberg, M. S. & Elman, J. L. (1999). Do infants learn grammar with statistics or algebra?. *Science, 284*, 433.
- Shastri, L. & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: a connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences, 16*, 417–494.
- Sporns, O., Gally, J. A., Reeke, G. N., & Edelman, G. M. (1989). Reentrant signaling among simulated neuronal groups leads to coherency in their oscillatory activity. *Proceedings of the National Academy of Sciences, 86*, 7265–7269.

# Resource-adaptive Selection of Strategies in Learning from Worked-Out Examples

Peter Gerjets, Katharina Scheiter & Werner H. Tack

Collaborative Research Center 378: Resource-Adaptive Cognitive Processes

University of the Saarland

D-66041 Saarbruecken/Germany

{pgerjet, katharis, tack}@cops.uni-sb.de

## Abstract

Most tasks can be pursued by using different strategies (Logan, 1985; Reder & Schunn, 1998). In this paper we focus on strategies of learning from worked-out examples. Within a resource-oriented framework these different strategies can be classified according to their costs and benefits. These features may determine which strategy will be selected for accomplishing a task in situations with certain resource limitations. We investigate specific hypotheses about strategic adaptations to resource limitations (e.g., time pressure or lack of prior knowledge) within a hypertext-based learning environment. A comparison of the strategy selection of good and poor learners is used to assess the degree of subjects' resource adaptivity. Ideas for modeling resource-adaptive selection of strategies within the ACT-R architecture are discussed.

## Resource-Adaptive Selection of Strategies

According to Reder and Schunn (1998) individual performance differences in learning and problem-solving tasks may not only depend on the *variability of cognitive parameters* (e.g. speed of processing, working-memory capacity) or on interindividual differences in the *availability of strategies* for solving the same task. Instead subjects may differ with regard to their *ability of shifting strategies* as a consequence of changes in task demands or other situational parameters. Therefore, the *adaptive selection of strategies* should be of major importance for success in learning and problem solving. Theoretically, *associative approaches* explain strategy selection as a reaction to cues related to certain strategies (cf. Reder & Schunn, 1998). On the contrary, *rational approaches* assume that subjects choose strategies according to their costs and benefits in terms of resource demands and expected utility (cf. Payne, Bettman & Johnson, 1993; Logan, 1985). In our paper we prefer a rational approach which is based on a wide conception of resources comprising all internal and external means that are useful or necessary for solving a specific task. We focus on internal resources like prior knowledge and external resources like learning time and external information. The *costs* of adopting a specific strategy increase with its resource demands. Besides differences in costs, strategies may additionally differ with respect to their *benefits* (e.g., effectiveness in solving the task at hand, success in solving subsequent tasks, acquisition of different kinds of knowledge).

To describe processes of strategy selection within a resource-based framework *two different types of resource adaptivity* have to be distinguished: (a) On the one hand evolution may have forced cognitive systems to generally employ *resource-adapted strategies*, i.e. strategies that do not lead to optimal task performance but that are compatible with the usual limitations of processing resources. According to this assumption resource-adapted behavior will be even displayed in situations with relatively high resource availability. A well known theoretical approach to resource-

adapted behavior that may be applied to strategy selection is the concept of *satisficing* (Simon, 1990). According to this concept bounded rational agents do not select the most effective strategy for solving a task but rather set a specific aspiration level (probably associated with the value of the respective goal) and select a strategy that exceeds it. (b) On the other hand cognitive systems may be *resource-adapting* in that the strategies employed to pursue a certain task are additionally constrained by the configuration of resources currently available for the agent. If resources like knowledge, time or external information are restricted, strategies should be adopted that are less demanding with respect to these resources. These strategic shifts may compensate for performance impairments expectable without such adaptations. Severe limitations of specific resources or certain combinations of resource limitations may prove impossible to compensate. Taken together it can be postulated that subjects generally choose satisficing strategies even when no strong resource limitations are present. Additionally they should adapt to specific limitations by choosing strategies that are more frugal with respect to these limited resources. Therefore, our main aim is to determine whether subjects working under certain resource limitations employ the same strategies as subjects without such limitations or whether they adapt to these limitations in a useful way.

## Learning from Worked-Out Examples

In our empirical work we are specifically interested in strategies of learning from worked-out examples from the domain of probability word problems. Worked-out examples are instances of a certain problem type together with a detailed solution. They facilitate the learning of abstract procedures for later problem solving (Cummins, 1992; Catrambone, 1998) and the solving of novel problems by analogy (Gick & Holyoak, 1983; Reed, 1999). Prerequisites for the use of examples for knowledge acquisition and application are the generation of suitable example representations and the initiation of appropriate cognitive processes working on these representations. For our purposes strategies of learning from worked-out examples can be described on two dimensions: *rare versus frequent use* of examples and *brief versus extensive use* of examples. Van Lehn and Jones (1993) found that better learners preferred a rare use of examples and tried to solve training problems on their own. While good learners only inspected examples for getting specific information, poor learners referred back to examples as often as possible.

Beyond differences in the frequency of example use learners can use examples more or less extensively depending on the degree of example elaboration during learning. These elaborations may comprise the abstract deep structure of an example problem, the subgoal structure of its solution (Ca-

trambone, 1998), or the similarities and differences of the example compared to other examples from the same problem type (Cummins, 1992). Chi, Bassok, Lewis, Reimann & Glaser (1989) found that learners who elaborated examples during study substantially differed in their performance from learners that didn't elaborate on example problems. According to Chandler and Sweller (1991, p. 294) these results "indicate the importance to learning of an ability to properly process worked examples". Therefore, strategy selection in learning from worked-out examples may have a major influence on the quality of knowledge acquired.

## Hypotheses

Based on the underlying theoretical framework we derived five experimental hypotheses about resource-limitations in learning from worked-out examples and about adaptive strategy selection: (a) Learning may be impaired if relevant resources like learning time, prior knowledge or external information are severely limited. (b) Different resource limitations may not act additively but interact with each other if more than one resource is limited. Therefore, different combinations of resource-limitations may result in different patterns of performance impairments. (c) Strategy shifts may help compensate for performance impairments associated with certain resource configurations, though some combinations of resource limitations may prove impossible to compensate by strategic choices. Hence, good and poor learners may differ in strategic variables under some but not under all resource configurations. (d) It can be expected that subjects select faster but less accurate processing strategies if learning time is limited. In the case of limited prior knowledge or external information, subjects may select less information-demanding strategies even if these strategies involve time-consuming inferences. (e) The dimensions rare - frequent use of examples and brief - extensive use of examples should be useful to characterize strategies for learning from worked-out examples and to describe relevant strategy shifts.

To investigate these hypotheses we conducted a series of three experiments in which subjects' had to work on a learning and problem-solving task from the domain of probability word problems. We developed a hypertext system to serve as experimental environment that allowed us to log subjects' strategic decisions in great detail. With regard to our hypotheses a question of central importance is whether possible strategic differences between experimental conditions can be interpreted as adaptive. To answer this question we employed the method of contrasting *strategic differences between experimental conditions* with *strategic differences between good and poor learners* within experimental conditions. This approach allows us to decide whether subjects who learn under a specific configuration of resource limitations change their behavior in a direction that can be identified as useful given this configuration of resources.

## Experiment 1

### Method

**Participants** The subjects were 46 students of the University of the Saarland (UdS), Germany who either participated for course credit or payment. Average age was 24.5 years.

**Materials and procedure** In the hypertext environment a short introduction to the domain of combinatorics was presented and subjects were instructed to solve a number of *probability word problems* following a *self-paced learning phase*. In the learning phase of the experiment subjects could retrieve *abstract explanations* of six solution principles from the domain of combinatorics (with their associated formula) by clicking on the respective links in the navigation bar. In the *test phase* the instructional information of the learning phase was no longer available. Three test problems were presented on the screen and one of the test problems had to be selected to begin with. In this experiment *no worked-out examples* were included because in the first step we wanted to study performance and strategy selection in learning with purely abstract information.

**Design and dependent measures** As independent variables time pressure and prior knowledge were manipulated by implementing three different learning conditions. In the *baseline condition with high resource availability* subjects possessed relatively high domain-specific prior knowledge and were instructed to take as much time as needed to understand the solution principles and then to begin with the test phase by clicking on the respective link. In the *condition with low prior knowledge* learning time was likewise unlimited, but subjects were rather unfamiliar with the domain of combinatorics. In the *condition with low learning time* we restricted the learning time of subjects with high prior knowledge to seven minutes (i.e., about two thirds of the mean learning duration in the condition without limitations). To induce time pressure subjects were informed that they would only be granted two thirds of the time usually needed for the learning phase. When the learning time (visible for the subjects on a digital clock) expired, the first page of the problem solving phase was automatically presented on the screen and subjects were instructed to begin working on the test problems. During problem solving there were no time limits. In the test phase the subjects had to mark the appropriate solution principle and the values of two variables for each of the three test problems in a multiple-choice form available in the hypertext environment. No calculations had to be made. One error was assigned for each wrong answer. Problem-solving time as well as total learning time, mean reading time per abstract page presented and frequency of retrieving abstract information pages were recorded by using logfiles. Following the test phase subjects had to pass a knowledge test with ten multiple-choice questions related to abstract concepts from the domain of combinatorics. One error was assigned for each wrong answer. Similar conceptual questions were posed as a pretest at the beginning of the experiment to control for domain-specific prior knowledge. Additionally, we registered subjects' last math grade as a general measure of mathematical ability which ranged from grade one (best) to grade six (worst).

## Results and Discussion

First we investigated whether the three learning conditions differ with regard to performance and strategy measures. For this reason, we used the *baseline condition with high resource availability* as a point of reference and contrasted its data with the two other conditions (see table 1).

Table 1: Means and significance of differences

Learning without worked-out examples	A: Low prior knowledge	B: Base-line	C: Low learn. time	Significance of Difference
Problem-solving errors	52.3 %	32.5 %	42.1 %	A >> B << C
Knowledge-test errors	35.4 %	8.9 %	17.1 %	A >> B = C
Math grade	2.2	2.1	1.7	A = B = C
Pretest errors	65.8 %	32.2 %	30.2 %	A >> B = C
Frequency / abst. info.	22	24	15	A = B > C
Mean time / abst. info.	84 sec.	69 sec.	31 sec.	A = B >> C
Total learning time	823 sec.	722 sec.	397 sec.	A = B >> C
Problem solving time	782 sec.	726 sec.	759 sec.	A = B = C

Note: >>:  $p \leq .05$ ; >:  $p \leq .10$ ; =:  $p > .10$  (p-values result from one-tailed t-tests)

A comparison with the *low-prior-knowledge condition* (A versus B) reveals strong differences in problem-solving errors and knowledge-test errors while there are no differences with regard to strategic measures. This may imply that subjects with low prior knowledge don't try to compensate for their performance impairments by increasing problem-solving time or learning time if only abstract information about the solution principles is available in the learning environment. Comparing the baseline condition with the *low-learning-time condition* (B versus C) yields similar differences in problem-solving errors while there are no differences with regard to knowledge-test errors. In addition, both conditions differ with respect to strategic measures. Compared to subjects in the baseline condition subjects under time pressure retrieve abstract information pages less frequently and spend less time on each abstract information page. This change in strategic behavior is not obligatory as subjects could as well have reacted to time pressure by only reducing the mean time reading abstract information but not the retrieval frequency (as they do in experiment 2).

In a second step we evaluated the adaptivity of strategy shifts in experiment 1 by comparing good and poor learners within the experimental conditions with regard to the strategy measures listed in table 1 (post-hoc median splits according to problem-solving performance). To rule out the hypothesis that differences between good and poor learners are caused by differences in prior knowledge or math grade we inserted these variables as covariates in the statistical comparison of good and poor learners. The respective analyses of covariance reveal that there are no differences with regard to strategy measures distinguishing between good and poor learners. This implies that subjects' strategic options (modifying the frequency or intensity of processing abstract information) are unsuitable for improving problem-solving in the conditions with purely abstract information. Accordingly, efficiency impairments caused by restrictions in either prior knowledge or learning time cannot be easily compensated by strategic shifts in this experiment. Therefore, no resource-adaptive processes of strategy selection could be evidenced here. We conducted experiment 2 to investigate whether strategies of information processing are of greater importance in example-based learning.

## Experiment 2

### Method

**Participants and materials** The subjects were 46 students of the UdS who either participated for course credit or pay-

ment. Average age was 24.5 years. In experiment 2 the hypertext environment was supplemented by a single worked-out example per solution principle.

**Design and dependent measures** The same three learning conditions as in experiment 1 were implemented. Time pressure was induced analogously to experiment 1 by restricting learning time to nine minutes. The learning environment was augmented by a single worked-out example for each solution principle. These examples as well as the abstract information of the learning phase were no longer available in the test phase. Dependent measures were problem-solving errors, knowledge-test errors, domain-specific prior knowledge, last math grade, mean reading time per example provided, frequency of example retrieval (number of clicks), mean reading time per abstract information page, frequency of abstract information retrieval (number of clicks), total learning time, and problem-solving time.

### Results and Discussion

Compared to the baseline condition with high resource availability subjects in the *low-prior-knowledge condition* show substantial performance impairments in problem solving and in the knowledge test (see A versus B in table 2). Furthermore, there are significant differences with regard to strategic measures between the two experimental groups. Subjects with low prior knowledge spend more time on learning and especially show an increased frequency of retrieving abstract information as well as an increased mean time reading these pages. There are, however, no differences concerning the use of examples between the two groups. Comparing the baseline condition with the *low-learning-time condition* (B versus C) yields similar differences in problem-solving errors and knowledge-test errors. With respect to strategic measures, subjects in the low-learning-time condition retrieve examples less frequently and spend less time reading examples and abstract information. Interestingly, subjects under time pressure retrieve abstract information more often than baseline subjects.

Table 2: Means and significance of differences

Learning with one worked-out example	A: Low prior knowledge	B: Base-line	C: Low learn. time	Significance of Difference
Problem-solving errors	55.4 %	32.0 %	54.5 %	A >> B << C
Knowledge-test errors	36.8 %	13.1 %	20.9 %	A >> B < C
Math grade	2.6	1.9	2.0	A > B = C
Pretest errors	66.1 %	33.3 %	32.3 %	A >> B = C
Frequency / example	8	7	3	A = B >> C
Mean time / example	56 sec.	44 sec.	14 sec.	A = B >> C
Frequency / abst. info.	22	13	16	A >> B < C
Mean time / abst. info.	62 sec.	45 sec.	37 sec.	A >> B > C
Total learning time	1047 sec.	809 sec.	516 sec.	A > B >> C
Problem solving time	600 sec.	606 sec.	571 sec.	A = B = C

Note: >>:  $p \leq .05$ ; >:  $p \leq .10$ ; =:  $p > .10$  (p-values result from one-tailed t-tests)

A comparison of good and poor learners within the experimental conditions reveals the following strategic differences: In the *baseline condition* good learners spend more time reading examples than poor learners. In the *low-prior-knowledge condition* there are no strategic differences between good and poor learners. This implies that the performance in this condition may not easily be improved by

strategic shifts. Nevertheless, subjects with low prior knowledge try to improve their performance by learning longer (increased frequency and time reading abstract information). This shift, however, only increases costs in terms of time investment but doesn't yield any benefits in terms of performance. Therefore, subjects in this condition don't behave resource-adaptive.

In the *low-learning-time condition* good learners invest more time reading examples than do poor learners. In the light of this finding, it can be recommended that subjects under time pressure should save time by reducing time for abstract information processing without simultaneously confining the processing of examples. As the data in table 2 reveal, subjects under time pressure do not follow this recommendation towards resource-adaptive behavior. They only show a slight reduction in the mean reading time per abstract information page while there is a substantial decrease in the mean time reading examples. The respective interaction is significant and indicates that no resource-adaptive strategy shift took place. To conclude, performance impairments due to lacking prior knowledge cannot be compensated by selecting different strategies. Therefore, subjects' attempts to improve performance are in vain. On the other hand, performance impairments due to time pressure may be compensated by focussing on example information. Unfortunately, subjects do not shift their strategies in this direction. Therefore, no resource-adaptive strategy selection could be found in experiment 2.

We finally compared all six conditions from experiment 1 and 2. Contrasting the two conditions with low prior knowledge doesn't reveal any decrease in problem-solving errors due to the provision of examples. However, subjects in the one-example condition need less time for problem solving which indicates a slight increase in overall efficiency. A similar pattern of results can be found for the two baseline conditions. Unexpectedly, subjects in the low-learning-time condition deteriorate significantly with regard to problem-solving errors when provided with one example. Their problem-solving time is decreased analogously to the two other resource conditions. The respective interaction between time pressure (with/ without) and example availability (with/ without) with regard to problem-solving errors is significant.

To sum up, in our experimental setting learning with examples doesn't seem to be more effective than learning with only abstract information. At least the mere provision of instructional examples is obviously not sufficient to improve learning. Rather, the availability of examples must be accompanied by an extensive example-processing. As the differences between good and poor learners in the baseline condition and in the low-learning-time condition reveal this is crucial to performance. Furthermore, we found first support for the assumption that different kinds of resources may interact with regard to their effects on learning and problem solving. The augmentation of abstract information with one worked-out example slightly improves problem solving (i.e., reduces problem-solving time) if prior knowledge is restricted while it can have detrimental effects on problem-solving errors in the case of time limitation. In order to test whether these effects can also be observed when providing more than one example we conducted a third experiment.

## Experiment 3

### Method

**Participants and materials** The subjects were 43 students of the UdS who either participated for course credit or payment. Average age was 24.7 years. In experiment 3 the hypertext environment was supplemented by three worked-out examples of varying complexity to illustrate the application of each solution principle to different problem situations.

**Design and dependent measures** The same three conditions as in experiment 1 and 2 were used in this experiment. Time pressure was induced by allowing 13 minutes for learning in the time-limited condition. Dependent measures were the same as in experiment 2.

### Results and Discussion

Compared to the baseline condition with high resource availability subjects in the *low-prior-knowledge condition* again show an increase in both types of error rates (see table 3, A versus B). With regard to strategic measures, subjects with low prior knowledge spend less time reading examples but simultaneously show an increase in time reading abstract information. Their time for problem solving is slightly decreased. Surprisingly, the comparison between the baseline condition and the *low-learning-time condition* (B versus C) shows that time pressure does not lead to impairments in problem solving like it did in experiment 1 and 2. There are, however, differences in knowledge-test errors as expected. Concerning strategic measures, subjects under time pressure spend less mean time reading examples and retrieve examples less frequently.

Table 3: Means and significance of differences

Learning with three worked-out examples	A: Low prior knowledge	B: Baseline	C: Low learn. time	Significance of Difference
Problem-solving errors	50.3 %	32.5 %	28.9 %	A >> B = C
Knowledge-test errors	33.7 %	13.6 %	22.0 %	A >> B < C
Math grade	2.8	2.0	2.5	A >> B = C
Pretest errors	59.6 %	29.4 %	35.6 %	A >> B = C
Frequency / example	13	17	7	A = B >> C
Mean time / example	25 sec.	32 sec.	9 sec.	A < B >> C
Frequency / abst. info.	28	23	21	A = B = C
Mean time / abst. info.	71 sec.	49 sec.	54 sec.	A > B = C
Total learning time	1179 sec.	1153 sec.	751 sec.	A = B >> C
Problem solving time	522 sec.	640 sec.	753 sec.	A < B = C

Note: >>:  $p \leq .05$ ; >:  $p \leq .10$ ; =:  $p > .10$  (p-values result from one-tailed t-tests)

Comparing good and poor learners within the *baseline condition* shows that good learners spend more time on learning (especially on abstract information pages) and more time on problem solving. In the *low-prior-knowledge condition* good learners' frequency of retrieving examples and of retrieving abstract information is increased as well as their mean time reading example pages. Hence, it would be resource-adaptive in this condition to study abstract information and example information more intensively and in a well-balanced way. However, subjects with low prior knowledge even show a reduced mean time reading examples compared to subjects in the baseline condition. Furthermore, there is a significant cross-interaction between prior-knowledge (with/ without) and retrieval frequency of



different instructional material (examples/ abstract information). This interaction shows that low-prior-knowledge subjects focus on the retrieval of abstract information instead of handling examples and abstract information in a well-balanced way. In the *low-learning-time condition* good learners show an increased frequency of retrieving abstract information and examples. Thus a useful recommendation to subjects working under time constraint could be to retrieve example information and abstract information in a well-balanced way. A significant cross-interaction between time pressure (with/ without) and retrieval frequency of different instructional material (examples/ abstract information) reveals that subjects under time pressure focus on the retrieval of abstract information instead of handling examples and abstract information in a well-balanced way. Their behavior can thus not be classified as resource-adaptive. However, this is the only condition in which time pressure does not lead to significant performance impairments. This unexpected finding can be explained by considering that subjects more or less ignored the examples provided and therefore could spend the same amount of time in processing abstract information as subjects without time pressure and without instructional examples (i.e., baseline condition in experiment 1). Accordingly, their performance is comparable to that condition.

Contrasting the results from experiment 2 and 3 reveals that subjects with three examples learning in the baseline condition and in the low-prior-knowledge condition do not perform any better than the respective subjects in the one-example conditions. As explained before, improvements under time pressure are presumably not attributable to the provision of three examples but rather to the fact that subjects ignore the examples to save time for processing abstract information. The augmentation of instructional resources to three examples therefore does not prove as beneficial as could be expected when considering theories of learning by analogy (Gick & Holyoak, 1983) or theories of learning from worked-out examples (Cummins, 1992; Quilici & Mayer, 1996). At least the *mere provision* of three examples is obviously not sufficient to improve learning. Rather, the provision of multiple examples must be accompanied by a balanced processing of example information and of abstract information in order to acquire the relevant knowledge for problem solving. As the differences between good and poor learners in each of the three-example conditions reveal this is crucial to performance. Contrary to subjects learning with one example who profit most from studying the example intensively subjects learning with three examples should equally focus on abstract information. This finding fits theoretical assumptions about schema abstraction and the acquisition of transferable knowledge according to which it is necessary to compare different examples with respect to relevant abstract properties to induce theoretical concepts that may be applicable to analogous problems (Cummins, 1992).

## General Discussion

Contrary to our *first hypothesis (a)* we found that limitations of relevant resources are not always associated with performance impairments and accordingly that the provision of

relevant resources is not always associated with performance improvements. E.g., the provision of additional instructional information doesn't always improve problem-solving. It can even lead to impairments if subjects are overwhelmed by information selection and integration. This interpretation is in line with the fact that subjects with low learning time suffer from the provision of one example and that they resign from the processing of examples when provided with three examples. Furthermore, as postulated in our *second hypothesis (b)* effects of resource limitations are not always additive, but may even be cross-interacting. For example, the augmentation of instructional resources by worked-out examples is slightly beneficial for subjects with low prior knowledge (decreased problem-solving time), while it can even have harmful effects for subjects with low learning time (increased problem-solving errors). Contrary to our *third hypothesis (c)* no cases of resource-adaptive strategy shifts could be identified. There are no patterns of differences between experimental conditions that can be classified as adaptive with respect to differences between good and poor learners within these experimental conditions. Our *fourth hypothesis (d)* stating that subjects with limited learning time should select faster but less accurate example processing strategies was confirmed in experiment 2 and 3. Contrary to our expectations, subjects with low prior knowledge do not adopt more time-consuming strategies of example processing. Finally, as predicted in the *fifth hypothesis (e)*, the dimensions brief versus extensive use (time per example provided) and rare versus frequent use (frequency of example retrieval) are important dimensions for describing strategies of learning from worked-out examples. This can be inferred from the differences between good and poor learners and between experimental conditions with respect to these variables.

In conclusion, our experiments show that strategic options to improve one's learning performance become the more numerous the more instructional material is provided. At the same time it could be demonstrated that one has to make use of these strategic options, i.e., adopt adequate strategies in order to benefit from this additional information.

## Cognitive Modeling Approach

In the next step we intend to develop a more detailed model of resource limitations and their influences on processes of strategy selection. Within a cognitive science framework high-level processes of executive control like strategy selection in learning and problem-solving may be best modeled by means of cognitive architectures that are designed as comprehensive theories of human cognitive abilities. As a theoretical basis for the cognitive modeling of strategy selection in learning from worked-out examples we will refer to the ACT-R architecture (Anderson & Lebiere, 1998) that is based on a rational analysis approach compatible with our framework of resource-adaptive strategy selection. If one defines strategies for performing tasks as sets of procedures or operations that may be adopted in order to implement a certain goal, strategies can be easily represented in ACT-R by sets of productions that are sufficient to solve a task successfully. Based on this representation, two mechanisms of

action control can be distinguished in ACT-R that are useful in modeling strategy selection.

On the one hand, processing in ACT-R is controlled by the *currently active goal*. Productions referring to other than the current goal cannot be selected for execution. Strategy selection by setting strategy-specific subgoals can be interpreted as a choice process that is based on discrete symbolic knowledge and may be useful to model more deliberate aspects of strategy selection. Accordingly, goal setting implies that the accomplishment of the current task is interrupted for a period of meta-level decision making.

On the other hand, control in ACT-R is determined by the *mechanism of conflict resolution* that selects one of the conflicting productions that are compatible with the current goal for the next processing step. Strategy selection based on conflict resolution may be described as a subsymbolic process embedded within the fundamental mechanisms of the architecture. Conflict resolution is assumed to be an automatic process that is not consciously accessible and accordingly is initiated without changes in the current goal of information processing. ACT-R's mechanism for conflict resolution is based on an estimation of the expected gain  $E$  of the conflicting productions. For every feasible production  $i$  the value of  $E$  is determined by the formula  $E = P G - C$  with  $P$  being the expected probability of goal achievement when using  $i$ ,  $G$  being the goal value, and  $C$  being the expected costs of goal achievement when using  $i$ . Within this framework the resource limitations studied in our experiments can be modeled as follows.

**Time pressure** In ACT-R the goal value  $G$  is operationalized by the maximum amount of time that may be invested for goal achievement. Costs of goal achievement  $C$  are likewise measured by the time needed for goal accomplishment. Based on these conventions the mechanism of conflict resolution inherently produces a speed-accuracy trade-off depending on the available time. Time pressure will result in a decrease of  $G$ , leading to a lower weight of success probability and a higher weight of processing costs in production selection. As a result, less effective but at the same time less costly strategies will be selected for task accomplishment. Thus ACT-R enables the modeling of subjects' adaptation to limitations in time resources by automatic strategy shifts.

**Lack of prior knowledge** Limitations in domain-specific knowledge can be represented as gaps in declarative knowledge, i.e., the appropriate conceptual apparatus to encode the instructional material. In ACT-R these limitations can be best represented by missing chunk-types (representing concepts). Thus gaps in prior knowledge cannot be compensated automatically. Rather a deliberative setting of a specific learning goal may be necessary to first initiate activities to acquire the required conceptual knowledge.

**Limited external information** If external information necessary for the execution of the production with the highest expected gain in the conflict set is lacking this production is automatically abandoned in ACT-R and a production with less expected gain is selected that matches the currently available information. Thus a task can be handled successfully as long as there are productions available whose information demands are satisfied by the current external in-

formation. Augmenting external information beyond these minimal requirements will improve performance if this information can be encoded correctly and if there are productions available that properly use this information within the time available. To model the interaction between time limitations and example availability with regard to problem-solving errors we assume that subjects under time pressure may lack the necessary time to process example information properly. This may explain why the provision of one worked-out example is harmful for low-learning-time subjects while subjects with sufficient learning time don't show any efficiency impairments.

## Acknowledgements

We thank Simon Albers, Tina Schorr, Julia Schuh, Markus Werkle and the German Research Foundation for their support.

## References

- Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Erlbaum.
- Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General*, 127, 355-376.
- Chandler, P. & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8, 293-332.
- Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P. & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145-182.
- Cummins, D. D. (1992). Role of analogical reasoning in the induction of problem categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 1103-1124.
- Gick, M. L. & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15, 1-38.
- Logan, G. D. (1985). Executive control of thought and action. *Acta Psychologica*, 60, 193-210.
- Payne, J., Bettman, J. R. & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge: Cambridge University Press.
- Quilici, J. L. & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, 88, 144-161.
- Reder, L. M. & Schunn, D. (1998). Bringing together the psychometric and strategy worlds: Predicting adaptivity in a dynamic task. In D. Gopher & A. Koriat (Eds), *Cognitive regulation of performance (Attention and performance XVII)*. Cambridge, MA: MIT Press.
- Reed, S. K. (1999). *Word problems*. Mahwah, NJ: Erlbaum.
- Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, 41, 1-19.
- VanLehn, K. & Jones, R. M. (1993). Better learners use analogical problem solving sparingly. In P. E. Utgoff (Ed.), *Machine Learning: Proceedings of the Tenth Annual Conference*. San Mateo, CA: Morgan Kaufmann.

# A Neural Network Model of Concept-influenced Segmentation

Robert L. Goldstone (rgoldsto@indiana.edu)  
Department of Psychology, Indiana University  
Bloomington, IN 47405 USA

## Abstract

Several models of categorization assume that fixed perceptual representations are combined together to determine categorizations. This research explores the possibility that categorization experience alters, rather than simply uses, descriptions of objects. Based on results from human experiments, a model is presented in which a competitive learning network is first given categorization training, and then is given a subsequent segmentation task, using the same network weights. Category learning establishes detectors for stimulus parts that are diagnostic, and these detectors, once established, bias the interpretation of subsequent objects to be segmented.

## Concept Learning and Perception

The current research explores the influence that learning a new concept has on the segmentation of objects into component parts. Recently a number of researchers have argued that in many situations, concept learning influences the featural descriptions used to describe a set of objects. Rather than viewing perceptual descriptions as fixed by low-level sensory processes, this view maintains that perceptual descriptions are dependent on the higher-level processes that use the descriptions (Goldstone, Steyvers, Spencer-Smith, & Kersten, 2000; Schyns, Goldstone, & Thibaut, 1998). Evidence for this view comes from the study of expert/novice differences (Lesgold et al., 1988), influences of acquired concepts on the interpretation of stimuli (Wisniewski & Medin, 1994), and influences of category learning on psychophysical measurements of perceptual sensitivity (Goldstone, 1994).

## Experiential Influences on Object Segmentation

One type of influence of concept learning on perceptual learning may be to alter how objects are segmented into parts. Objects often have more than one possible segmentation. The letter "X" can be viewed as comprised of two crossing diagonal lines, or as a "V" and an upside-down "V" that barely touch at their vertices. The segmentation of scenes into parts depends upon experience. Behrmann, Zemel, and Mozer (1998) found that judgments about whether two parts had the same number of humps were faster when the two parts belonged to the same object rather than different objects. Further work has found an influence of experience on subsequent part comparisons. Two stimulus components are interpreted as belonging to the same object if they have co-occurred many times (Zemel, Behrmann, Mozer, & Bavelier, 1999). Thus,

experience with particular feature combinations determines whether or not features will be integrated into a single object.

Pevtzow and Goldstone (1994; reported in Goldstone et al., 2000) explored the influence of category learning on segmentation with the materials shown in Figure 1. We pursued the idea that how psychologically natural a part is might depend on whether it has been useful for previous categorizations. Naturalness was measured by how quickly subjects could confirm that the part was contained within a whole object (Palmer, 1978). To test this conjecture, we gave participants a categorization task, followed by part/whole judgments. During categorization, participants were shown distortions of the objects A, B, C, and D shown in Figure 1. The objects were distorted by adding a random line segment that connected to the five segments already present. Subjects were given extended training with either a vertical or horizontal categorization rule. For participants who learned that A and C were in one category, and B and D were in another (a vertical categorization rule) the two component parts at the bottom of Figure 1 were diagnostic. For participants who learned that A and B belonged in one category, and C and D belonged to the other category (a horizontal rule), the components on the right were diagnostic.

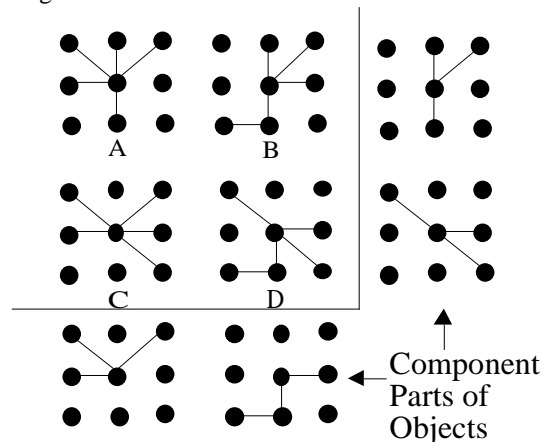


Figure 1: Pevtzow and Goldstone (1994) stimuli

During part/whole judgments, participants were shown a whole, and then a part, and were asked whether the part was contained in the whole. Participants were given both present and absent judgments, and examples of these judgments are shown in Figure 2. Note that the two parts shown in Figure 2 were both potentially diagnostic during

the earlier categorization training. Whether or not a part was diagnostic was independent of the appearance of the part itself, depending only on how the four objects of Figure 1 were grouped into categories.

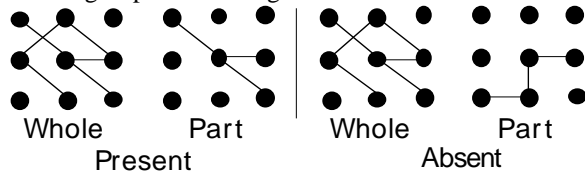


Figure 2: Part/whole present and absent judgments

The major result was that subjects were faster to correctly respond "present" when the part was diagnostic than when it was non-diagnostic. To the extent that one can find response time analogs of signal detection theory sensitivity and bias, this effect seems to be a sensitivity difference rather than a bias difference, because absent judgments also tended to be faster for diagnostic than nondiagnostic parts. Given that a category part that was diagnostic for the horizontal categorization group was nondiagnostic for the vertical group, these results indicate that it is not simply the physical stimulus properties that determine how readily a person can segment an object into a particular set of components; segmentation is also influenced by the learned categorical diagnosticity of the components

### Modeling Interactions Between Concept Learning and Segmentation

We model the result from these experiment using a modified competitive learning network (Rumelhart & Zipser, 1985). As with the experiment, the network is first given categorization training, and then is given a subsequent segmentation task, using the same network weights. The goal of the modeling is to show how categorization training can prime the segmentation network such that objects will tend to be segmented into parts that were previously diagnostic for categorization.

#### The Categorization Network

The categorization network has three layers of units: one representing the input patterns, one representing a bank of learned detectors, and one reflecting the category assignments of the inputs. Both the weights from the input patterns to the detectors and the weights from the detectors to categories are learned. The categorization task uses a modified unsupervised competitive learning algorithm (Rumelhart & Zipser, 1985), but includes a top-down influence of category labels that incorporates supervised learning. The network begins with random weights from a two-dimensional input array to a set of detector units, and from the detectors to the category units. When an input pattern is presented, the unit with the weight vector that is closest to the input pattern is the "winner," and will selectively adjust its weights to become even more specialized toward the input. By this mechanism, the originally homogenous detectors will become differentiated

over time, splitting the input patterns into categories represented by the detectors. The competitive learning algorithm automatically learns to group input patterns into the clusters that the patterns naturally form. However, given that we want the detectors to reflect the experiment-supplied categories, we need to modify the standard unsupervised algorithm. This is done by including a mechanism such that detectors that are useful for categorizing an input pattern become more likely to win the competition to learn the pattern. The usefulness of a detector is assumed to be directly proportional to the weight from the detector to the presented category which is provided as a label associated with an input pattern. The input-to-detector weights do not have to be set before the weights from detectors to categories are learned.

In addition to modifying the unsupervised development of hidden-layer detectors by considering their usefulness for categorization, a second modification of the standard competitive learning algorithm is required to fix one of its general problems in optimally making use of all detectors to cover a set of input patterns. This problem is that if multiple input patterns are presented that are fairly similar to each other, there will be a tendency for one detector to be the winner for all of the patterns. As a result, the winning detector's weight vector will eventually become similar to the average of the input patterns' activations, while the rest of the detectors do not learn at all. This situation is suboptimal because the input patterns are not covered as well as they would be if the unchanging detectors learned something. The standard solution to this problem is called "leaky learning" and involves adjusting both winning and losing detectors, but adjusting losing detectors at a slower rate (Rumelhart & Zipser, 1985). To understand the more subtle problem with this solution, imagine, for example, that four input patterns naturally fall into two groups based on their similarities, and the network is given four detectors. Ideally, each of the detectors would become specialized for one of the input patterns. However, under leaky learning, one detector will tend to become specialized for one cluster, a second will become specialized for the other cluster, and the remaining two detectors will be pulled equally by both clusters, becoming specialized for neither. Note that it does not help to supplement leaky learning by the rule that the closer a detector is to an input pattern, the higher its learning rate should be. There is no guarantee that the two "losing" units will evenly split such that each is closer to a different cluster.

Other researchers have noted related but not identical problems with competitive learning and have suggested solutions (Grossberg, 1987). Our current solution is to conceptualize competitive learning as not simply a competition among detectors to accommodate a presented input pattern, but also as a competition among input patterns to be accommodated by a given detector. Input patterns are presented sequentially to the network, and as they are presented, the closest input pattern to each detector is determined. The learning rate for a detector is set to a higher value for its closest input pattern than for other

inputs. In this manner, detectors that are not the winning detector for a pattern can still become specialized by becoming unequally influenced by different patterns. In addition, the learning rate for a detector when presented with an input pattern will depend upon how well the input is currently covered by existing detectors. This dependency is required to allocate detectors to input regions where they are required. Putting these considerations together, the activation of detector  $i$  when presented with pattern  $p$ , is

$$A_{i,p} = \sum_{h=1}^n I_{h,p} W_{i,h} + \sum_{j=1}^c STW_{j,i}$$

where  $I_{h,p}$  is the activation of input unit  $h$  for pattern  $p$ ,  $W_{i,h}$  is the weight from input  $h$  to detector  $i$ ,  $S$  is the strength of the top-down pressure on detector development,  $T$  is the teacher signal (if Pattern  $p$  belongs to Category  $j$  then  $T=1$ , otherwise  $T=-1$ ), and  $W_{j,i}$  is the weight from Detector  $i$  to Category Unit  $j$ . The second term increases the activation of a detector to the extent that it is useful for predicting the input pattern's categorization. The detector activation will determine which detector is the "winner" for an input pattern. As such, detectors that are useful for categorization will tend to become winners, thus increasing their learning rate.

Input-to-detector weights are learned via top-down biased competitive learning using the following equation for changing weights from input pattern  $h$  to Detector  $i$ :

$$\Delta W_{i,h} = \begin{cases} M(I_{h,p} - W_{i,h}) & \text{if } \forall x(A_{i,p} \geq A_{x,p}) \\ N(I_{h,p} - W_{i,h})K_p & \text{if } \forall y(A_{i,p} \geq A_{i,y}) \\ O(I_{h,p} - W_{i,h})K_p & \text{otherwise} \end{cases} \text{otherwise}$$

where  $M$ ,  $N$ , and  $O$  are learning rates ( $M > N > O$ ), and  $K_p$  is the distance between pattern  $p$  and its closest detector. This distance is inversely related to the cosine of the angle between the vector associated with the closest detector and  $p$ . This set of learning rules may appear non-local in that all detectors are influenced by the closest detector to a pattern, and depend on previous presented inputs. However, the rules can be interpreted as local if the pattern itself transmits a signal to detectors revealing how well covered it is, and if detectors have memories for previously attained matches to patterns. When an input pattern is presented, it will first activate the hidden layer of detectors, and then these detectors will cause the category units to become activated. The activation of the category unit  $A_j$  will be

$$A_j = \sum_{i=1}^d A_i W_{j,i}$$

where  $d$  is the number of detectors. Detector-to-category weights are learned via the delta rule  $\Delta W_{j,i} = L(T - A_j)A_i$  where  $L$  is a learning rate and  $T$  is the teacher signal described above.

We formed a network with 2 detectors units and 2 category units, and presented it with four input patterns. We gave the network four patterns that were used in experiments with human subjects. These patterns are not identical to the patterns shown in Figure 1, but are of the

same abstract construction. When the patterns were categorized as shown in Figure 3A, such that the first two patterns belonged to Category 1, and the second two patterns belonged to Category 2, then on virtually every run, the detectors that emerged were those reflecting the diagnostic segments -- those segments that were reliably present on Category 1 or Category 2 trials. The picture within a detector unit in Figure 3 reflects the entire weight vector from the 15 X 15 input array to the detector. When the same patterns are presented, but are categorized in the orthogonal manner shown in Figure 3B, then different detectors emerge that again reflect the category-diagnostic segments. In both cases, each detector will have a strong association to one and only one of the category units. This is expected given that one of the factors influencing the development of detectors was their categorical diagnosticity. For the results shown here, and the later simulations to be reported, the following parameter values were chosen:  $M=0.1$ ,  $N=0.05$ ,  $O=0.02$ , and  $S=0.1$ . Activation values were between  $-1$  and  $+1$ . One hundred passes through the input materials were presented to the network. In the example shown in Figure 3, only 30 passes with each of the

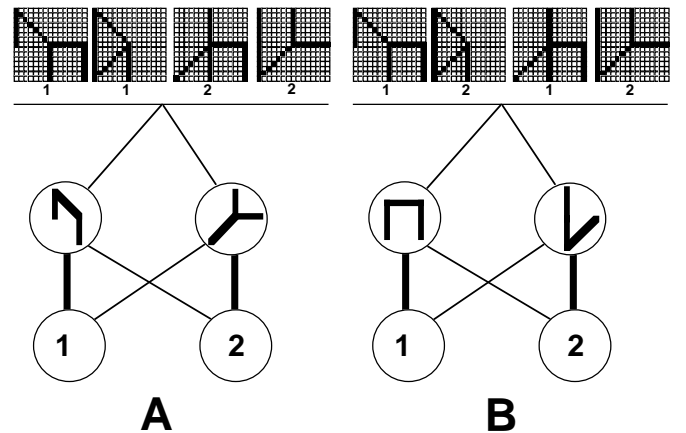


Figure 3: Categorization-dependent detectors are acquired

4 patterns was required for the complete specialization of detectors to input patterns.

### The Segmentation Network

The basic insight connecting categorization and segmentation tasks is that segmentation can also be modeled using competitive learning, and thus the two tasks can share the same network weights and consequently influence on each other. Competitive learning for categorization sorts complete, whole input patterns into separate groups. Competitive learning for segmentation takes a single input pattern, and sorts the pieces of the pattern into separate groups. For segmentation, instead of providing a whole pattern at once, we feed in the pattern one pixel at a time, so instead of grouping patterns, the network groups pixels together. Thus, each detector will compete to cover individual pixels of an input pattern such that the detector with the pixel-to-detector weight that is closest to the pixel's

actual value will adapt its weight toward the pixel's value, and inhibit other detectors from so adapting. With this technique, if the pattern in Figure 4 is presented to the network, the network might segment it in the fashion shown in Panel A. Panels A-D show the weights from the 15 X 15 input array to each of two detectors, and reflect the specializations of the detectors. The two segments are complements of each other — if one detector becomes specialized for a pixel, the other detector does not.

Unfortunately, this segmentation is psychologically implausible. No person would decompose the original figure into these parts. To create psychologically plausible segmentations, we modify the determination of winners. Topological constraints on detector creation are incorporated by two mechanisms: A) input-to-detector weights "leak" to their neighbors in an amount proportional to their proximity in the 15 X 15 array, and B) input-to-detector weights also spread to each other as a function of their orientation similarity, defined by the inner-product of four orientation filters. The first mechanism produces detectors that tend to respond to cohesive, contiguous regions of an input. The second mechanism produces detectors that follow the principle of good continuation, dividing the figure "X" into two crossing lines rather than two kissing sideways "V"s, because the two halves of a diagonal line will be linked by their common orientation. Thus, if a detector wins for pixel X (meaning that the detector receives the more activation when Pixel X is on than any other detector), then the detector will also tend to handle pixels that are close to, and have similar orientations to, Pixel X. The segmentation network, augmented by spreading weights according to spatial and orientational similarity, produces segmentations such as the one shown in Panel B of Figure 4.

Although the segmentation in Panel B is clearly superior to Panel A's segmentation, it is still problematic. The pixels are now coherently organized in line segments, but the line segments are not coherently organized into connected parts. Spreading weights according to spatial similarity should ideally create segmentations with connected lines, but such segmentations are often not found because of local minima in the harmony function (the value N defined on the next page). Local minima occur when a detector develops specializations for distantly related pixels, and these specializations develop into local regions of mutually supporting pixels. Adjacent regions will frequently be controlled by different detectors. Each of the detectors may have sufficiently strong specializations for local regions that they will not be likely to lose their specialization, due to the local relations of mutual support.

Our solution to local minima is to incorporate simulated annealing, by which randomness is injected into the system, and the amount of randomness decreases as a function of time. Unlike standard annealing techniques, we reduce the amount of randomness in the system over time, but do so by basing the amount of randomness on the current structural goodness of a solution (Hofstadter & Mitchell, 1994).

The segmentation network works by fully connecting a 15 X 15 input array of pixel values to a set of N detectors. Although ideally the value of N would be dynamically determined by the input pattern itself, in the current modeling, we assume that each object is to be segmented into two parts (as did Palmer, 1978). When an input pattern is presented, the pixels within it are presented in a random sequence to the detectors, and the activation of Detector i which results from presenting Pixel p is

$$A_{i,p} = \sum_{h=1}^n I_h W_{i,h} S_{h,p}$$

where  $I_h$  is the activation of Pixel h,  $W_{i,h}$  is the weight from Pixel h to Detector i, and S is the similarity between pixels h and p. As such, detectors are not only activated directly by presented pixels, but are also activated indirectly by pixels that are similar to the presented pixels. Thus, a detector will be likely to be strongly activated by a certain pixel if it is already activated by other pixels similar to this pixel.

The similarity between two pixels h and p is determined by

$$S_{h,p} = T \frac{\sum_{i=1}^n G_{ih} G_{ip} L_{i,h,p}}{n} + U e^{-D_{h,p} C}$$

Where T and U are weighting factors,  $G_{ih}$  is the response of orientation filter i to Pixel h,  $L_{i,h,p}$  is the degree to which Pixels h and p fall on a single line with an orientation specified by filter i,  $D_{h,p}$  is the Euclidean distance between Pixels h and p, and C is a constant that determines the steepness of the distance function. Four orientation filters were applied, at 0, 45, 90, and 135 degrees. The response of each filter was

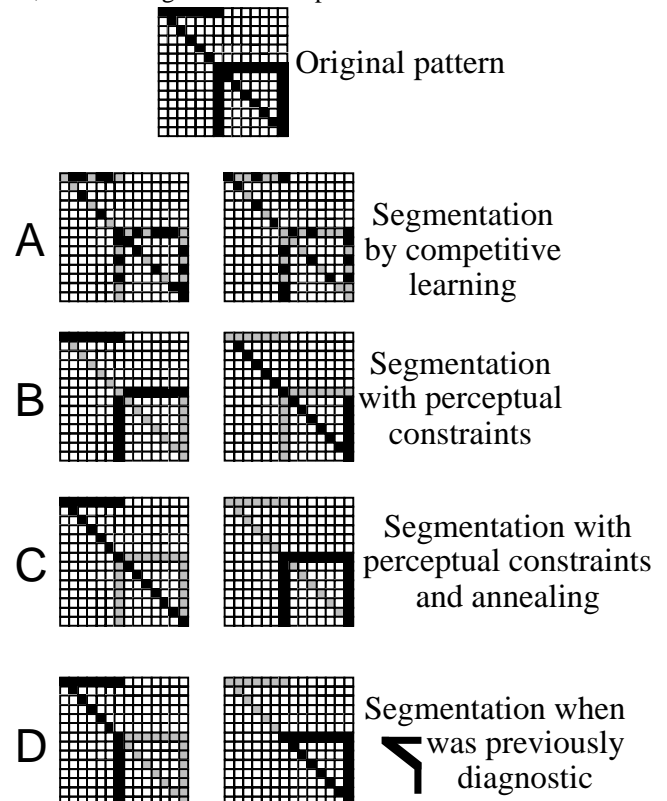


Figure 4. Segmentations of the original figure with incremental improvements from A-D.

found by finding the inner product of the image centered around a pixel and a 5 X 5 window with the image of one of the four lines. Thus, the greater the overlap between the line and the image, the greater will be the output of the filter for the line. The alignment of two pixels along a certain direction was found by measuring the displacement, in pixels, between the infinite length lines established by the two pixel/orientation pairs.

Pixel-to-Detector weights are learned via competitive learning:

$$\Delta W_{i,p} = \begin{cases} M(I_p - W_{i,p}) + \text{Random}(-N,+N) & \text{if } \forall x(A_{i,p} \geq A_{x,p}) \\ \text{Random}(-N,+N) & \text{otherwise} \end{cases}$$

Where M is a learning rate, and Random(-N,+N) generates Gaussian random noise between + and - N. The amount of noise, N, in adjusting weights is a function of the harmony across all detectors relative to R, the maximal harmony in the system:

$$N = R - \sum_{i=1}^n \sum_{p=1}^m \sum_{h=1}^m I_h I_p W_{i,h} W_{i,p} S_{h,p}$$

As such, if similar pixels in similar states have similar weights to detectors, then the harmony in the system will be high, and the amount of noise will be low. Thus, the amount of randomness in the weight learning process will be inversely proportional to the coherency of the current segmentation. These learning equations allow the network to regularly create the segmentation shown in Panel C of Figure 4.

In the simulations of the segmentation network to be reported, no attempt was made to find optimally fitting values of the constants. T and U were set at 0.5, M was set at 0.1, and C was set to 1.

### Combining the Networks

Considered separately, the categorization and segmentation networks each can be considered to be models of their respective tasks. However, they were also designed to interact, with the aim of accounting for the results from Pevtzow and Goldstone's (1994) experiments with human subjects. The segmentation network, because it shares the same input-to-detector weights that were used for the categorization network, can be influenced by previous category learning. Detectors that were diagnostic for categorization will be more likely used to segment a pattern because they have already been primed. Thus, if a particular shape is diagnostic and reasonably natural, the network will segment the whole into this shape most of the time, as shown in Panel D Figure 4. In short, category learning can alter the perceived organization of an object. By establishing multi-segment features along a bank of detectors, the segmentation network is biased to parse objects in terms of these features. Thus, two separate cognitive tasks can be viewed as mutually constraining self-organization processes. Categorization can be understood in terms of the specialization of perceptual detectors for particular input patterns, where the specialization is influenced by the diagnosticity of a segment for

categorization. Object segmentation can be viewed as the specialization of detectors for particular parts within a single input pattern. Object segmentation can isolate single parts of an input pattern that are potentially useful for categorization, and categorization can suggest possible ways of parsing an object that would not otherwise have been considered.

In order to model the results from the earlier human experiments, the network was first trained on distortions of the patterns A, B, C, and D shown in Figure 1, with either a horizontal or vertical categorization rule. As with the human experiment, the distortions were obtained by adding one random line segment to each pattern in a manner that resulted in a fully contiguous form. Following 30 randomly ordered presentations of distortions of the four patterns, the segmentation network was then presented with the original object shown in Figure 5. Segmentations were determined by examining the stable input-to-detector weight matrix for each of the two detector units.

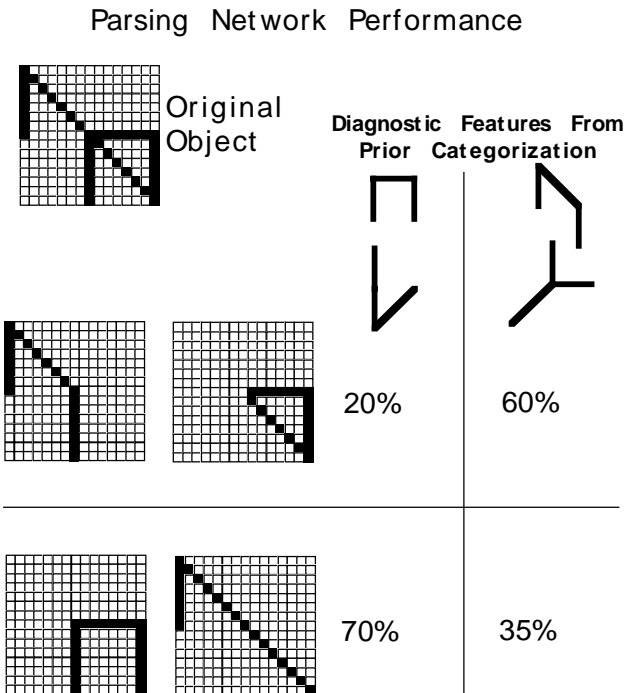


Figure 5. The segmentation of an ambiguous object is influenced by prior category learning.

One hundred subjects were simulated in each of the two pre-segmentation categorization conditions. As the results from Figure 5 indicate, the segmentation of the ambiguous original object is influenced by category learning. In particular, the original object tends to be segmented into parts that were previously relevant during category learning (column percentages do not add up to 100% because of rarely occurring alternative segmentations). As such, the results from Pevtzow and Goldstone (1994) are predicted under the additional assumption that response times in a

part/whole task are related to the likelihood of generating a segmentation that includes the probed part.

In a subsequent test of the networks, the actual wholes used by Pevtzow and Goldstone (1994) in their part/whole task were presented to the segmentation network. Each whole was presented 200 times, 100 times preceded by each of the two possible categorization rules. Out of the 24 whole objects tested, segmentations involving categorization-relevant parts were produced more often than segmentations involving irrelevant parts for 19 of the objects. This comparison controls for any intrinsic differences in naturalness between segmentations of a whole object because the parts that are categorization-relevant for half of the simulated subjects are irrelevant for the other half. As such, the results from Figure 5 generalize to the actual materials used in the experiment. Human subjects and the simulation were exposed to same image-based materials, rather than presenting a digested and abstracted stimulus representation to the simulation.

## Conclusions

A pair of neural networks were presented that learned to group multiple objects into categories, and learned to group parts from a single object into segments. More importantly, the computational modeling provides a mechanism by which one type of grouping influences the other. Category learning causes detectors to develop, and once these detectors have developed, there is a tendency to use the detectors when segmenting an object into parts.

Future work will be necessary compare the model to other existing models that allow for experience-dependent visual object segmentation (e.g. Behrmann et al., 1998; Mozer, Zemel, Behrmann, & Williams, 1992). Two extensions of the model would clearly be desirable: 1) allowing the model to determine for itself how many segments a pattern should be decomposed into, and 2) allowing the computed segmentation of a single pattern to influence its categorization. The latter extension is required to fit human experimental evidence suggesting that not only does category learning influence segmentation, but the perceived segmentation of an object influences its categorization (Schyns et al, 1998; Wisniewski & Medin, 1994).

The computational model, and associated experimental results, support theories that propose that categorization does not simply employ fixed descriptions such as geons, textons, holons, oriented lines segments, or spatial filters, but also creates new object descriptions. The primary advantage of such a state of affairs is that the perceptual system can become tuned and specialized to environmental demands. Cognitive science researchers who have proposed particular fixed sets of primitives have been clever, and have designed primitives that are useful for representing words, objects, and events. However, everyday people may be almost as clever as these researchers have been, and may be able to come up with their own sets of elements tailored to important categorizations (Schyns et al, 1998). Once created, these elements are then used for interpreting

subsequently encountered objects. To the person who has a hammer, the world looks like a nail, and to the person who has learned that a particular configuration is relevant for categorization, the world looks like it is composed out of that configuration.

## Acknowledgments

This research was funded by National Institute of Health Grant R01MH56871-01A2, and a Gill fellowship.

## References

- Behrmann, M., Zemel, R. S., & Mozer, M. C. (1998). Object-based attention and occlusion: Evidence from normal participants and a computational model. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1011-1036.
- Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*, 178-200.
- Goldstone, R. L., Steyvers, M., Spencer-Smith, J., & Kersten, A. (2000). Interactions between perceptual and conceptual learning. In E. Diettrich & A. B. Markman (Eds.) *Cognitive Dynamics: Conceptual Change in Humans and Machines*. (pp. 191-228). Lawrence Erlbaum and Associates.
- Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, *11*, 23-63.
- Hofstadter, D. R., & Mitchell, M. (1994). The Copycat project: A model of mental fluidity and analogy-making. In K. J. Holyoak and J. A. Barnden (Eds.) *Advances in Connectionist and Neural Computation Theory, Volume 2*. (pp. 31-112). Norwood, NJ: Ablex.
- Lesgold, A., Glaser, R., Rubinson, H., Klopfer, D., Feltovich, P., & Wang, Y. (1988). Expertise in a complex skill: Diagnosing x-ray pictures. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise*. (pp. 315-335). Hillsdale, NJ: Erlbaum.
- Mozer, M. C., Zemel, R. S., Behrmann, M., & Williams, C. K. I. (1992). Learning to segment images using dynamic feature binding. *Neural Computation*, *4*, 650-665.
- Palmer, S. E. (1978). Structural aspects of visual similarity. *Memory & Cognition*, *6*, 91-97.
- Pevtzow, R., & Goldstone, R. L. (1994). Categorization and the parsing of objects. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. (pp. 717-722). Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, *9*, 75-112.
- Schyns, P. G., Goldstone, R. L., & Thibaut, J. (1998). Development of features in object concepts. *Behavioral and Brain Sciences*, *21*, 1-54.
- Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, *18*, 221-281.
- Zemel, R. S., Behrmann, M., Mozer, M. C., & Bavelier, D. (1999). Experience-dependent perceptual grouping and object-based attention. Unpublished manuscript.



# The Determinants of Basic-Level Performance

Frédéric Gosselin (GOSELIF@PSY.GLA.AC.UK)  
Philippe G. Schyns (PHILIPPE@PSY.GLA.AC.UK)

Department of Psychology, University of Glasgow  
58 Hillhead St., Glasgow G12 8QB UK

## Abstract

SLIP (*Strategy Length & Internal Practicability*) is a new model of basic-level performance that postulates two computational constraints on the basic-levelness of a category: the number of feature tests required to place the input in a category (*Strategy Length*) and the ease with which these tests are performed (*Internal Practicability*). This article reports three experiments that examined the validity of SLIP in two-level taxonomies of computer-synthesized artificial objects. Experiment 1 isolated *strategy length*, Experiment 2, *practicability*, and Experiment 3 explored the interactions of these factors. Whereas SLIP predicted the RT of these experiments, two established basic-level models of basic-level performance, Jones's (1983) *category feature-possession* and Corter and Gluck's (1991) *category utility*, did not.

What distinguishes your cellular phone, your fountain pen, your computer, your house, and other everyday objects of yours from those of your neighbors is often a combination of features. For example, to identify your pink Porsche 911 in a parking lot also comprising a pink Toyota Tercel and a lime Porsche 911, you must examine both the *color* and the *shape* of the cars. This is so because real-world things share many features. The hierarchical organization of categories is a direct consequence of this sharing of features.

In a seminal article, Rosch, Mervis, Gray, Johnson & Boyes-Braem (1976) distinguished between three natural levels of categorization hierarchy (or taxonomy), the subordinate (e.g., "Porsche 911"), basic (e.g., "car") and superordinate (e.g., "vehicule"), from the most specific to the most general. Of all these levels, they showed that the basic was the best in many respects. People tend to: designate things by their basic-level names; list many more features at the basic level than at any other level; decide more rapidly that things are members of their basic categories than of any others; and so on.

SLIP (*Strategy Length & Internal Practicability*) is designed to model one of the most important index of basic-level performance: categorization speed. It postulates an ideal categorizer that performs the fewest possible number of features tests to classify things. Its name is derived from the fact that its attention slips off its ideal track once in a while.

Going back to the parking lot example, you had to check both the *color* and the *shape* of cars to find your pink Porsche 911. Fewer tests would have not lead to a definitive decision (because there were also a lime Porsche 911 and a pink Toyota Tercel). We call this optimal series of tests a *strategy*. Two aspects of strategies fully determine the response time of SLIP: their *length* and their *internal practicability*. Strategy length is simply the minimal number of tests required to complete a

strategy. In the parking lot example, strategy length is equal to 2 (one test on *shape*; one test on *color*). The longer a strategy associated with a category, the more time it will take to categorize an object in this category. The second factor of SLIP—*internal practicability* (or *practicability*, for short)—is simply the ease with which a particular test in a strategy can be executed (e.g., the number of features that uniquely define this category). The greater the practicability of a category, the less time it will take to verify that an object belongs to this category. SLIP integrates strategy length and internal practicability to predict categorization time (see the Appendix for formal details).

In Gosselin and Schyns (1997, 1999) we demonstrated that the principles of SLIP better predict the results of 22 classic basic-level experiments (from Rosch et al., 1976; Murphy and Smith, 1982; Mervis & Crisafi, 1982; Hoffmann & Ziessler, 1983; Murphy, 1991; Lassaline, 1990; Tanaka & Taylor, 1991; Johnson & Mervis, 1997; and Gosselin & Schyns, 1998) than the leading models (e.g., Jones's, 1983, *category feature-possession* and Corter & Gluck's, 1991, *category utility*).

No matter how successful, these simulations are *a posteriori* accounts of data. The validity of SLIP would be better tested with a direct examination of strategy length and internal practicability<sup>1</sup>. In three experiments, we isolated the possible role of these factors on basic-levelness. Specifically, Experiment 1 isolated the effect of strategy length, Experiment 2, the effect of internal practicability, and Experiment 3 the interactions of the two factors.

## Experiment 1

Experiment 1 examines the effect of *strategy length* on basic-levelness. Strategy length is the minimum number of required feature tests to perform a given categorization.

Experiment 1 is set up as a category verification task of two two-level taxonomies of artificial objects (see Figure 2). The taxonomies are designed to induce orthogonal patterns of categorization speed across conditions. In the HIGH\_FAST taxonomy, overlap of geons between categories defines a shorter strategy at the higher than at the lower level. In the LOW\_FAST taxonomy, different geon arrangements yield the reverse situation—i.e. longer strategies at the top than at the bottom level. SLIP predicts that shorter strategies are completed faster than longer strategies, irrespective of categorization levels.

<sup>1</sup> Furthermore, it must be noted that our data set of 22 published experiments is itself biased to mid-then-high-then-low level. Any model that predicts this RT sequence will be 58% right, irrespective of the actual hierarchy.

Hence, on the basis of only strategy length, SLIP predicts orthogonal categorization performance across taxonomies.

## Method

### Participants

Twenty University of Glasgow students with normal or corrected vision were paid to participate in the experiment.

### Stimuli

Stimuli were computer-synthesized chains of four geons (see Figure 1) similar to those used in Tarr, Bülthoff, Zabinski and Blaz (1997) in the context of object recognition. We designed stimuli with a three-dimensional modeling software on a Macintosh computer.



Figure 1. A four-geon chain used in Experiment 1.

Five geons defined the categories of the HIGH\_FAST taxonomy. One different geon defined each one of three high-level categories. Each one of six possible low-level categories was further specified by one of the two remaining geons. Figure 2 illustrates this taxonomy.

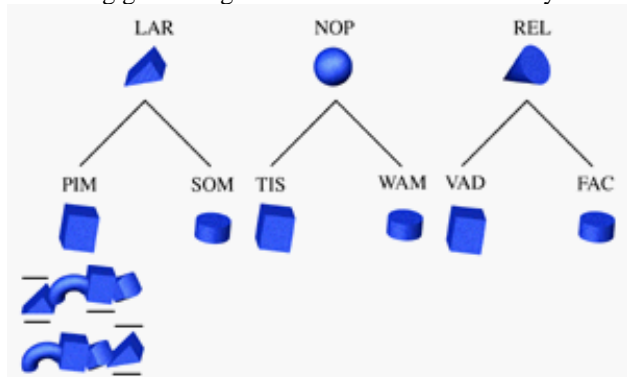


Figure 2. The HIGH\_FAST taxonomy of Experiment 1. The geons specify the defining information of each category. The bottom geon chains are the two PIM exemplars (they are also LAR exemplars) used in the experiment.

In Figure 2, strategy length equals 1 for the higher-level categories. A length 1 strategy means that only one feature needs to be tested (the feature defining each high-level category) to determine the membership of the objects at this level. Strategy length equals 2 at the lower-levels, because these categorizations require two feature tests. The longer strategies arise from the overlap of features across lower-level categories. Shortly put, in the taxonomy of Figure 2, lower-level categorizations require conjunctions of features to be tested. This difference between strategy lengths across the levels of a taxonomy is the backbone of Experiment 1. To create the actual

experimental stimuli, we further added two geons that served as *fillers* to obtain six four-geon objects. Fillers were identical across objects and so could not be used to categorize them. We created two exemplars per low-level category by changing the location of the diagnostic geons in the chain (see Figure 2, the bottom geon chains).

Nine geons defined the LOW\_FAST taxonomy. A unique combination of two geons (sampled from a set of three) defined each one of three top-level categories (see Figure 3). High-level strategies had length 2 because a two-geon conjunction had to be tested. A unique diagnostic geon further specified the bottom-level categories. Bottom-level categories had length 1 strategies because a single feature test determined membership. Figure 3 shows the LOW\_FAST taxonomy. We added one filler to generate six four-geon chains. From these, we created two exemplars per category (see Figure 3).

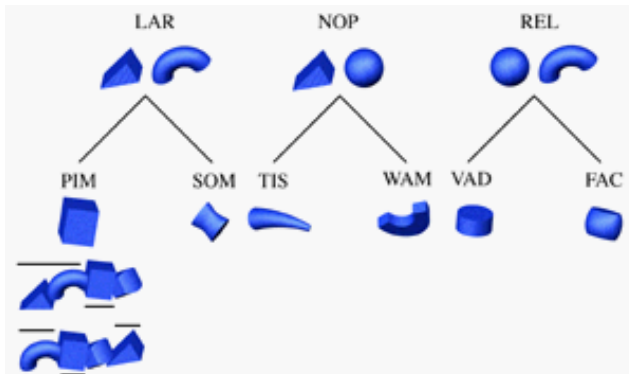


Figure 3. The LOW\_FAST taxonomy of Experiment 1. The geons specify the defining information of each category. The bottom geon chains the two PIM exemplars (they are also LAR exemplars) used in the experiment.

### Procedure

The procedure followed closely that of Murphy (1991). In a learning phase, participants were evenly split between the learning of the HIGH\_FAST and LOW\_FAST taxonomies. We instructed participants to learn the names and the defining geon(s) of nine categories. Participants saw their taxonomy on a sheet of paper (see Figures 2 and 3); this learning phase was not constrained in time.

We tested participants' knowledge of the taxonomy by asking them to list the features associated with each category name. Learning criterion was to list twice in a row, without any mistake, the defining features of each category. Corrective feedback was provided.

When subjects knew the taxonomy, a category verification task measured categorization time at each level. Each trial began with the presentation of a category name. Subjects could then press the "continue" computer keyboard button to see the list of all learned definitions on the screen (each definition corresponded to a set of geons per category). Participants had to identify the list associated with the previously shown category name. This insured that subjects accessed the representation of this category. After a 200 ms delay, an object appeared on the screen. Subjects had to decide—as fast as they possibly could—whether or not the named category and the object matched by pressing the "yes" or "no" computer keyboard key. We recorded response latencies. Note that low-level categories are more numerous than high-level categories. We normalized the number of positive and negative trials

with the constraint of equating the number of trials per level.

### Results and discussion

We performed the analysis of RTs on the correct positive trials that were within two standard deviations from the means. Table 1 reports the mean RTs at the low- and high-levels for the two taxonomies tested (see Observations in Table 1).

Table 1. Mean RTs for the Positive Trials of All Experiments As Well As Predictions of SLIP, Category Feature-Possession, and Category Utility (Erroneous Predictions Are Shaded).

	Model	Level	
		Low	High
Exp. 1, HIGH_FAST	Observation	1,256	896
	Possession	2	3
	Utility	.195	.222
	SLIP	6.4	3.2
Exp. 1, LOW_FAST	Observation	948	1,240
	Possession	1	3
	Utility	.25	.333
	SLIP	3.2	6.4
Exp. 2, HIGH_FAST	Observation	788	660
	Possession	1	3
	Utility	.375	.500
	SLIP	3.2	2.286
Exp. 2, LOW_FAST	Observation	740	774
	Possession	3	1
	Utility	.624	.500
	SLIP	2.286	3.2
Exp. 3, EQUAL	Observation	672	680
	Possession	1	5
	Utility	.176	.260
	SLIP	1.714	1.714
Exp. 3, SL_DOWN	Observation	920	1,058
	Possession	1	5
	Utility	.250	.333
	SLIP	1.714	3.429
Exp. 3, IP_UP	Observation	928	775
	Possession	1	5
	Utility	.250	.333
	SLIP	6.857	3.429

A two-way (GROUP x STRATEGY LENGTH) ANOVA of the RTs with repeated measures on one factor (STRATEGY LENGTH) revealed a main effect of STRATEGY LENGTH,  $F(1, 18) = 77.08, p < .0001$ , (mean length 1 strategies = 922 ms verification time; mean length 2 strategies = 1248 ms verification time), meaning that participants systematically verified length 1 strategies faster than length 2 strategies, irrespective of the considered level (low vs. high). Neither the interaction between GROUP and STRATEGY LENGTH,  $F(1, 18) = .84, ns$ , nor the main GROUP effect,  $F(1, 18) = .02, ns$ , were significant. The error rate was low overall and was not correlated with RT ( $r = -.17, ns$ ), ruling out a speed-accuracy trade-off.

Remember that SLIP predicts that length 1 strategies should be completed faster than length 2 strategies, irrespective of categorization level (see SLIP in Table 1 for numerical predictions with  $S = .25$ ). The data reported here confirms that strategy length, rather than categorization level, determines participants RTs.

## Experiment 2

Practicability refers to the ease with which the features identify a category at any level of a taxonomy. A category has high practicability if many of its defining features are uniquely diagnostic of this category (or if the features occupy few positions across exemplars). It will have low practicability if only one feature defines the category (or if the features can occupy many positions across exemplars). Practicability has so far been the only factor under study in basic-level experiments (see Gosselin & Schyns, 1997). Never has it been shown, however, that practicability could affect the basic-levelness of all categorization levels.

Experiment 2 isolates practicability in a two-level taxonomy using objects similar to those of Experiment 1. All strategies had length 1 but the high and low levels differed in practicability. In the HIGH\_FAST condition, high-level strategies had greater practicability than low-level strategies. The opposite applied to the LOW\_FAST condition, with low-level strategies having higher practicability. SLIP predicts that categories with higher practicability will be verified faster, irrespective of their level in the taxonomy.

### Method

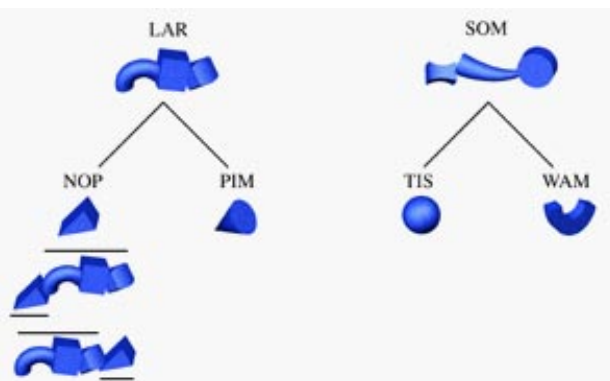
#### Participants

Twenty students from University of Glasgow with normal or corrected vision were paid to participate in the experiment.

#### Stimuli

Stimuli were similar to those of Experiment 1: four-geon chains synthesized with a three-dimensional modeling software on a Macintosh computer.

The HIGH\_FAST condition used 10 diagnostic geons. Three different geons defined each one of two high-level categories; one different geon further defined each low-level category (see Figure 4). We generated two exemplars per category by changing the location (either rightmost or leftmost of the chain) of the three geons



defining the high-level categories (see Figure 4).

Figure 4. The HIGH\_FAST taxonomy of Experiment 2. The geons specify the defining information of each category. The bottom geon chains are the two NOP exemplars (they are also LAR exemplars) used in the experiment.

The LOW\_FAST condition involved fourteen diagnostic geons. A single diagnostic geon defined each one of two high-level categories, and three different geons further defined each one of four low-level categories (see Figure 5). As before, we created two category exemplars by changing the location (either far right or far left of the object) of the triplets defining the low-level categories (see Figure 5).

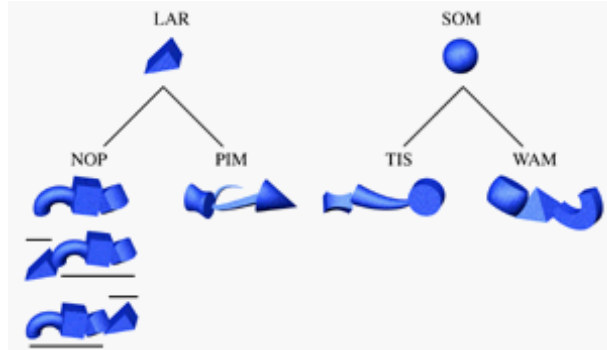


Figure 5. The LOW\_FAST taxonomy of Experiment 2. The geons specify the defining information of each category. The bottom geon chains are the two NOP exemplars (they are also LAR exemplars) used in the experiment.

Practicability is greater for high-level categories in the HIGH\_FAST condition and for the low-level categories in the LOW\_FAST condition because more unique features are associated with the top- and bottom-level categories, respectively. SLIP predicts a faster verification performance for categories with higher practicability (high in HIGH\_FAST and low in LOW\_FAST) irrespective of the level of the taxonomy considered.

#### Procedure

The procedure followed in all respects that of Experiment 1: Participants were randomly assigned to the HIGH\_FAST and LOW\_FAST conditions. They were taught their respective taxonomy before being measured on the categorization speeds of its levels. Each one of 280 trials consisted in the initial presentation of a category name followed by an object. Participants had to decide as fast as they possibly could whether the two matched and we recorded response latencies.

#### Results and discussion

We analyzed only the correct positive trials RTs within two standard deviations from the means. Table 1 shows the mean RTs at the low and high-levels for the HIGH\_FAST and for the LOW\_FAST taxonomies.

A two-way (GROUP x PRACTICABILITY) ANOVA on the RTs with repeated measures on one factor (PRACTICABILITY) revealed a main effect of practicability,  $F(1, 18) = 16.83$ ,  $p = .001$  (mean verification time = 700 ms for high practicability strategies; 781 ms for low practicability strategies). Out of 20 participants, only three did not respond faster to the greater practicability categories. Neither the GROUP x PRACTICABILITY interaction,  $F(1, 18) = 5.53$ , ns, nor

the main GROUP effect,  $F(1, 18) = .06$ , ns, were significant. The error rate was low overall and was not correlated with RT ( $r = .05$ , ns), ruling out a speed-accuracy trade-off.

In sum, SLIP predicted that greater practicability strategies should yield faster categorization decisions (see SLIP in Table 1 for numerical predictions with  $S = .25$ ). The results of Experiment 2 reveal that this factor determined RTs at different categorization levels.

### Experiment 3

Experiments 1 and 2 respectively revealed that strategy length and internal practicability—the two computational determinants of SLIP—can contribute independently to faster categorizations at any level of a taxonomy. Experiment 3 further explores how these two factors interact to determine performance.

There are many possible interactions to investigate and we will not investigate them all. Instead, we have selected to examine three scenarios that selectively change the fastest categorization level by selectively modifying either strategy length or internal practicability.

In the EQUAL scenario, strategies at the high and low-levels have an equal length of 1 and a constant practicability. SLIP predicts that in these circumstances, categorization speeds should be equal across levels. EQUAL is our baseline condition. In the SL\_DOWN scenario, practicability is constant across levels, but whereas low-level strategies have length 1, high-level strategies have length 2. SLIP predicts faster categorizations at the lower level. The IP\_UP scenario preserves the difference in strategy lengths, but it changes the fastest categorizations to the higher level by decreasing the practicability of the low level.

Together, EQUAL, SL\_DOWN and IP\_UP illustrate how the faster categorization level can go up and down a taxonomy by changing strategy length or the internal practicability, the two factors of SLIP.

#### Method

##### Participants

Thirty students from University of Glasgow with normal or corrected vision were paid to participate in the experiment.

##### Stimuli

Stimuli were similar to those of Experiments 1 and 2: geon chains designed with a 3D-object modeling software.

Nine diagnostic geons entered the composition of categories in the EQUAL, SL\_DOWN and IP\_UP conditions. In EQUAL, one geon defined each one of the nine categories of the taxonomy (see Figure 6). We added four fillers to each defining geon to form a total of six six-geon chains. We placed the geons defining the high-level categories at the far left of the chains, and those defining the low-level categories at the far right (see Figure 6).

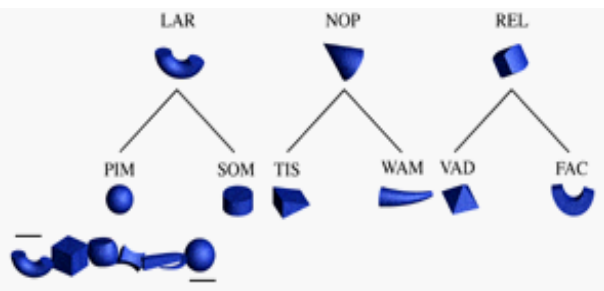


Figure 6. This illustrates the EQUAL taxonomy of Experiment 3. The geons specify the defining information of each category. The bottom geon chain is the PIM exemplar (it is also a LAR exemplar) used in the experiment.

In SL\_DOWN, a unique combination of two of the nine geons defined each top-level category. The addition of one different geon further defined each lower-level category. (SL\_DOWN employed the Experiment 1, LOW\_FAST taxonomy but with a different set of geons). We produced six six-geon chains by adding three fillers. We placed the geon pairs defining the high-level categories at the far left of the chains, and those defining the low-level categories at the far right.

These chains also served to construct the exemplars of condition IP\_UP. Here, we generated four exemplars per category by changing only the location in the chain of the single geon defining the low-level categories (one of the four rightmost positions in the six-geon chains).

#### Procedure

The procedure was almost identical to that of experiments 1 and 2. Participants were randomly assigned to one of three conditions (EQUAL, SL\_DOWN, and IP\_UP). Following a learning of their taxonomy, they did 240 verification trials. Each trial consisted in the presentation of a category name followed by an object. Participants had to decide whether these matched and we measured response latencies.

#### Results and discussion

We performed the analysis of RTs on the positive, correct trials that were within two standard deviations from the means. Table 1 shows the mean RTs.

A two-way (GROUP x LEVEL) ANOVA with repeated measures on LEVEL revealed a significant interaction between GROUP and LEVEL,  $F(2, 27) = 11.85, p < .001$ , simple main effects of GROUP(SL\_DOWN) by LEVEL,  $F(1, 27) = 10.58, p = .003$ , GROUP(IP\_UP) by LEVEL,  $F(1, 27) = 13.09, p = .001$ , and GROUP(EQUAL) by LEVEL,  $F(1, 27) = .04, ns$ . The error rate was low overall and was positively correlated with RT ( $r = .31, p < .05$ ), ruling out a speed-accuracy trade-off.

SLIP predicted all the results observed in Experiment 3 (see SLIP in Table 1 for numerical predictions with  $S = .25$ ). Participants categorized equally fast at both levels in EQUAL. Increasing the strategy length of the higher level in SL\_DOWN induced faster categorizations of the lower level. Diminishing practicability at the lower level then made the high level faster. Thus, the two computational factors of SLIP predicted speed of categorization in taxonomies.

#### General Discussion

SLIP (Strategy Length & Internal Practicability) is a new model of basic-level performance. Three verification experiments tested the two computational determinants of the model: strategy length and internal practicability. In Experiment 1, strategy length was shown to decide basic-levelness. In Experiment 2, practicability was shown to be a second determinant of basic-level performance. In Experiment 3, interactions between strategy length and internal practicability in SLIP predicted the observed RTs.

SLIP performance can be compared to that of two well-established measures of basic-levelness, *category feature-possession* (Jones, 1983) and *category utility* (Corter & Gluck, 1990). The predictions of the models are given in Table 1. The scores of both category utility and category feature-possession should be inversely proportional to RTs; SLIP's scores should be directly proportional to RTs. The best predictor is SLIP with seven correct RT patterns out of seven, followed by category feature-possession with a hit rate of 5/7 (the mistakes have been shaded in Table 1), and trailed by category utility with a 4/7 hit rate.

It is instructive to decompose these scores into strategy length and internal practicability scores. For the conditions testing only practicability (Experiment 2, HIGH\_FAST and LOW\_FAST, and Experiment 3, EQUAL and IP\_UP), category feature-possession and category utility both predict 3/4 of all RT patterns. We have demonstrated elsewhere (Gosselin & Schyns, 1999) that these models are biased to faster responses at higher levels.

Interestingly, for the conditions testing only strategy length (Experiment 1, HIGH\_FAST and LOW\_FAST, and Experiment 3, EQUAL and SL\_DOWN) category feature-possession and category utility only predict 4/8 and 2/8 of the RTs, respectively. (Note that Experiment 3, EQUAL, is included in the break-down into strategy length and internal practicability; it is an extreme case of both.) This confirms the argument that category feature-possession and category utility neglect strategy length as a specific factor of basic level performance (Gosselin & Schyns, 1997). This is a serious problem because attributes do overlap between categories in the real-world, and so strategy length is an important factor of categorization performance outside the laboratory.

To the extent that any model of categorization implements computational constraints (even if these are not well specified), the conclusion is that those of SLIP are closest to those underlying the speed of access to the categories of a taxonomy.

#### Acknowledgements

The first author was supported by scholarships from the Fonds pour la formation de Chercheurs et l'Aide à la Recherche (FCAR) and from the University of Glasgow during this research. This research was funded by ESRC grant R000237901 to the second author.

#### Reference

- Corter, J. E. & Gluck, M. A. (1992). Explaining basic categories: Features predictability and information. *Psychological Bulletin, 111*, 291-303.
- Gosselin, F. & Schyns, P. G. (1997). Debunking the basic level. *Proceedings of the nineteenth annual conference*

- of the cognitive science society (pp.277-282). New Jersey: Lawrence Erlbaum.
- Gosselin, F. & Schyns, P. G. (1998). The contingency of parts in object concept. *Proceedings of the twentieth annual conference of the cognitive science society* (p.1222). New Jersey: Lawrence Erlbaum.
- Gosselin, F. & Schyns, P. G. (1999). *Why do we SLIP to the basic-level? Computational constraints and their implementation*. Manuscript submitted for publication.
- Hoffmann, J. and Ziessler, C. (1983). Objectidentifikation in kunstlichen begriffshierarchien. *Zeitschrift fur psychologie*, 194, 135-167.
- Johnson, K. E. and Mervis, C. B. (1997). Effects of varying levels of expertise on the basic level of categorisation. *Journal of Experimental Psychology: General*, 126, 248-277.
- Jones, G. V. (1983). Identifying basic categories. *Psychological Bulletin*, 94, 423-428.
- Lassaline, M. E. (1990). *The basic level in hierarchical classification*. Unpublished master's thesis, University of Illinois.
- Mervis, C. B. and Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development*, 53, 258-266.
- Murphy, G. L. & Smith, E. E. (1982). Basic level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior*, 21, 1-20.
- Murphy, G. L. (1991). Parts in objects concepts: Experiments with artificial categories. *Memory & Cognition*, 19, 423-438.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382-352.
- Tanaka, J. W. & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23, 457-482.
- Tarr, M. J., Bülthoff, H. H., Zabinski, M. & Blanz, V. (1997). To what extent do unique parts influence recognition across changes in viewpoint? *Psychological Science*, 8, 282-289.

## Appendix

A category is defined by a list of features. Typically, some of these features are unique to this category and some overlap with the defining features of other categories. An optimal strategy is the shortest series of tests on the features defining the category. We posit that SLIP categorizers always use optimal strategies. We call redundant features, or set of redundant features, the collection of features which, individually, provide exactly the same information as to the category membership of objects. In other words, testing one, two, or more redundant features does not provide more information.

Formally, we will say that a strategy is a series of sets of redundant features. It has succeeded whenever all sets of redundant features have been completed in a specific order. And a set of redundant features is completed as soon as a test on the presence of one of its redundant features has been performed.

This usually happens after a succession of misses. The probability of having  $t-1$  successive misses is given by  $(1 - \Psi_j)^{t-1}$  where  $\Psi_j$  – when redundancy of sets of features and the number of possible configurations that these can take in objects are taken into account – is equal to  $C_j(I-S) + C_jSR$ ; that is, the practicability of set of redundant features  $j$  or the probability that it will be completed after a single attempt.  $S$  is the probability of a random slip (it was arbitrarily set to .5 throughout the simulations), and  $C_j$  is the probability that the target features will be in the expected configuration (1 / number of configurations). Thus the first term of  $\Psi_j$  is the probability that the SLIP categorizer will guess the feature configuration correctly and that it

will not slip.  $R_j$  is the probability that a random slip will result in a diagnostic test ([cardinality of  $j$ ] / [number of features in objects]). The second term of  $\Psi_j$  is the probability that the categorizer will slip, but that it will guess the correct configuration and will perform a diagnostic feature test.

The probability of a hit is simply 1 minus the probability of a miss. Thus, the probability that the set of redundant features  $j$  will be completed after  $t$  trials is

$$(1 - \Psi_j)^{t-1} \Psi_j,$$

and the probability that a strategy of length  $n$  will have succeeded after  $t$  trials in a certain configuration of hits and misses is

$$\prod_{j=1}^n (1 - \Psi_j)^\Phi \Psi_j,$$

where  $\Phi$  is a function of  $j$  (it will remain unspecified) which gives the number of misses for the  $j$ th set of redundant features for that particular configuration. Usually, more than one such configuration exist. In fact, the number of possible configurations is easy to compute. The last hit necessarily happens at the  $t$ th trial; the  $n-1$  other hits, however, can happen anywhere in the  $t-1$  trials left, in order. Therefore, the number of possible configurations is the number of combinations of  $t-1$  items taken  $n-1$  by  $n-1$  that is,

$$\lambda = \binom{t-1}{n-1} = \frac{(t-1)!}{(t-n)!(n-1)!}.$$

We can now give the global shape of the probability that a strategy of lengths  $n$  will succeed after  $t$  trials:

$$\sum_{i=1}^{\lambda} \prod_{j=1}^n (1 - \Psi_j)^\omega \Psi_j,$$

where  $\omega$  is a function of  $i$  and  $j$  that specify the number of misses for the  $j$ th set of redundant features for the  $i$ th configuration of hits and misses. We call this the Response Time Function (RTF). We still have to specify  $\omega$ . We will establish a connection between this function and multinomial expansions. The multinome  $(a_1 + a_2 + \dots + a_n)^{t-n}$

expands into  $\lambda$  different terms, and the sum of the  $n$  exponents of each term is equal to  $t-n$ . It follows that  $\omega$  gives the  $j$ th exponent of the  $i$ th term in this multinomial expansion.

As a global measure of basic-levelness, we use  $t\_mean$ , the mean number of tests required to complete a strategy. When internal practicability is constant within a strategy (this is true for all experiments reported in this article), the RTF is a Pascal density function and, thus,

$$t\_mean \text{ is equal to } \frac{n}{\Psi}.$$

# Latent Semantic Analysis Captures Causal, Goal-oriented, and Taxonomic Structures

Arthur Graesser (a-graesser@memphis.edu)

Department of Psychology, University of Memphis, CAMPUS BOX 526400  
Memphis, TN 38152 USA

Ashish Karnavat (akarnavat@hotmail.com)

Department of Computer Science, University of Memphis, CAMPUS BOX 526429  
Memphis, TN 38152 USA

Victoria Pomeroy (vpomeroy@memphis.edu)

Department of Psychology, University of Memphis, CAMPUS BOX 526400  
Memphis, TN 38152 USA

Katja Wiemer-Hastings (kwiemer@cc.memphis.edu)

Department of Psychology, University of Memphis, CAMPUS BOX 526400  
Memphis, TN 38152 USA

and the Tutoring Research Group

## Abstract

Latent Semantic Analysis (LSA) has been used to represent the domain of computer literacy in AutoTutor, a fully automated computer tutor. The analyses in the present study support the claim that the 200-dimensional LSA space captures aspects of the structured mental models that underlie computer literacy. Knowledge structures were constructed that contained causal networks, goal/plan/action hierarchies, and taxonomic hierarchies. The proximity of a pair of nodes (i.e., concept, state, event, action, goal) in these structures predicted the cosine similarity scores that are routinely computed in LSA analyses.

## Representing World Knowledge with Conceptual Graph Structures

World knowledge has traditionally been captured by knowledge structures throughout the history of cognitive science, artificial intelligence, and discourse processes. The knowledge structure structures include semantic networks, taxonomies, causal networks, planning networks, ontological trees, spatial region hierarchies, and various other classes of conceptual graph structures (Golden, 1997; Graesser & Clark, 1985; Kiel, 1979; Lehmann, 1992; Lenat, 1995; Norman & Rumelhart, 1975; Schank & Abelson, 1977; Trabasso, van den Broek, & Suh, 1989; Sowa, 1983). A knowledge structure contains a set of categorized nodes that refer to concepts, events, processes, states, actions, goals, and other ontological classes. The nodes are connected by

relational arcs that also are assigned to various categories, e.g., is-a, has-as-parts, cause, reason, enables, contains, etc. A particular package of knowledge may incorporate spatial composition, causal networks, goal hierarchies, taxonomic hierarchies and other viewpoints. All of these viewpoints allegedly can be represented as a set of categorized nodes that are integrated by a set of directed, relational arcs.

It is a time consuming, methodical task to map out knowledge structures for a domain of knowledge. Developers of expert systems and other knowledge based systems would require a decade to perform the knowledge engineering that is needed for a system of reasonable scope with widespread practical applications (Lenat, 1995). There are authoring tools that guide either experts or novices in the building of the knowledge structures (Williams, Hultman, & Graesser, 1998). The structures are built in a principled fashion that caters to the constraints of the composition rules, so guidance is needed to prevent illegal compositional structures. All of this takes training and experience that can be measured in months or years. However, conceptual graph structures are powerful theoretical entities because they support the intelligent procedures and processes that operate on the representations, as in the case of retrieval, classification, summarization, problem solving, question asking, question answering, and so forth.

The distance between two nodes in a conceptual graph structure is frequently regarded as a metric of conceptual relatedness. That is, the conceptual relatedness between

nodes A and B decreases as a function of the number of arcs that exist on a legal path between A and B. For example, if 1 arc separates A and B on a causal chain, then A and B are strongly related, compared to the case where 4 arcs separate two nodes on a causal chain. The structural proximity between any two nodes that are connected by a legal path of arcs is designated as its structural-proximity (A, B).

### Representing World Knowledge with Latent Semantic Analysis

Researchers have more recently turned to Latent Semantic Analysis (LSA) because it provides an approximation of the representation of world knowledge, but in a very short period of time -- measured in weeks, days or even hours. LSA is a statistical representation of a body of world knowledge that is reflected in a large corpus of textual documents (Landauer & Dumais, 1997; Landauer, Foltz, & Latham, 1998). LSA capitalizes on the fact that particular words appear in particular texts (called "documents"); the cooccurrence of words in documents reflects the constraints that exist in world knowledge. The input to LSA is a cooccurrence matrix that specifies the number of times that word  $W_i$  occurs in document  $D_j$ . These frequencies are adjusted with a logarithm transformation that also corrects for the base rates of words appearing across documents. A word is a distinctive index for a document to the extent that its occurrence in the document is above the base rate for that word across documents. A standard statistical method, called singular value decomposition, reduces the large  $W \times D$  cooccurrence matrix to  $K$  dimensions (typically, 100 to 500 dimensions). Each word, sentence, or text ends up being represented as a weighted vector on the  $K$  dimensions.

The similarity or conceptual relatedness between two bags of words (A and B) is computed as a geometric cosine (or dot product) between the two vectors. The values normally range from 0 to 1. This LSA match between two language strings is designated as its LSA-match (A, B). The LSA match can be high even though there are few, if any words in common between the two strings. LSA allegedly goes well beyond simple string matches because the meaning of a language string is partly determined by the company (other words) that each word keeps (Landauer & Dumais, 1997).

The empirical success of LSA has been promising and sometimes remarkable. Landauer and Dumais (1997) created an LSA representation with 300 dimensions from 4.6 million words that appeared in 30,473 articles in Grolier's Academic American Encyclopedia. They submitted to the LSA representation the synonym portion of the TOEFL test, a test developed by the Educational Testing Service to assess how well non-native English speakers have mastered the words in the English language. The test has a four-alternative, forced choice format, so there is a 25% chance of answering the questions correctly. The LSA model selected the alternative that had the highest match with a

comparison word. The LSA model answered 64.4% of the questions correctly, which is essentially equivalent to the 64.5% performance for college students from non-English speaking countries. LSA has had remarkable success in capturing the world knowledge that is needed to grade essays of students (Foltz, 1996), to assign texts to students of varying abilities to optimize learning (Wolfe, Schreiner, Rehder, Laham, Foltz, Kintsch, & Landauer, 1998), and to provide effective feedback in the training of summarization skills (E. Kintsch, W. Kintsch, Laham, Landauer, DePaula, Schreiner, Stahl, & Steinhart, 2000). There are now LSA-based graders of essays that assign grades to essays with the validity and reliability of human experts in composition (Foltz, 1996). In our research on computer literacy, LSA has been quite successful in evaluating the quality of college students' answers to deep reasoning questions and to the contributions of learners during the tutorial interactions with AutoTutor (Graesser, Wiemer-Hastings, Wiemer-Hastings, Harter, Person, & the TRG, 2000; Wiemer-Hastings, Wiemer-Hastings, Graesser, and the TRG, 1999).

The success of LSA is quite remarkable given that it was never designed to capture many of the traditional problems in language understanding systems, such as word order, syntax, quantification, and negation. There are other corpus-based probabilistic models that capture word order and syntax (Burgess, Livesay, & Lund, 1998; Charniak, 1993) but the present study focuses on the capabilities of LSA.

At this point, there is a great deal of uncertainty about what is being represented in the  $K$ -dimensional spaces of LSA. One optimistic possibility is that the  $K$  dimensions reflect ontological categories, semantic features, and structural compositions of mental models that would be directly adopted in structural theories of world knowledge representation. For example, a simple and straightforward assumption would be that particular banks of the  $K$  dimensions of LSA would have a one-to-one or many-to-one mapping onto ontological categories (Chi, Slotta, & de Leeuw, 1994; Keil, 1979), to conceptual primitives (Miller & Johnson-Laird, 1976; Norman & Rumelhart, 1975; Schank & Abelson, 1977), or to the domain-specific features that are associated with a particular topic. Very few researchers would go out on the limb and propose an elegant mapping between the  $K$  dimensions of LSA and sophisticated theories of world knowledge. However, most researchers would seriously entertain the possibility of weaker correspondences. At the other end of the continuum, there are researchers who believe that the  $K$  dimensions have nearly an arbitrary mapping to the attributes of mature theories of world knowledge (Landauer & Dumais, 1997).

A somewhat different question addresses whether the LSA space is capable of recovering aspects of the deeper mental models that underlie text (Forbus, Gentner, & Law, 1995), or what is sometimes called situation models (Kintsch, 1998). Foltz, Britt, and Perfetti (1996) reported evidence



that suggested that LSA does capture mental model representations to some extent, whereas Perfetti (1998) has expressed doubts that LSA captures the representations and processes of psychological models. LSA may capture shallow knowledge rather than deep knowledge. That is, LSA may capture the sort of word associations that are reflected in the archives of dictionaries and encyclopedias, but may not penetrate the deeper mental models. On the other hand, LSA may be successful in capturing aspects of the deeper situation model. An accomplished expert on some topic certainly does know how to use the right bags of words at the right time; the systematic use of words in particular documents may be recovered in the LSA solution spaces. At this point in the science, however, there is not enough empirical evidence to support one position or another.

The present study hopes to shed additional light on what is captured by the LSA representations. An LSA space has been developed in the domain of computer literacy. This LSA representation has been used in a fully automated computer tutor, called AutoTutor (Graesser, Franklin, Wiemer-Hastings, & the TRG, 1998; Graesser et al., in press; Graesser et al., 2000; Wiemer-Hastings, Graesser, Harter, & the TRG, 1998). In addition to the LSA space, AutoTutor has dozens of conceptual graph structures that capture knowledge in a more structured form. The present study examines whether the structural composition of the conceptual graph structures can predict the LSA match scores. That is, is there a significant correlation between structural proximity and LSA match scores when we examine taxonomic hierarchies, causal networks, and goal structures? A positive correlation would support the claim that LSA spaces to some extent recover aspects of the mental models. A zero correlation supports the claim that the K dimensional LSA spaces have an unsystematic mapping onto structural theories of knowledge representation.

### **Corpus of Texts and LSA Space on Computer Literacy**

A 200-dimensional LSA space was developed for the domain of computer literacy during the development of AutoTutor. The corpus of included (a) two books on computer literacy, (b) 30 articles that focus on hardware, operating systems, and the internet, and (c) AutoTutor's curriculum script of lessons, example problems + solutions, and questions + answers. An LSA analysis requires the preparation of a document by word (D x W) co-occurrence matrix. Each cell in the matrix specifies the number of occurrences of word  $W_i$  in Document  $D_j$ . In order to prepare the DxW matrix, the researcher needs to define what constitutes a document unit. A single document was defined as (a) a paragraph in the case of the textbooks and 30 articles and (b) a sentence that conveys a lesson, a good

answer, or piece of a solution in the case of the curriculum script. An LSA analysis was performed on the 2.3 MB corpus of documents, yielding a solution with 200 dimensions.

The 200-dimensional LSA was validated in our assessments of AutoTutor (Graesser et al., 2000; Wiemer-Hastings et al., 1999). For example, Wiemer-Hastings et al. (1999) analyzed how well the LSA space on computer literacy could accurately evaluate a sample of 192 answers to the questions in the curriculum script. College students enrolled in the computer literacy course answered the questions in the curriculum script by typing in their answers into a web cite facility. The data were collected after the college students had read the relevant chapters in the book and had received a lecture on each macrotopic (i.e., hardware, operating system, Internet). Trained experts (such as graduate research assistants) also rated the 192 answers to the questions. The results of correlational analyses revealed that the LSA did an excellent job evaluating the quality of student answers. The correlation between LSA's answer quality scores and the mean quality scores of the experts was .49. This correlation is indistinguishable from the .51 correlation between the ratings of the two intermediate experts (i.e., the individuals who normally grade exams in a college computer literacy course). Graesser et al. (2000) reported that AutoTutor's LSA component did an excellent job discriminating the ability of learners who interact with AutoTutor in a multi-turn tutorial dialog. LSA was capable of discriminating different classes of student ability (good, vague, erroneous, versus mute students) and in tracking the quality of contributions in tutorial dialog.

The LSA space in AutoTutor was adopted in the present study. We computed the LSA-match scores between pairs of nodes in the conceptual graph structures that had been prepared for topics on hardware, operating systems, and the internet.

### **Conceptual Graph Structures on Topics in Computer Literacy**

AutoTutor's architecture includes a set of conceptual graph structures on the various topics in the curriculum script. A typical structure contains approximately 10 to 30 nodes. We randomly selected 12 conceptual graph structures in the present analysis, including 4 structures for hardware, 4 for operating systems, and 4 for the internet.

The 12 knowledge structures were composed by applying the conceptual graph structure (CGS) representations developed by Graesser (Graesser & Clark, 1985; Graesser et al., 1992; Graesser, Wiemer-Hastings, & Wiemer-Hastings, in press; Williams, Hultman, & Graesser, 1998). The CGS's have 5 node categories: concepts, states, events, goals, and style specifications. There are 22 basic arc categories. The composition of these conceptual graph structures is not arbitrary, but is based on formal and conceptual constraints

that have been studied for several decades in artificial intelligence (Lehmann, 1992). The categories of nodes and arcs are sufficient for implementing computational models of question answering which have been validated in experiments on adults (Baggett & Graesser, 1995; Graesser & Hemphill, 1991; Graesser, Lang, & Roberts, 1991).

Three types of knowledge structures were directly analyzed in the present study: taxonomic hierarchies, causal networks, and goal hierarchies. A node was included in the present analysis if and only if it was part of any of these three types of structures. The composition of these three types of structures is specified below.

### Taxonomic Hierarchies

Concept nodes are connected by *is-a* arcs. For example, the concepts Norton Antivirus, utility program, and tool would be connected by two *is-a* arcs:

(concept-1: Norton Antivirus) –*is-a*→  
 (concept-2: utility program) –*is-a*→  
 (concept-3: tool)

The structure distance is 1 between concepts 1 and 2 and between concepts 2 and 3; the structural distance is 2 between concepts 1 and 3.

### Causal Networks

State and event nodes are connected by arcs that signify Cause, Enables, Subprocess, and Implies (see Graesser & Clark, 1985 and Graesser, Wiemer-Hastings, & Wiemer-Hastings, in press for more complete definitions of arcs). Some of these categories of nodes and arcs are illustrated in the following chain.

(state-1: the operating system is stored on the hard disk) – Enable→  
 (event-2: the operating system is loaded onto the computer) –Subprocess→  
 (event-3: the operating system gets into RAM) –Cause→  
 (event-4: the CPU executes instructions)

The structural distance is 1 between nodes 1&2, 2&3, and 3&4, is 2 between nodes 1&3 and 2&4, and is 3 between nodes 1&4.

### Goal-structures

Goal nodes are connected to other nodes by virtue of arcs that signify Reason, Manner, Initiate, and Outcome. For example, the following three goal nodes form a goal hierarchy via a Reason arc.

(goal-1: user types in command) –Reason→  
 (goal-2: user starts word processing software) –Reason→  
 (goal-3: user writes article)

The goals are triggered by various events and states in the world by virtue of Initiate arcs, whereas Outcome arcs specify whether or not the goals are achieved.

### Scaling of Pairs of Nodes on Structural Proximity

Pairs of nodes in the 12 conceptual graph structures were scaled on structural proximity with respect to taxonomic hierarchies, causal networks, and goal structures. A node in a structure was included in the analysis if and only if it was part of one or more of these three types of structures. When considering all 12 conceptual graph structures, there were 536 pairs of nodes in the analysis. A pair of nodes (A and B) was scaled on causal proximity by computing the reciprocal of the structural distance on a legal causal path between A and B (i.e., 1/distance). Thus, if two nodes have a structural distance of 1, 2, 3, versus 4 arcs on a legal path, then the causal proximity scores would be 1.00, .50, .33, and .25, respectively. If there is no legal causal path that connects A and B, the causal proximity score is 0. Goal proximity and taxonomic proximity were computed in a similar fashion for all 536 nodes. The mean proximity scores were .07, .31, and .40 for the taxonomic, causal, and goal proximity scores, respectively; the corresponding standard deviations were .26, .40, and .45.

### Relationship Between LSA Match Scores and Structural Proximity Scores

The analyses uncovered a robust relationship between the LSA match scores and the structural proximity scores. Consider first the causal proximity scores. The mean LSA match scores were .47, .35, and .24 when the causal proximity scores were 1.00, .50, and .33 or lower (but not 0), respectively. When analyzing the goal proximity scores, the LSA match scores were .53 and .42 for goal proximity scores of 1.00 and .50 or lower (but not 0), respectively. The taxonomic proximity scores rarely went lower than 1.00 when considering nonzero values, so we could not isolate a sensitive gradient for this proximity score. The overall mean LSA match score was .44 (SD = .30).

A multiple regression was conducted to assess the extent to which the LSA match scores could be predicted by the taxonomic, causal, and goal proximity scores. The three predictor variables together explained a significant 9% of variance in the LSA match scores,  $F(3, 532) = 16.46, p < .05, R^2 = .09$ . All three predictors had a significant unique impact on the LSA scores, with beta weights of .14, .31, and .47 for taxonomic, causal, and goal proximity, respectively.

We performed some follow-up multiple regression analyses that statistically controlled for some potential extraneous variables. One extraneous variable was the length of the node descriptions, as defined by the number of words in the pair of nodes. Those who have conducted research on LSA have reported that lengthier descriptions have a slight tendency to produce higher LSA matches when

two bags of words are compared (Rheder, Schreiner, Wolfe, Laham, Landauer, & Kintsch, 1998; Wiemer-Hastings et al., 1999). The mean length of the node descriptions in our sample was 10.62 words in the node pair ( $SD = 4.03$ ), or 5.31 words per node. A second extraneous variable was the number of nouns that overlap between the pair of nodes. Overlapping nouns are analogous to argument overlap in propositional theories of text processing (Graesser, Millis, & Zwaan, 1997; Kintsch, 1998); the fact that constituents refer to the same entity is one important foundation for coherence in discourse processing. However, from the standpoint of the present analyses, we would not be particularly surprised if the LSA match scores could be explained by mere noun overlap because it is analogous to a keyword overlap. The mean number of overlapping nouns in a node pair was .71 ( $DS = .67$ ).

Table 1 presents the results of the multiple regression analysis that predicted LSA match scores as a function of the three structural proximity scores, length, and noun overlap. The five predictor variables accounted for a significant 55% of the variance in LSA match scores,  $F(5, 530) = 128.05$ ,  $p < .05$ ,  $R^2 = .55$ . When considering the two extraneous variables, noun overlap had a robust impact on the LSA match scores whereas length had no significant effect. Although noun overlap was robust, the three structural proximity variables still had a significant unique impact on the LSA match scores in the multiple regression analyses. Interestingly, we did not find the noun overlap scores to be correlated very highly with the taxonomic, causal, and goal proximity scores,  $r = -.18$ ,  $.25$ , and  $-.02$ , respectively. Overlap in predicates was also analyzed but the correlations were also modest or nonsignificant. These results support the claim that the structural proximity scores have an impact on LSA match scores over and above keyword matches.

Table 1: Multiple regression analyses that predict LSA match scores

Predictor Variable	beta-weight	t-score
Taxonomic proximity	.14	4.08 *
Causal proximity	.11	2.17 *
Goal proximity	.15	2.94 *
Length (number of words)	-.02	.75
Noun overlap	.72	23.20 *

\* significant at  $p < .05$ .

## Conclusions

The results of this study support the claim that LSA captures aspects of the mental models that underlie computer literacy. The content of the mental models includes taxonomic structures, causal networks, and goal/plan/action hierarchies. The LSA match scores between pairs of nodes in the conceptual graph structures can be predicted by taxonomic, causal, and goal structural proximity. The structural proximity scores predict LSA match scores over and above noun overlap, keyword overlap, and the number of words in the node descriptions.

Aside from demonstrating that LSA captures aspects of mental models, we have demonstrated that LSA can be useful for performing semantic and conceptual analyses on relatively short verbal descriptions. Researchers have sometimes claimed that LSA is only useful when analyzing lengthier verbal descriptions on the order of a paragraph. The present study supports the claim that LSA can be useful for compositional analyses on individual words and short sentences of 5-6 words. Additional research is needed to identify the limits of LSA in recovering different aspects of semantics and world knowledge.

## Acknowledgements

This research was funded by the National Science Foundation (SBR 9720314) in a grant awarded to the first author of this manuscript. The following members of the Tutoring Research Group at the University of Memphis conducted research on this project: Ashraf Anwar, Laura Bautista, Myles Bogner, Tim Brogdon, Patrick Chipman, Scotty Craig, Rachel DiPaolo, Evan Drumwright, Stan Franklin, Max Garzon, Barry Gholson, Art Graesser, Doug Hacker, Peggy Halde, Derek Harter, Jim Hoeffner, Xiangen Hu, Jeff Janover, Ashish Karnavat, Bianca Klettke, Roger Kreuz, Kristen Link, Shulan Lu, Johanna Marineau, William Marks, Lee McCaulley, Brent Olde, Para Orfanides, Natalie Person, Victoria Pomeroy, Penelope Price, Sonya Rajan, Akshay Thota, Mat Weeks, Holly Yetman White, Shannon Whitten, Katja Wiemer-Hastings, Peter Wiemer-Hastings, Shonijie Yang, and Zhaohua Zhang.

## References

- Baggett, W.B., & Graesser, A.C. (1995). Question answering in the context of illustrated expository text. *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 334-339). Hillsdale, NJ: Lawrence Erlbaum.
- Burgess, C., Livesay, K., & Lund, K. (1998). Exploring context space: Words, sentences, discourse. *Discourse Processes*, 25, 211-257.
- Charniak, E. (1993). *Statistical language analysis*. Cambridge, MA: Cambridge University Press.

- Chi, M.T.H., Slotta, J.D., & de Leeuw, N. (1994). From things to processes: A theory of conceptual change for learning science concepts. *Learning and Instruction*, 4, 27-43.
- Foltz, P.W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, and Computers*, 28, 197-202.
- Foltz, P.W., Britt, M.A., & Perfetti, C.A. (1996). Reasoning from multiple texts: An automatic analysis of readers' situation models. In G. Cottrell (Ed.), *Proceedings of the 18<sup>th</sup> Annual Cognitive Science Conference* (pp. 110-115). Mahwah, NJ: Erlbaum.
- Forbus, K., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19, 141-205.
- Golden, R.M. (1997). Causal network analysis validation using synthetic recall protocols. *Behavior Research Methods, Instruments, and Computers*, 29, 15-24.
- Graesser, A.C., & Clark, L.C. (1985). *Structures and procedures of implicit knowledge*. Norwood, NJ: Ablex.
- Graesser, A.C., Franklin, S., & Wiemer-Hastings, P. and the TRG (1998). Simulating smooth tutorial dialog with pedagogical value. *Proceedings of the American Association for Artificial Intelligence* (pp. 163-167). Menlo Park, CA: AAAI Press.
- Graesser, A. C., Gordon, S. E., & Brainerd, L. E. (1992). QUEST: A model of question answering. *Computers and Mathematics with Applications*, 23, 733-745.
- Graesser, A. C. & Hemphill, D. (1991). Question answering in the context of scientific mechanisms. *Journal of Memory and Language*, 30, 186-209.
- Graesser, A. C., Lang, K. L., & Roberts, R. M. (1991). Question answering in the context of stories. *Journal of Experimental Psychology: General*, 120, 254-277.
- Graesser, A.C., Millis, K.K., & Zwaan, R.A. (1997). Discourse comprehension. *Annual Review of Psychology*, 48, 163-189.
- Graesser, A.C., Wiemer-Hastings, P., & Wiemer-Hastings, K. (in press). Constructing inferences and relations during text comprehension. In T.Sanders, J. Schilperoord, & W. Spooen (Eds.), *Text representation: Linguistic and psycholinguistic aspects*. Amsterdam: Benjamins.
- Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., and the TRG (in press). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*.
- Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., and the TRG (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*.
- Kiel, F.C. (1979). *Semantic and ontological development: An ontological perspective*. Cambridge, MA: Harvard University Press.
- Kintsch, E., Kintsch, W., Laham, D., Landauer, T.K., DePaula, R., Schreiner, M.E., Stahl, G., & Steinhart, D. (2000). Learning how to summarize using LSA-based feedback. *Interactive Learning Environments*.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, MA: Cambridge University Press.
- Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
- Landauer, T.K., Foltz, P.W., Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.
- Lehmann, F. (1992)(Eds.). *Semantic networks in artificial intelligence*. New York: Pergamon.
- Lenat, D.B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38, 33-38.
- Miller, G.A., & Johnson-Laird, P.N. (1976). *Language and perception*. Cambridge, MA: Harvard University Press.
- Norman, D.A., & Rumelhart, D.E. (1975). *Explorations in cognition*. San Francisco, CA: Freeman.
- Perfetti, C.A. (1998). The limits of cooccurrence: Tools and theories in language research. *Discourse Processes*, 25, 363-377.
- Rheder, B., Schreiner, M.E., Wolfe, M.B.W., Laham, D., Landauer, T.K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, 25, 337-354.
- Schank, R.C., & Abelson, R. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Erlbaum.
- Sowa, J.F. (1983). *Conceptual structures: Information processing in mind and machine*. Reading, MA: Addison-Wesley.
- Trabasso, T., van den Broek, P.W. & Suh, S. (1989). Logical necessity and transitivity of causal relations in the representation of stories. *Discourse Processes*, 12, 1-25.
- Wiemer-Hastings, P., Graesser, A.C., Harter, D., and the TRG (1998). The foundations and architecture of AutoTutor. *Proceedings of the 4th International Conference on Intelligent Tutoring Systems* (pp. 334-343). Berlin, Germany: Springer-Verlag.
- Wiemer-Hastings, P., Wiemer-Hastings, K., and Graesser, A. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. *Artificial Intelligence in Education* (pp. 535-542). Amsterdam: IOS Press.
- Williams, K.E., Hultman, E., & Graesser, A.C. (1998). CAT: A tool for eliciting knowledge on how to perform procedures. *Behavior Research Methods, Instruments, & Computers*, 30,565-572.
- Wolfe, M.B.W., Schreiner, M.E., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W., Landauer, T.K. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes*, 25, 309-336.

# Rational Assessments of Covariation and Causality

**Gustaf Gredebäck (Gustaf.Gredeback@psyk.uu.se)**

Department of Psychology, Uppsala University  
Box 1225, SE-751 42, Uppsala, Sweden

**Anders Winman (Anders.Winman@psyk.uu.se)**

Department of Psychology, Uppsala University  
Box 1225, SE-751 42, Uppsala, Sweden

**Peter Juslin (Peter.Juslin@psy.umu.se)**

Department of Psychology, Umeå University  
SE-901 87, Umeå, Sweden

## Abstract

Are human contingency judgments based on associationistic principles such as cue competition or on normative principles as specified by rational-cognitive models? In this study, participants learned to predict an outcome from several simultaneously presented cues. They were asked to judge the cues in regard to causal power or statistical concepts such as probability or relative frequency. Uniform application of associationistic principles implies cue-interaction effects of blocking (Experiment 1) and conditioned inhibition (Experiment 2) for all judgments. A rational-cognitive framework predicts cue-interaction effects for causality judgments, but not for probability and relative frequency judgments. The results support the rational-cognitive framework on all accounts.

## Introduction

The ability to detect causal relations in the environment is of utter importance to all organisms. Fortunately, at first glance, at least, we seem to adjust well to such demands. We readily formulate hypotheses about plausible causal relationships and contingencies. Research deriving from the associationist tradition, however, suggests that this optimistic view is unwarranted. It is proposed that, because of cue-interaction effects, our representations are distorted. This study pits a more optimistic view of human contingency judgment based on the metaphor of the mind as an “intuitive scientist” against this associationist view. It does so by comparing causality judgments with probability and relative frequency judgments in an inductive contingency judgment task.

Imagine that you suffer from an allergic reaction, which you believe originates from eating shellfish. It seems reasonable to assume that this hypothesis of causality originates in your recollection of similar events, in this case your memory of eating shellfish and suffering from allergic reactions. Of course, most meals that you have eaten contained neither shellfish nor resulted in allergic reactions. Still your memory tells you that the allergic reaction on numerous occasions co-occurred with dishes that included shellfish. You also recall dinners which included shellfish but which did not lead to allergic reaction, and times when the allergy sprung up in the absence of shellfish. One popular notion is that memories of previous events are categorized in what resembles a 2×2 matrix. In this *contingency matrix* the presence and absence

of a predictor event and an outcome event constitute the two axes.

Table 1: A contingency matrix.

	Outcome	No outcome
Predictor present	A	B
Predictor absent	C	D

According to a normative model, judgments of covariation are based on conditional contingency: that is, on the probability of the outcome (e.g., allergy) in the presence and absence of the predictor (e.g., shellfish). Formally, this can be expressed by the  $\Delta p$  algorithm:

$$\Delta p = \left(\frac{a}{a+b}\right) - \left(\frac{c}{c+d}\right), \quad (1)$$

where  $a$ ,  $b$ ,  $c$ , and  $d$  are the number of times that events A, B, C, and D in the contingency matrix have occurred. A positive contingency is perceived if Cell A and D contain more occurrences than Cell B and C ( $\Delta p > 0$ ). Similarly, a negative contingency is indicated by a negative  $\Delta p$  and a zero  $\Delta p$  indicates no contingency. Perceptions of causal relations are thought to be based on observed covariation registered in a form equivalent to the contingency matrix, and computed by the  $\Delta p$  algorithm. Chapman and Robbins (1990) examined two effects that in a conspicuous way violate the normative  $\Delta p$  model: the cue interaction effects referred to as *blocking* and *conditioned inhibition*.

## Cue Interaction Effects

The allergic reaction used in the example exemplifies a simple causal relation where one potential cause is evaluated with respect to its assumed effect ( $C \rightarrow E$ ). However, such occurrences of unambiguous sole cues seldom arise beyond the realms of clinical test scenarios. Instead, we often evaluate complex situations where multiple potential causes may produce an effect ( $C_1, C_2 \dots C_n \rightarrow E$ ). In Experiment 1 of Chapman and Robbins (1990) participants examined the relationship between the change in prize of four individual stocks (predictors) and the stock market as a whole (outcome). In the first phase, either stock A increased in value, followed by an increase in the value of the market, or stock C increased without a market increase. In the second phase, either of pairs of stocks AB or

$CD$  increased, in both cases followed by a market increase. On one third of the trials in Phases 1 and 2 no stocks increased and the market remained unchanged. After each phase, participants rated the extent to which an increase in each stock predicts a market increase on a scale from -100 to 100. (See Table 2.)

Table 2: The design in Chapman and Robbins (1990).

Phase 1	Phase 2	Test 2
$A+, C-, \emptyset-$	$AB+, CD+, \emptyset-$	$B < D$

Note. (+) = outcome, (-) = no outcome,  $\emptyset$  = predictor absent.

If the normative  $\Delta p$  model is correct the frequency of occurrences between each predictor and the outcome is mapped within separate contingency matrices. This means that each predictor is evaluated in isolation with contingency judgments based on the recollection of frequencies. Because the Stocks  $B$  and  $D$  appear an identical number of times, always in the presence of the outcome, these two stocks should receive identical ratings of predictability in Test 2.

As it turns out, the ratings for individual predictors *interact*. In this case the interaction is *blocking*: of two cues with identical contingencies with the outcome, systematically lower ratings are given to the cue presented with a previously established predictor than to the cue presented with a non-predictor. In a second experiment, Chapman and Robbins examined *conditioned inhibition*. In this design, a predictive cue is followed by the outcome, except when it occurs with a second cue. The second cue, referred to as the inhibitor, is rated lower in predictability than a control cue that has the same objective contingency with the outcome. Cue-interaction such as blocking and conditioned inhibition casts doubt on the  $\Delta p$  model as a descriptive model of human judgment of causation and covariation. The question is whether these results are best explained by an *associationist* or, what we refer to as, a *rational-cognitive* framework.

### Associationist Framework

Cue-interaction is routinely observed in studies of animal learning. This has been taken to indicate that human behavior in contingency judgment tasks is best described by a “grand theory” of learning based on the principles that apply to animal conditioning. Associationistic models therefore adhere to a strong analogy between conditioned ( $CS$ ) and unconditioned ( $UCS$ ) stimuli in classical conditioning and predictors and outcomes in human contingency judgments. Contingency judgments are seen as being based on the associative strength of the relationship between conditioned stimuli (predictors) and the unconditioned stimuli (outcomes). In a multiple-cue task, each  $CS$ - $UCS$  association is based on the informative strength of  $CS_x$  with respect to the  $US$ , in competition with all the  $CS$ s present.

The model most often called upon to explain cue interaction effects is the *Rescorla-Wagner model* (Rescorla & Wagner, 1972), hereafter referred to as the R-W model.

Formally, the model states that:

$$\Delta V_x^{n+1} = \alpha_x \beta_1 (\lambda_1 - V_{total}^n), \quad (2)$$

where  $\Delta V_x^{n+1}$  is the change in associative strength ( $V$ ) of  $CS_x$  as the result of pairing it with  $UCS_I$  on trial  $n+1$ ,  $\alpha_x$  is a learning rate parameter representing the associability of  $CS_I$ , and  $\beta_1$  is the corresponding parameter for  $UCS_I$ ,  $\lambda_1$  is the maximum associative strength that the  $UCS$  can support, called the asymptote ( $\lambda_1 = 1$  in the presence of the  $UCS_I$ , 0 in its absence), and  $V_{total}^n$  is the total associative strength of all  $CS$ s on trial  $n+1$ .

Equation 2 describes the change in associative strength of a  $CS$  as a function of the current associability of the  $UCS$  and  $CS$ , in relation to the remaining associability of the  $UCS$  ( $\lambda_1 - V_{total}^n$ ). A consequence of Equation 2 is that learning will occur only when an outcome is unexpected or surprising in the light of one’s expectations.

According to the R-W model, the blocking described by Chapman and Robbins (1990) is due to cue competition. Once a  $CS_I$ - $US$  association is established, any new  $CS_2$  that is presented with the previous association ( $CS_I, CS_2 \rightarrow US$ ) will not become associated with the outcome. Equation 2 states that the change in associative strength on trial  $n+1$  is defined by the difference between the asymptote ( $\lambda$ ) and the total associative strength ( $V_{total}^n$ ) on trial  $n$ . Since  $CS_I$  already predicts the outcome there is no room for  $CS_2$  to become associated with the  $US$ . In comparing  $CS_2$  with another stimulus  $CS_3$  that has an identical outcome contingency the R-W model thus predicts that  $CS_2$  will be valued as less associative than  $CS_3$ .

Conditioned inhibition is seen as the opposite of excitation. According to the R-W model this phenomenon is expressed by ( $\lambda_1 - V_{total}^n < 0$ ). Since the asymptote itself can never be negative, the expression will only be true when  $V_{total}^n$  is larger than zero, that is, when some excitation already has occurred. Say that stimulus  $CS_I$  leads to an outcome  $E$ , while stimulus compound  $CS_p, CS_2$  leads to absence of the outcome. If later tested individually,  $CS_I$  receives an excitatory value, while  $CS_2$  receives an equally strong inhibitory (negative) value, making the total associative value equal to zero. A number of alternative models have surfaced within the associationistic tradition (e.g., Gluck & Bower, 1988; Pearce, 1994; Van Hamme & Wassermann, 1994). Siegel and Allan (1996) singled out the R-W model as the most successful model and, for the purpose of this paper, the R-W model will represent the associationist framework.

### Rational-Cognitive Framework

A number of theoreticians (Waldmann & Holyoak, 1992; Cheng, 1997) propose that the observation of cue interaction effects does not pose a threat to models based on the  $\Delta p$  algorithm. These results merely indicate the inapplicability of the contingency matrix model to situations involving multiple causes. Imagine that someone claims that alcohol consumption causes lung cancer. In support for this claim it is noted that consumers of alcohol more often suffer from lung cancer than others. With the knowledge that alcohol consumption often is accompanied by smoking this line of reasoning may strike you as odd. Instead it

seems reasonable to assume that smoking is the real cause for the increased risk of cancer.

The difference is that your antagonist is considering a simple causal relationship (unconditional contingency) while you apply a more complex analysis of causal relationships, taking into account the *conditional contingency*. This means that the causal relationship is viewed in the context of the presence and absence of alternative causes. In the smoking example, it is nearly impossible to incorporate all of the potential alternative causes. However, the idea is that, like “intuitive scientists” (Kelley, 1967), we have the capacity to control, at least, for “likely” alternative causes (see also Spellman, 1996).

According to a cognitive-rational approach people store information about events in frequency format. This information is available for different forms of analyses by application of cognitive algorithms. Complex causal relationships require a more sophisticated algorithm than the simple  $\Delta p$  rule (Cheng & Novick, 1990; Cheng, 1997). In *Power PC theory* (Cheng, 1997), for example, the strength of a causal factor is estimated by the conditional contingency: that is, the contingency when other potential causes are controlled for<sup>1</sup>. According to this view, cue interaction is viewed as a consequence of the *participants’ attempts to control for alternative causes*. In Chapman and Robbins (1990), Experiment 1, for example, the participants may have arrived at the conclusion that the causal relationship between Stock *B* and the outcome event (*E*) is uncertain, since its effect is nullified once control for Stock *A* is performed by a conditional contrast; ( $p(E|AB) - p(E|A) = 0$ ). Applying the same algorithm to Stock *D* indicates that *D* is a strong causal factor ( $p(E|CD) - p(E|C) = 1$ ). The controlling for alternative causes would thus lead to results that coincide with the blocking effect predicted by the R-W model.

The same line of reasoning is applicable to conditioned inhibition. Participants conclude that a stimulus has negative causal power (i. e. power to prevent an effect) due to its conditional contingency with respect to other causes (for a discussion, see Cheng, 1997). Because people are assumed to act like scientists in applying rational arguments and interpreting patterns of covariation in terms of unobservable causes, we refer to this as the rational-cognitive framework.

### The Effect of Judgment Type

If the assumption of veridical representation of frequencies in models like power PC-theory is correct, we should not expect interaction effects for judgments of probability or relative frequency (at least, to the extent that that probability ratings are based on representations of relative frequencies). This is a crucial difference between the R-W model and power-PC theory. An orthodox interpretation of the R-W model, which presupposes that the same processes underlie judgments of causality and covariation, predicts that there

should be no difference between conceptually distinct ways of probing for the relationship between cue and outcome. The judgments are mapped from associative strengths and we expect interaction effects for judgments of causality, probability, and frequency alike. In contrast, the rational-cognitive framework presupposes correct representations of environmental frequencies, which deviate from causality ratings in predictable ways. In short: With power PC-theory there is a distinction between judgments of causality and covariation, with the R-W model there is not.

Cue interaction has occasionally been reported also with judgments of probability or frequency (Chapman, 1991, Price & Yates, 1995). Nevertheless, the main body of empirical findings on blocking and conditioned inhibition rests on assessments of often vaguely defined judgment scales (e.g., predictability). We know of no study of conditioned inhibition with relative frequency or probability judgments. In this study, we thus compared judgments of a) the *causal power* of the predictor on the outcome, b) the *probability* of the outcome given the event, and c) the *relative frequency* of the outcome given that the predictor was present on the trial, in a between-subjects design. The associationist account suggests cue-interaction effects with all three judgments. The rational-cognitive approach implies cue interaction effects for causality judgments, but an absence of this effect with the other two judgments.

### Experiment 1: The Blocking Effect

Experiment 1 of Chapman and Robbins (1990) examined the blocking effect in a multiple cue task involving stock market predictions (Table 2). Participants were asked to rate the “predictability” of each stock on a scale from -100 to 100. The present experiment applies the design from Chapman and Robbins’ Experiment 1, with the addition of three new groups. In addition to predictability (to replicate their results), the participants judged either explicit causality, explicit probability, or relative frequency.

### Method

**Participants.** Participants were 64 undergraduate students from Uppsala University. They received either course credit or a movie ticket in exchange for their participation.

**Materials and procedure** The experiment was divided in two learning-phases ( $L_1, L_2$ ) and two test-phases ( $T_1, T_2$ )<sup>2</sup>, appearing in the order  $L_1, T_1, L_2, T_2$ . In each learning-phase participants were to assess whether the stock market as a whole would change in value based on the individual movement of four fictional stocks (see Table 2).  $L_1$  contained 36 trials ( $12 \times A+$ ,  $12 \times C-$ , and  $12 \times \emptyset-$ ).  $L_2$  contained 76 trials ( $24 \times AB+$ ,  $24 \times CD+$ ,  $24 \times \emptyset-$ ). In each test phase, participants assessed the relationship between the increase of each separate stock and an increased stock market. One group rated the stocks according to their *predictability* on

<sup>1</sup> It is important to note, however, that covariation is merely one component of the process of causal induction in power PC theory. Another important component is an a priori framework for interpreting input in terms of causal mechanisms (Cheng, 1997).

<sup>2</sup> Only the most central results, that is, those of the second test phase are included in the present paper.

a scale from -100 to 100, in accordance with Chapman and Robbins (1990). The other three groups rated either relative frequency (*In what percentage of the occasions on which the stock X increased, did the outcome occur?*), probability (*Given that stock X increases, what is the probability of the outcome?*), or causality (*To what degree does stock X cause the outcome?*) with respect to an increased market on a scale from 0 to 100 (a bi-directional scale does not apply to frequency and probability). The judgments were varied between groups.

## Results

The results from Experiment 1 are presented in Table 3. A blocking index was calculated by subtracting ratings to predictor *B* from ratings given to predictor *D*, where a positive score indicates a blocking effect. As illustrated in Figure 1A, both the mean predictability blocking index 27.2 and the mean causality blocking index 24.3 show significant blocking. In contrast, the mean probability blocking index 5.7 and the mean frequency blocking index -15.2 show no sign of the blocking effect (the latter is even negative).

Table 3: The average rating of stimuli *A* through *D* during test Phase II. 95% confidence intervals within parentheses.

	Predictor			
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
Predictability	53.1 (21.4-84.8)	35.9 (10.2-61.7)	-10.6 (-35.3-14.0)	63.1 (35.9-90.2)
Causality	89.3 (78.133-100.4)	49.3 (34.7-63.8)	27.1 (10.1-44.1)	73.6 (55.4-91.8)
Probability	69.7 (55.1-84.2)	52.3 (43.4-61.3)	36.7 (19.5-53.8)	58.0 (43.2-72.8)
Frequency	87.5 (70.8-104.2)	98.4 (95.4-101.5)	71.6 (52.0-91.1)	83.2 (67.0-99.5)

We replicate the findings by Chapman and Robbins (1990) with blocking effects for predictability judgments, and the effect is even more consistent with the causality scale. These results support the idea that the judgment labeled predictability is interpreted as a mix of causality and covariation. Within the same settings we fail to observe cue interaction effects for probability and relative frequency judgments. This predicted pattern is significant as concluded from a planned comparison of means analysis,  $F(1, 57) = 9.8, p < .01^3$ .

The results thus support theories such as power PC theory with respect to its analysis of causal reasoning, as well as the more general notion in rational-cognitive models that (roughly) veridical representations of event frequencies are preserved. At the same time the results are in opposition to the R-W model. To extend and validate the result from the

<sup>3</sup> Because the predictability group used a bi-directional scale whereas the other groups used unidirectional scales, the scores for all groups were standardized within each condition before they were entered into the planned comparison.

first experiment, a second experiment was conducted in order to examine conditioned inhibition.

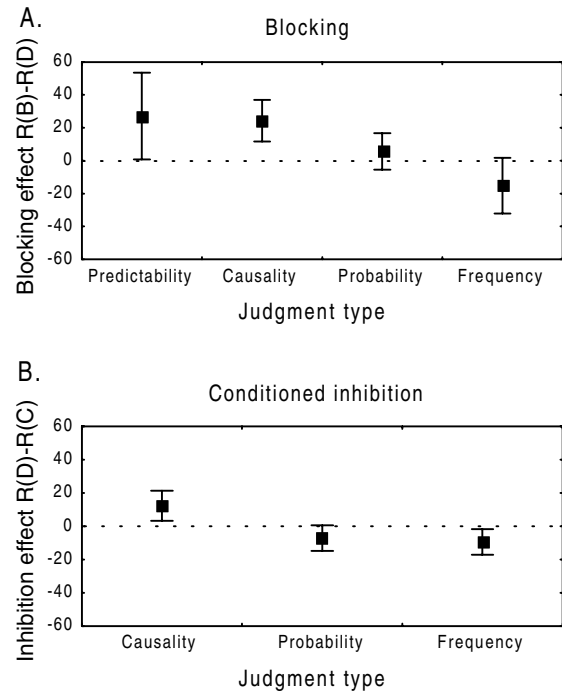


Figure 1: Panel A: Blocking effect in Experiment 1 as a function of judgment. Panel B: Conditioned inhibition in Experiment 2 as a function of judgment.

## Study 2: Conditioned Inhibition

Experiment 2 of Chapman and Robbins (1990) was the first study to examine the conditioned inhibition effect in humans. The results showed clear signs of conditioned inhibition. Williams, Sagness, and McPhee (1994) report several failures to replicate Chapman and Robbins' (1990) cue interaction effects. They reasoned that these failures might be due to the way participants approach the task. People can either interpret a stimulus compound (*A, B*) as one distinct stimulus (*AB*) or as the combination of the two stimuli (*A and B*). The former is termed a *configural encoding* while the latter is referred to as an *elemental encoding*. In a multiple cue task conditioned inhibition can not be obtained with configural encoding: the effect demands that the participants view each stimuli in isolation. Williams et al. therefore tried to experimentally encourage participants to engage in elemental encoding.

After a couple of fruitless attempts to replicate the effect we therefore abandoned the design of Chapman and Robbins (1990). Inspired by Williams et al.'s (1994) Experiment 2, we attempted to promote elemental strategies in favor of configural strategies. In order to take every measure to obtain an effect we made some additional changes. In the original task the outcome always occurs with the positive predictor alone, but never with this predictor in conjunction with the inhibitor. The deterministic design may promote the learning of explicit rules, which



might diminish a true effect. Furthermore, floor effects might mask a real effect since both the inhibitor and the control cue can be expected to be rated at the lower end of the scale.

We therefore made the task probabilistic. The outcome occurred with a probability of .95 in the presence of the positive predictors alone, and with a probability of .3 in the presence of the positive predictor in conjunction with the inhibitor. This modification deals with both of the unfortunate characteristics of the Chapman and Robbins (1990) design. In addition, separate single presentations of the negative cue which have been found to increase an effect (Williams, 1995) were added. Finally, the content of the task was changed. In the original task, the content consists of stocks that change. It could be argued that this content does not encourage inhibition since it may be hard to create a mental model of a causal mechanism of how a particular stock hinders the outcome to occur. We used a task of evaluating experimental fertilizers (cf. Spellman, 1994) where it is easier to form a model of how a particular substance may hinder growth<sup>4</sup>. To summarize: Study 2 was designed to investigate whether conditioned inhibition effects will occur also for probability and relative frequency judgments or if a dissociation will be observed between these and judgments of causality.

## Method

**Participants** Seventy-five undergraduates from Uppsala University took part in the study. They received a movie ticket or course credit in exchange for their participation

**Materials and procedure** The experiment was divided in two tasks. The first (pretraining) involved one learning phase and one test phase, the second (main experiment) was divided in two learning-phases ( $L_1$ ,  $L_2$ ) and two test-phases ( $T_1$ ,  $T_2$ ), appearing in the order  $L_1$ ,  $T_1$ ,  $L_2$ ,  $T_2$ . Both the pretest and the experiment involved the prediction of whether a plant would produce flowers or not after an observation of which fertilizers that had been added to an irrigation fluid. The learning phase of the pretraining was identical to the *Explicit Condition* of Williams et al. (1994). It consisted of  $12 \times X+$ ,  $12 \times XY+$ ,  $12 \times Y-$  and  $12 \times Z-$  trials. When finished, participants were asked to rate each fertilizer with respect to the outcome according to either causality, probability, or relative frequency<sup>5</sup> on a scale from 0-100. The purpose of the pretreatment was to encourage an elemental encoding and results of this phase were not investigated further. In Phase  $L_1$  of the main experiment either of fertilizers  $A$  through  $E$  were added to the liquid 20 times each. In the case of fertilizers  $A$  and  $B$  the plant produced flowers in 19/20 (95%) of the occasions.

<sup>4</sup> A change of content should not affect the result according to associative models, which imply independence of content. In cognitive-rational models on the other hand content may play an important role since prior causal models potentially can affect interpretation according to these.

<sup>5</sup> Due to the similar results for predictability and causality ratings in Study 1 and the deviating scale for the predictability ratings, the latter were dropped in Study 2.

Fertilizers  $C$ ,  $D$ , and  $E$  were coupled with the outcome on 6/20 (30%) occurrences. In  $L_2$  three constellations of fertilizers were followed by the outcome with a base rate of 19/20. These were  $A$ ,  $B$ , and  $AB$ <sup>6</sup>. Fertilizer  $E$ , as well as fertilizer combinations  $AC$  and  $DE$  was coupled to the outcome with a base rate of 6/20. Thus, in this design, predictor  $C$  is the inhibitor and predictor  $D$  is the control, with exactly the same contingency with the outcome and number of occurrences. Table 4 describes the conditioned inhibition design in the experiment. After the learning phases, participants rated the relationship between each fertilizer and the outcome based on causality, probability, or frequency with the same scale as described in Study 1. Throughout Study 2, each participant only made judgments for one of the three scales.

Table 4: Conditioned inhibition design in Experiment 2.

Phase 1	Phase 2	Test 2
A+ B+ C- D- E-	A+ B+ E- AB+	C<D?
	AC- DE-	

Note. (+) indicates a probability of outcome of .95, (-) indicates a probability of outcome of .3.

## Results

Table 5 presents the results of Experiment 2. To repeat, conditioned inhibition is observed if stimulus  $C$  is rated lower than stimulus  $D$ . A significant conditioned inhibition effect was found in the causality group (average difference between  $D$  and  $C = 12.4$ : See Figure 1B). In neither of the other groups is there conditioned inhibition. Both groups have higher ratings for the  $C$  than the  $D$  predictor (difference between  $D$  and  $C = -9.5$  in the frequency group and  $-7.2$  in the probability group.). In fact, this reversed difference is significant in the frequency group. A planned comparison shows that the predicted difference between the causality group and the probability and frequency groups is significant ( $F(1, 72) = 13.7, p < .001$ ). The reason for the significant difference in the opposite direction in the frequency group is unclear, but interestingly the trend was the same in Experiment 1. An explanation, (undeniably speculative) could be that higher level deductive reasoning influence frequency ratings; Maybe participants reason that since predictor  $A$  occurred often together with the outcome and predictor  $C$  often occurred in conjunction with predictor  $A$ , then predictor  $C$  probably occurred quite often together with the outcome. Note, however, that the observed significance in no way indicates that the frequency ratings are severely distorted. A comparison between true frequencies vs. rated probabilities and frequencies show that these agree approximately (although the ratings are moderately regressive). In neither case are the true frequencies excluded by the confidence intervals for the ratings, making it impossible to

<sup>6</sup> The conjunction of the positive predictors  $AB$  was included in order to eliminate the possibility that participants learned a rule implying that a conjunction of any two predictors was followed by a decreased probability of the outcome.

reject the hypothesis that these are made on basis of undistorted representations of the true frequencies. These results are in line with predictions of a rational-cognitive model, with conditioned inhibition effects only for judgments of causality.

Table 5: The average rating of stimuli A through E during test Phase II (95% confidence intervals within parentheses).

	Predictor				
	A	B	C	D	E
Causality	83.9 (78.1-89.7)	88.2 (82.0-94.5)	23.80 (16.3-31.3)	36.2 (28.8-43.5)	36.0 (27.8-44.2)
Probability	82.0 (73.9-90.1)	90.1 (85.0-95.1)	32.8 (24.3-41.2)	25.6 (17.9-33.2)	33.3 (23.8-42.8)
Frequency	79.8 (71.8-87.7)	87.4 (81.6-93.2)	33.2 (24.9-41.5)	23.7 (16.3-31.2)	29.9 (21.4-38.4)

## Discussion

In this paper, we have contrasted two different frameworks for the processes that underlie human contingency judgment. An associationist account which stresses the similarity to the processes derived from learning in animals, as epitomized in the R-W model, and one rational-cognitive account that relies on the metaphor of the mind as an intuitive scientist.

The rational-cognitive account implies that the participants can appreciate a distinction between judgments that concern the causal power of a factor, and judgments that pertain to covariation, such as probability and relative frequency. On this view, blocking and conditioned inhibition arise from appropriate considerations of the confounding between multiple potential causes. This reasoning is compatible with—and indeed presupposes—availability of accurate information about frequencies. An orthodox interpretation of the R-W model, presuming the same process behind judgments of causality and covariation, suggests no effect of the judgment type manipulation. On any account, the model does not provide an explanation for the observed effect. The results from two separate experiments, with fairly disparate designs covering the two most well known cue interaction effects clearly favor the rational-cognitive account. The participants seem to appreciate the distinction between a judgment of causality and judgments of probability and relative frequency.

These results suggest that, functionally the same behavior may be implemented by different mechanisms in different organisms. The same behavior that is computed by associationist processes in lower animals may be the results of high-level reasoning in humans. This conclusion may come as no surprise: Regardless of our ontogenetic sophistication we all share the challenge of dealing with a complex and uncertain environment, and the evolutionary and adaptive pressures we face may thus be very similar in the end.

## References

- Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 837-854.
- Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition*, *18*, 537-545.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.
- Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*, 545-567.
- Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*, 225-244.
- Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on motivation* (Vol. 15, pp. 192-238). Lincoln: University of Nebraska Press.
- Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, *101*, 587-607.
- Price, P. C., & Yates J. F. (1995). Associative and rule-based accounts of cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 1639-1655.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: II. Current theory and research* (pp. 64-99). New York: Appleton-Century-Crofts.
- Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review*, *3*, 314-321
- Spellman, B. A. (1996). Conditionalizing causality. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 167-206). San Diego: Academic Press.
- Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonrepresentation of compound stimulus elements. *Learning and motivation*, *25*, 127-151.
- Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*, 222-236.
- Williams, D. A. (1995). Forms of inhibition in animal and human learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *21*, 129-142.
- Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 694-709.

# Function-Follows-Form Transformations in Scientific Problem Solving

**Todd W. Griffith**

Computer Science Department  
Bucknell University  
Lewisburg, PA 17837  
(570) 577-3721  
tgriffth@bucknell.edu

**Nancy J. Nersessian and Ashok Goel**

College of Computing  
Georgia Institute of Technology  
Atlanta, Georgia 30332-0280  
(404) 894-5612  
{nancyn,goel}@cc.gatech.edu

## Abstract

This paper presents a pattern of reasoning called “function-follows-form” (FFF) uncovered through a study of scientific problem solving. In the study we modeled eleven “think-out-loud” problem-solving protocols taken by John Clement (1989). Our work involved computationally modeling the reasoning processes of eleven scientists each attempting to solve the same problem about springs. We describe experiments with two computational systems, ToRQUE and ToRQUE2, which were used to model salient reasoning found in the protocols, and we show how the use of function-follows-form reasoning patterns enables exploration and conceptual change.

## Introduction

Our research identifies and elucidates a pattern of reasoning we call function-follows-form (FFF) reasoning. We have shown that this pattern of reasoning plays an important role in exploratory problem solving, and may lead to significant change to a subject’s mental models. Here we present specification of FFF resulting from experiments with two successive computational systems called ToRQUE and ToRQUE2. The study involved modeling the problem solving of eleven scientists each attempting to solve the same problem about springs. We used “think-out-loud” protocols collected by John Clement (1989) and performed experiments testing the fidelity of our computational model with the protocols.

This research represents a melding of disciplines with the goal of understanding complex scientific problem solving. We have combined techniques from history and philosophy of science, cognitive psychology, and artificial intelligence to study the problem solving of scientists. The focus of our effort discussed here was to capture the salient aspects of problem solving for each of the scientists in the form of a general competence model, encoded in a computational system (i.e. ToRQUE2).

## Background

As a first attempt at developing an interpretation of scientific problem solving Nersessian and Greeno (1992) examined an extensive expert problem-solving protocol obtained in a “think-out-loud” interview conducted by John Clement (1989). In particular, they were interested in the second protocol (S2), because it exhibited many of the characteristics of James Clerk Maxwell’s problem-solving practices in the construction of the electromagnetic field

concept. As they interpret this protocol, the subject uses what they call “constructive modeling” to satisfy himself that his initial answer to a problem was the correct answer. They saw this process as primarily one of arriving at a model that is of the same kind with respect to the salient features of the spring problem. They argue that while this example is much more constrained than historical cases of scientific discoveries, it is still complex enough to require dealing with the many quite difficult modeling issues historical discoveries present.

Clement’s own analysis of S2’s reasoning focuses on a process he calls modeling via “bridging analogies”. He characterizes this process as one in which the subject “produces models via a successive refinement process of hypothesis generation, evaluation, and modification or rejection” (p.358, Clement 1989). It is the specific nature of the construction and “successive refinement” process that led Nersessian & Greeno to interpret S2’s reasoning as a form of constructive modeling, and subsequently led to our computational theory of generative modeling (Griffith *et al* 1996, Griffith *et al* 1997, Griffith 1999).

## The Problem

According to Clement, S2 was a computer scientist who had some training in physics. He had also passed comprehensive examinations in mathematics in the area of topology, which is highly significant to our interpretation of the protocol session.

In the protocol, S2 is asked to solve the following problem about springs:

“... a weight is hung from a spring. The original spring is replaced with a spring made of the same kind of wire; with the same number of coils; but with coils that are twice as wide in diameter. Will the spring stretch from its natural length more, less, or the same amount under the same weight? (Assume the mass of the spring is negligible compared to the mass of the weight.) Why do you think so?”

In our interpretation, S2 began the problem-solving session with an intuitive understanding that the stretch of a spring is due to its flexibility. Then he derived a new understanding that a spring maintains constant slope when stretched through torsion in the spring’s wire. So, although this is a more modest outcome of scientific reasoning than evidenced in historical cases, for S2 it was an instance of highly creative problem solving leading to conceptual change. To

find a satisfactory explanatory model for the problem solution, S2 had to generate a novel representation of how a spring works. He did so by generating a series of successive models through what we call FFF transformations.

## Methodology

This research focuses on investigating the reasoning processes found in all eleven protocols in order to place S2's creative problem-solving in a context. In so doing, we highlight the reasons that lead to his discovery of torsion as a central causal element in the function of a spring. The additional protocols show scientists attempting to solve the spring problem. All the scientists were expert problem solvers, though none were experts in the domain. The protocols were modeled in two sets. The first set of five protocols (S1-S5) was used to build and refine the ToRQUE system. The second six protocols (S6-S11) were used to evaluate the refined ToRQUE2 system. Experiments were conducted at each stage of development in order to evaluate hypotheses with respect to the methods and knowledge used by the subjects. The first set of experiments, used to refine the systems performance with respect to the first five protocols, involved the ablation and reconfiguration of tasks, methods, and knowledge in order to determine what aspects of the system enabled accurate modeling of the first five subjects. The second set of experiments were also ablation and reconfiguration experiments. For these experiments the system was left unchanged but was "reconfigured" to account for each of the remaining six subjects. This means that reasoning elements such as tasks, methods, or knowledge structures were removed or reordered but not added, and that no reimplementations were done on the ToRQUE2 system during the testing phase. Both sets of experiments looked at the choice of knowledge structures and reasoning methods used by the system, as well as the ordering and availability of knowledge and methods. The system was evaluated based on its ability to accurately model the salient reasoning of subjects.

## Ontologies for Function-Follows-Form

Function-Follows-Form transformations are based upon a series of ontological commitments with respect to the control of processing and the representation of knowledge, each of which is based upon past computational results. The language for the control of processing is called the task, method, knowledge language which is based upon a TMK architecture, while the language for representing physical systems is called the structure, behavior, function (SBF) language which was first developed as part of the theory of adaptive modeling. A reasoning packet comprises patterns from each of these languages.

*The TMK Language:* The Task, Method, Knowledge (TMK) architecture is a theory of control of processing that was first developed by Goel & Chandrasekaran (1992) in an analysis of the methods used for addressing complex tasks. This work was continued in (Goel *et al* 1994, Punch, Goel,

& Brown 1995, Goel *et al* 1996). The theory posits that high-level tasks such as conceptual design can be broken down into a hierarchy of methods and subtasks. Each task or subtask may have one or more methods that can be applied to solve the task. It also posits that each method specifies the sub-tasks that it spawns and control information for the ordering of those sub-tasks. Using multiple methods enables the architecture to account for a variety of reasoning strategies for addressing any one task, where a *strategy* is some sub-hierarchy of the task-method tree whose root is a method.

One advantage of the TMK architecture is that knowledge can have a direct effect on which method is selected to accomplish a particular task. For the purpose of modeling multiple subjects this feature is particularly important. In general each subject has different initial knowledge conditions. This means that one wants the system to be able to select different methods based on that knowledge in order to exhibit different reasoning traces. The TMK architecture allows for this kind of variation. The goal from a modeling perspective is to correctly specify the knowledge structures, reasoning strategies, and ordering of strategies, such that for any initial knowledge condition the TMK model is able to accurately account for the reasoning.

*The SBF Language:* As an initial attempt to address the issues from the Maxwell case and the Clement protocols, we attempted to model the Clement protocols using a computational theory of device design called "adaptive modeling" (Goel 1991b, 1996). This attempt led to the development of new design considerations and ultimately to a new computational theory of scientific problem solving. The theory of "adaptive modeling" takes its name from the perspective it adopts on conceptual device design. Conceptual design generally refers to the preliminary phase of the design process. The problem-solving task in this phase takes a specification of the functions of the desired device as input. It has the goal of giving a high-level specification of a structure for the device as output, where the structure can deliver the desired functions.

Kritik and IDeAL are operational knowledge systems that instantiate the theory of adaptive modeling, enable experiments with it, and provide well-defined AI languages. Built in the late eighties, Kritik integrated case-based and model-based reasoning for modeling evolutionary design of simple physical devices (Goel 1989, 1991a, 1992; Goel & Chandrasekaran 1989, 1992). The specific hypothesis in the Kritik experiments was that since the design task is a function  $\rightarrow$  structure mapping, the inverse structure  $\rightarrow$  function map of old designs may guide the adaptation of an old design to achieve a new functional specification. The structure  $\rightarrow$  function map of a device design in Kritik is specified as a Structure  $\rightarrow$  Behavior  $\rightarrow$  Function model. In an SBF model of a device the behavior mediates between function and structure: it captures teleological and compositional knowledge of a device, and provides a

functional and causal explanation of the how the structure of the device delivers its functions.

The IDeAL system builds upon the Kritik system in several significant ways. Perhaps the most significant contribution of the IDeAL system is the addition of a theory for cross-domain analogy called model-based analogy (MBA) (Bhatta 1995, Bhatta & Goel 1993, Bhatta & Goel 1997). This theory enables the system to apply abstract information that it learns in one domain such as that of electric circuits to another domain such as that of heat exchangers.

Kritik and IDeAL both focus on the task of conceptual design. In design the goal is the description of some artifact that serves a particular purpose, i.e., it has some desired function. For this reason the Kritik and IDeAL systems focus on functionally driven transformation processes. The task in scientific discovery is often one in which changes to the function are only realized after a structural change has taken place. The ToRQUE systems make use of this kind of transformation – form-based or “function-follows-form” transformations (see Griffith *et al* 1997, 1999).

In this research we have identified a series of transformational knowledge patterns that can be used to accomplish form-based transformations. We have called these patterns generic structural transformations (GSTs) because they are generic with respect to the models to which they may be applied and because they are first applied to the structure of the model and then propagated to the behavior. We have described two strategies for carrying out form-based transformations. The first is called Structure-Based Model Transformation (SBMT) and the second is called Limiting Case Analysis (LCA).

### **Function-Follows-Form Reasoning Packets**

One important task in artificial intelligence is identifying patterns of reasoning that are generic to a variety of problems. In this research we have identified several reasoning patterns using the TMK language. These reasoning patterns are packets of tasks, methods, and knowledge that frequently appear together. The most promising of these TMK reasoning packets is the FFF packet, which appears to be a general process used by expert reasoners to solve exploratory problems.

One important issue in both the historical and protocol studies is to find the function of a particular physical system given its form. For example, in the S2 case the task is to find the amount the spring will stretch given the diameter. Thus far we have developed a computational system, ToRQUE2, that models S2's discovery of torque in springs. A key computational characteristic of ToRQUE2 is its application of structural transformations to the structural and topological elements of SBF models to generate new models. To achieve FFF transformations ToRQUE2 uses the GST knowledge structures. After retrieving an initial source analog via model-based analogy, ToRQUE2 evaluates the model by attempting to reduce the differences between the

target and the analog model. This evaluation process involves retrieving generic models of physical principles (GPPs) which can explain away the differences or applying GSTs to transform the target or source models. These adaptations bring new knowledge to the task that may lead ToRQUE2 towards or away from the initial goal. ToRQUE2 discovers the GPP of torque while attempting to reduce the differences between a circular and an imaginary square coil.

*S2 protocol: line 121:* Now that's interesting...Just looking at this it occurs to me that when force is applied here [end segment], you not only get a bend on this segment, but because there's a pivot here [referring to a connection in the hexagonal coil], you get a torsion effect.. around there..[a center segment]

Through ToRQUE2 we have established that in the S2 case “function-follows-form” transformations play a significant role in the exploratory process. We hypothesize that “function-follows-form” transformations also play a significant role in Maxwell's exploration of electromagnetism.

In the following sections we present the function-follows-form reasoning packet by showing the task pattern, method pattern, and knowledge patterns that are used to carry out the reasoning, which taken together form a reasoning packet.

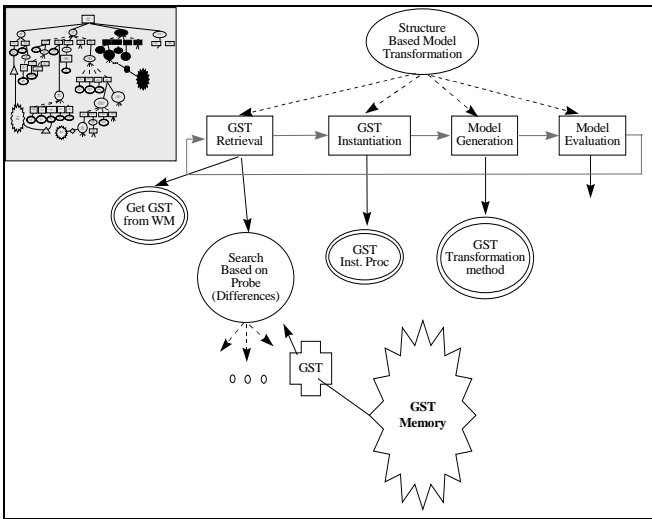
### **Function-Follows-Form Task Pattern**

In (Griffith 1999) we hypothesize that the ordering of high-level reasoning strategies proceeds from a strategy of model-based search through a process of analogy and, failing that, to processes of transformation. We are also hypothesizing that the FFF reasoning packet is used only under certain conditions. The conditions under which a method takes place is a part of its task pattern. The task pattern for FFF can be defined formally with respect to the models in memory, the target model, problem description, and the solution.

The formal task pattern for FFF is: given (1) a target model that is an element of a set of models available to the agent, (2) some problem with respect to that target model, and (3) that no solution can be generated using a search method or an analogy method, return a new model that contains a solution to the problem, such that using analogical transfer from this new model back to the target model provides a solution to the problem. The task pattern defines the problem to consist of input (the problem and the target model) and output (the solution to the problem). It also defines the situation in which the task is performed – in this case, after attempting model-based search and model-based analogy. We see this task pattern in several of the subjects including S2, S6, and S8. We also see this pattern in Maxwell's reasoning.

## Function-Follows-Form Method Pattern

The method pattern for the FFF transformation is found in the SBMT hierarchy. The method pattern in a TMK reasoning packet includes the hierarchy of subtasks that the method spawns, the ordering of these subtasks, and the knowledge that the method acquires during its processing. These aspects of the method are shown in Figure 1. The dashed lines indicate the subtasks that are spawned and the solid black lines indicate that a method or procedure is selected. The gray lines show the ordering of subtasks. The rectangles are subtasks in the method hierarchy. The single-line ellipses are methods, and the double-line ellipses are procedures. Memory is indicated by the star or seal



**Figure 1:** Function-Follows-Form Method Pattern

figure and the plus sign is a specific piece of knowledge that gets retrieved from memory.

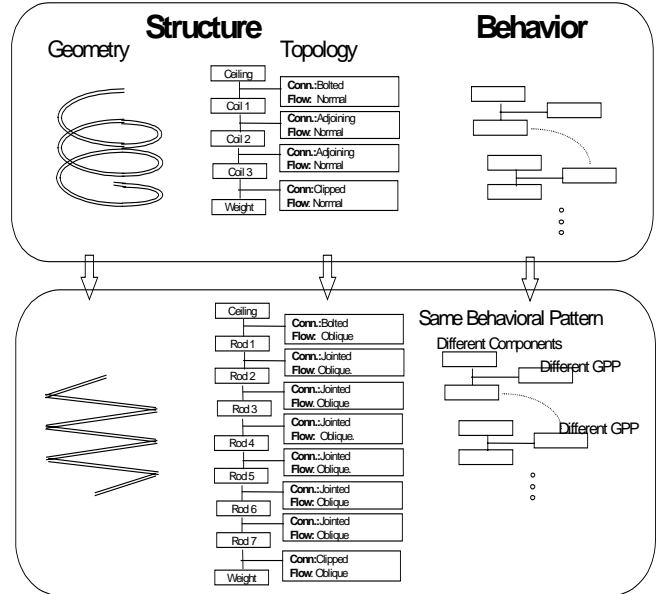
The pattern shows that SBMT is a memory-based process in which the retrieval of particular GSTs occurs when no GSTs are available in working memory. The process then instantiates the retrieved GST for the particular problem-solving situation, and attempts to generate a new model by applying the GST to the model. The process ends with an evaluation of the model that could result in the recognition of GPPs or in recursive application of the SBMT method.

## Function-Follows-Form Knowledge Patterns

The knowledge patterns for the FFF reasoning packet include SBF models as well as GSTs. GSTs are the active knowledge element in FFF. GSTs contain indexing information that allows them to be retrieved based on differences between analog models and information that indicates when they can be applied to a model. Most importantly they contain the processing information for transforming the structure (including geometry and topology) as well as the behavior of SBF models.

In Figure 2 we see one application of the function-follows-form reasoning packet. This reasoning packet

shows how the 3D-to-2D GST is applied to the spring model. First, the geometry of the spring is transformed, which results in changes to the structure. The changes lead to behavior changes in the spring. Each new component's behavior is consolidated to form a new behavioral pattern.



**Figure 2:** Function-Follows-Form in S2 Protocol

In Figure 3 we see how the function-follows-form reasoning packet is applied to our Maxwell's model construction. The top figures represent one stage in Maxwell's reasoning about the electromagnetic aether. He envisioned the aether as composed of a group of fluid vortices. The rotating circles is our representation of a cross section of a set of vortices packed together in the aether. The bottom figure shows Maxwell's representation of his model of the aether after the application of a function-follows-form transformation that changes the structure by adding "idle wheel particles" to solve the problem of friction between the vortices. The structural changes in the form of the aether result in behavioral changes. It is this model of the aether that Maxwell uses to construct the equations for electromagnetic interactions.

The significant point here is that the FFF reasoning packet was first discovered with respect to the Clement protocols, and then identified as potentially significant in interpreting Maxwell's case.

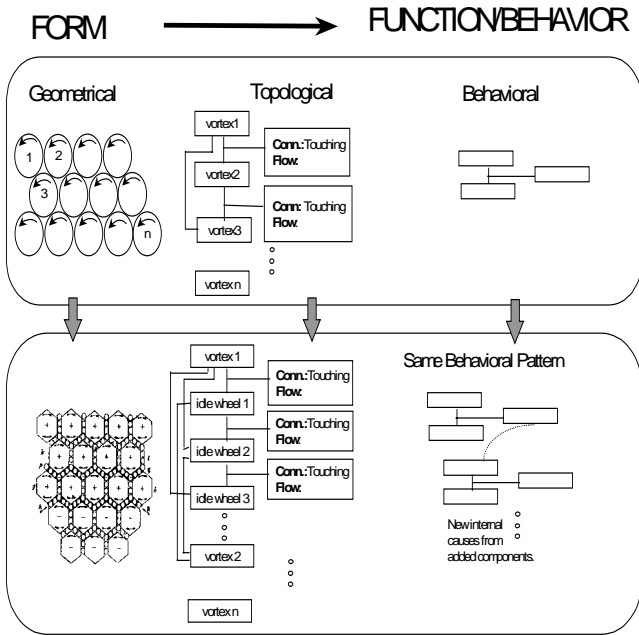
## The ToRQUE System

In this section we show how the computational model instantiated in ToRQUE2 captures the salient reasoning processes of subjects by presenting a walkthrough of the steps taken by the ToRQUE2 system when configured with our interpretation of S2's initial knowledge state.

The primary task of the ToRQUE2 system is to solve a problem. The problem in this situation can be characterized as finding a relationship between a structural concept (Cs)

(e.g. diameter) in the model and a behavioral concept (Cb) (e.g. amount of stretch).

To achieve exploration in TMK requires a working memory of target models (WMT), analogs (WMA), and

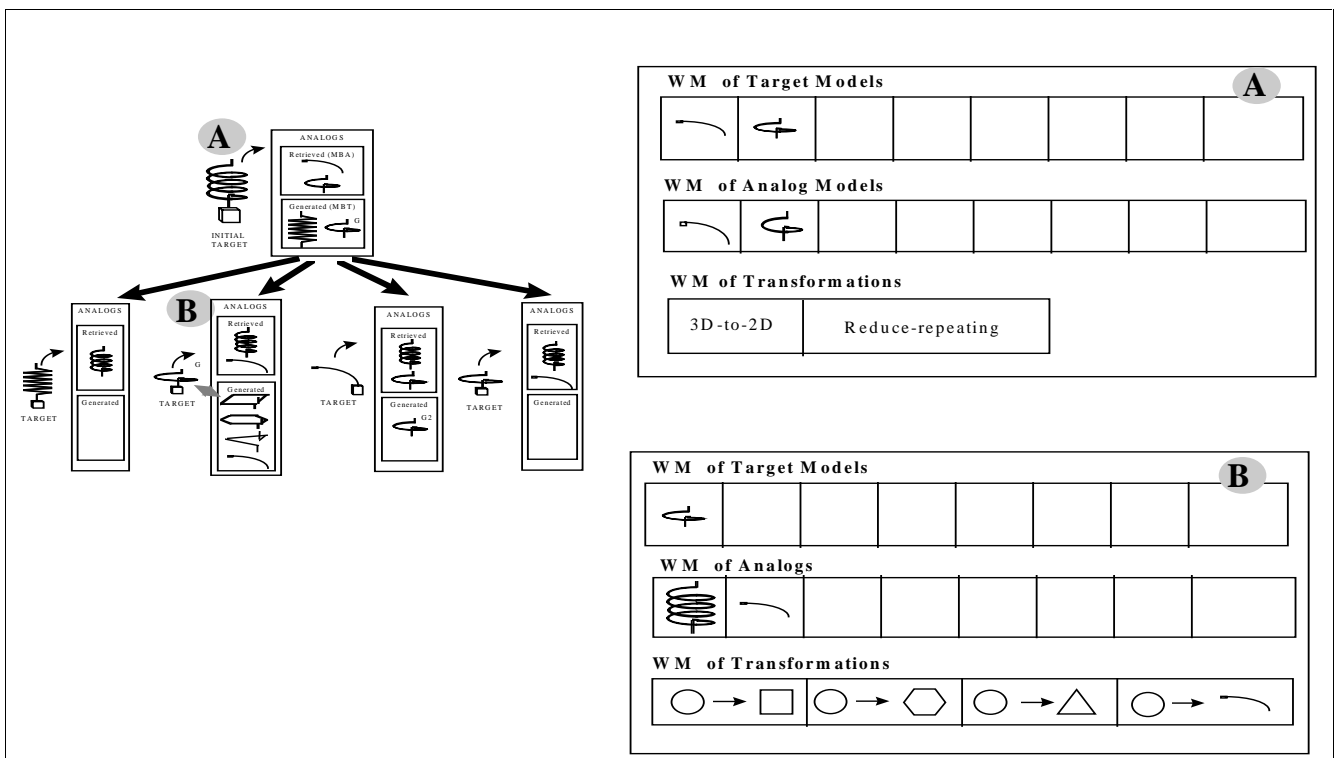


**Figure 3:** Hypothesized Function-Follows-Form in Maxwell

GSTs (WMGST). As an agent addresses its task they may come to a point where they do not know how to proceed. Past reasoning stored in working memory allows the agent to pick a GST that is related to the reasoning at hand or to reasoning that has occurred recently. This serves to

constrain the randomness of the selection of a GST. In ToRQUE2 working memory is captured in a data structure which has a last-in-first-out (LIFO) structure. Figure 4 shows two snapshots of working memory structures. The snapshot labeled (A) shows the WM during the first model-based analogy process prior to attempting any transformations. Snapshot B shows what transformations are placed on the structure when the circular coil becomes the target model. The transformation structure between A and B is the transformations performed between these snapshots. All the transformations that are retrieved are ordered and placed onto this WMGST structure. Thus one can think of this structure as using the last transformation which the agent was thinking about but did not apply. Not all transformations can be used on all models so many transformations may be rejected prior to being applied, e.g., a circle-to-square transformation is only possible if the target model is circular. Also, previously explored target models are removed from the structure such as when a coil retrieves a spring as an analog.

The exploration process proceeds through the interaction of Model-Based Analogy (MBA) and Structure-Based Model Transformation (SBMT) with the working memory structures, WMA and WMGST. MBA retrieves a set of analog models to solve the particular problem one of which is selected and the rest of which are placed in WMA. The answer that is produced from these analogs is evaluated by attempting to reduce the differences between it and the target model. One method of reducing these differences is to apply SBMT to the source or target analogs. Similarly GSTs are indexed and retrieved by these differences and one GST is applied while the remaining are placed in the



**Figure 4:** Snapshot of Working Memory Structures at Two Places in the Program State

WMGST structure. As reasoning progresses, a collection of transformations are placed into WM. In this way WM is not being used as a repository for knowledge that is currently being addressed, but as a repository for knowledge which has been retrieved but which has not been considered.

The left portion of Figure 4 depicts the models that are initially retrieved using the spring model and the models that are generated after transforming the initial target model in various ways. After the system retrieves an analog model it then evaluates that model by attempting to reduce the differences between itself and the analog model. These differences are used as indexes into a memory of generic structural transformations (GSTs). The SBMT process then applies the retrieved transformations to the target model to generate additional models. Notice that the models that are generated may be similar to retrieved analog models. These models, however, are not identical and so we have marked the generated coil model with a G. The figure shows the models that are retrieved as analogs for the spring model. These models were retrieved as functional analogs to the spring because they each supply a restorative force. Generic models such as GPPs are knowledge abstractions that can reduce the differences between two models by recognizing that the features of the analog model are also present in the target model.

One significant outcome of the ToRQUE2 experiments is that ToRQUE2 is able to model the competences exhibited by the test subjects (S6-S11) to a surprising degree of accuracy without changing anything except for the starting knowledge conditions. This means that the system could model the test subjects:

- ◆ without additional knowledge structures,
- ◆ without additional reasoning strategies,
- ◆ without altering the control architecture, and
- ◆ without altering the ordering of the strategies

Altering the starting knowledge conditions includes one or more of the following:

- ◆ removal of knowledge structure,
- ◆ removal of reasoning strategy, or
- ◆ removal of an index to a knowledge structure.

This means that the ToRQUE2 system is a representative instantiation of a general competence model for the spring problem. This means our model covers a representative subset of the possible knowledge and strategies one might use to solve the spring problem, such that it can account for both paths to the solution and paths to failure by the scientists. This lends support to our claim that “function-follows-form” transformations enable S2’s conceptual change, because it is one configuration of a representative problem-solving model.

### Acknowledgements

We would like to thank John Clement for generously supplying us with all of the “think-out-loud” protocols.

### References

- Bhatta, S.R. & Goel, A.K. (1993) Learning Generic Mechanisms from Experiences for Analogical Reasoning. In *Proc. Fifteenth Annual Conference of the Cognitive Science Society*, Boulder, Colorado, July 1993, Hillsdale, NJ: Lawrence Erlbaum.
- Bhatta, S. R. & Goel, A. K. (1994) Model-Based Discovery of Physical Principles from Design Experiences. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, Special Issue on Machine Learning in Design, 8(2), May 1994.
- Clement, J. (1989) Learning via Model Construction and Criticism: Protocol Evidence on Sources of Creativity in Science, In *Handbook of Creativity: Assessment, Theory and Research*, Glover, G., Ronning, R., & Reynolds, C. (Eds.), chapter 20, New York, NY:
- Goel, A., K. (1989) Ph.D. Dissertation, *Integrating Case-Based Reasoning and Model-Based Reasoning for Adaptive Design Problem Solving*, Department of Computer and Information Science, The Ohio State University.
- Goel, A., K. (1991a) A model-based approach to case adaptation. In *Proc. of the Thirteenth Annual Conference of the Cognitive Science Society*, Chicago.
- Goel, A., K. (1991b) Model revision: A theory of incremental model learning. In *Proc. of the Eighth International Conference on Machine Learning*, Chicago.
- Goel, A.K. & Chandrasekaran, B. (1992) Case-Based Design: A Task Analysis. In C. Tong and D. Sriram (eds) *Artificial Intelligence Approaches to Engineering Design, Volume II: Innovative Design* (San Diego: Academic Press).
- Griffith, T.W. (1999) A Computational Theory of Generative Modeling in Scientific Reasoning, Atlanta, Georgia Institute of Technology, Ph.D. Dissertation, GIT-COGSC199.
- Griffith, T.W., Nersessian, N.J., & Goel, A.K. (1996) The Role of Generic Models in Conceptual Change, *Proceedings of the Cognitive Science Society*, 18 (Hillsdale, NJ: Erlbaum).
- Griffith, T.W., Nersessian, N.J., Goel, A.K., and Clement, J. (1997). Exploratory Problem-Solving in Scientific Reasoning. *The Nineteenth Annual Conference of the Cognitive Science Society*, Stanford University, Lawrence Erlbaum
- Nersessian, N.J. (1992) How Do Scientists Think? Capturing the Dynamics of Conceptual Change in Science, In *Cognitive Models of Science*, ed. R.N. Giere. Minneapolis, MN: University of Minnesota Press.
- Nersessian, N.J. (1995) Should Physicists Preach What They Practice? Constructive Modeling in Creating Scientific Understanding, *Science & Education*, vol. 4.
- Nersessian, N.J. & Greeno, J. (1992) Constructive Modeling in Scientific Discovery (unpublished manuscript).
- Punch, W.F., Goel, A. K., & Brown, D.C. (1995). A Knowledge-Based Selection Mechanism for Control with Application in Design, Assembly and Planning. *International Journal of Artificial Intelligence Tools*.
- Stroulia, E. & Goel, A.K. (1992) Generic Teleological Mechanisms and their Use in Case Adaptation, In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (Hillsdale, NJ: Lawrence Erlbaum)
- Stroulia, E. & Goel, A.K. (1995) Functional Representation and Reasoning in Reflective Systems. *Journal of Applied Intelligence, Special Issue on Functional Reasoning*, 9(1).



# Teacakes, Trains, Taxicabs and Toxins: A Bayesian Account of Predicting the Future

Thomas L. Griffiths & Joshua B. Tenenbaum  
Department of Psychology  
Stanford University  
Stanford, CA 94305-2130 USA  
{gruffydd,jbt}@psych.stanford.edu

## Abstract

This paper explores how people make predictions about the future. Statistical approaches to predicting the future are discussed, focussing on the method for predicting the future suggested by J. R. Gott (1993). A generalized Bayesian form of Gott's method is presented, and a specific psychological model suggested. Three experiments show that the predictions people make about the future are consistent with a Bayesian approach.

Despite the difficulty of predicting the future, people happily do it every day. We are confident about being able to predict the durations of events, how much time we will need to get home after work, and how long it will take to finish the shopping. In many cases we have a great deal of information guiding our judgments. However, sometimes we have to make predictions based upon much less evidence. When faced with new situations our decisions about how much longer we can expect events to last are based on whatever evidence is available. When the only information we possess concerns how long a particular event has lasted until now, predicting the future becomes a task of induction.

In this paper we explore the question of how people predict the future when told only about the past. We examine a simple statistical method of predicting the future, and consider how such a method could be made sufficiently flexible to be useful in everyday situations. The resulting Bayesian model makes strong predictions about the effects of providing further information, the symmetry of this form of reasoning, and how it should be affected by prior knowledge. We test these predictions empirically.

## The Copernican Anthropic Principle

A simple solution to the problem of predicting the future was recently proposed by the cosmologist J. Richard Gott III (1993). Gott's method is founded upon what he calls the "Copernican anthropic principle", which holds that

... the location of your birth in space and time in the Universe is privileged (or special) only to the extent implied by the fact that you are an intelligent observer, that your location among intelligent observers is not special but rather picked at random (1993, p. 316)

Gott extends this principle to reasoning about our position in time – given no evidence to the contrary, we should not assume that we are in a "special" place in time. This means that the time at which an observer encounters a phenomenon should be randomly located in the total duration of that phenomenon.

Denoting the time since the start of a phenomenon  $t_{past}$ , and its total duration  $t_{total}$ , Gott forms what he terms the "delta  $t$  argument". Define the ratio

$$r = \frac{t_{past}}{t_{total}} \quad (1)$$

and assume that this is a random number between 0 and 1. It is possible to form probabilistic predictions about the value of  $r$ . For example,  $r$  will be between 0.025 and 0.975 with a probability  $P = 0.95$ , meaning that

$$\frac{1}{39}t_{past} < t_{future} < 39t_{past} \quad (2)$$

with 95% confidence, where  $t_{future} = t_{total} - t_{past}$ . Similarly,  $r$  will be less than 0.5 with probability  $P = 0.5$ , so  $t_{past} < t_{future}$  with 50% confidence.

This method of reasoning has been used to predict a wide range of phenomena. Gott (1993) tells of his visit to the Berlin Wall in 1969 ( $t_{past} = 8$  years). Assuming that his visit was randomly located in the period of the wall's existence, the 95% confidence interval for  $t_{future}$  would be 2.46 months to 312 years. The wall fell 20 years later, consistent with these predictions. Gott made similar calculations of  $t_{future}$  for Stonehenge, the journal *Nature*, the U.S.S.R., and even the human race. Subsequent targets of the principle have included Broadway musicals and the Conservative government in Britain (Landsberg, Dewynne, & Please, 1993).

## What's Bayes got to do with it?

Gott's (1993) method for predicting the future yields interesting predictions in a wide range of situations. It is simple, but could prove useful in forming effective plans and expectations about future events. On this basis, it would be plausible for people to apply similar principles when making judgments concerning time.

Despite the attractiveness of this claim, there may be good reasons why Gott's (1993) method would not belong in our cognitive armory. One reason could be the restrictive assumptions of such an inference. In many

cases in the real world where it might be desirable to predict the future, we know more than simply how long a process has been underway. In particular, our interaction with the world often gives us some prior expectations about the duration of an event. For example, meeting a 78 year-old man on the street, we are unlikely to think that there is a 50% chance that he will be alive at the age of 156.

Prior knowledge is not the only kind of information that Gott's (1993) method neglects. In some cases, our predictions are facilitated by the availability of multiple encounters with a phenomenon. For example, if we were attempting to determine the period that passes between subway trains arriving at a station, we would probably have several trips upon which to base our judgment. If on our first trip we discovered that a train had left the station 103 seconds ago, we might assume that trains run every few minutes. But, after three trips yield trains that have left 103, 34, and 72 seconds ago, this estimate might get closer to 103 seconds. And after ten trains, all leaving less than 103 seconds before we arrive, we might be inclined to accept a value very close to 103 seconds.

These limitations suggest that Gott's (1993) formalization lacks the flexibility that would be required of a method for predicting the future in real world situations. Such a method must allow for the influence of prior knowledge, and reflect the effect of multiple examples. Bayesian inference may provide a means of satisfying both of these requirements. Bayes Theorem states that

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)} \quad (3)$$

where  $h$  is some hypothesis under consideration, and  $d$  is the observed data. By convention,  $P(h|d)$  is referred to as the posterior probability of the hypothesis,  $P(h)$  the prior probability, and  $P(d|h)$  the likelihood of the data given that hypothesis. In the case where the hypotheses are continuous,  $P(d)$  can be obtained by summing across  $P(d|h)P(h)$  for all hypotheses, giving

$$P(h_i|d) = \frac{P(d|h_i)P(h_i)}{\int_{h \in H} P(d|h)P(h)dh} \quad (4)$$

where  $H$  is the set of all hypotheses.

Conveniently, some work extending Gott's (1993) method into a Bayesian framework already exists. In responding to a criticism offered by Buch (1994), Gott (1994) noted that his method for predicting the future could be expressed in Bayesian terms. Using the prior  $P(t_{total}) \propto \frac{1}{t_{total}}$  and the likelihood  $P(t_{past}|t_{total}) = \frac{1}{t_{total}}$  yields the same results as his original formulation of the delta  $t$  argument.

These values are not chosen arbitrarily. The use of  $\frac{1}{t_{total}}$  for the prior is motivated by sound statistical principles (Press, 1989), and provides a scale-invariant method for distributing probability over hypotheses. In many applications it is referred to as the uninformative prior, as it is appropriate when an inference is guided by no specific prior knowledge. The use of  $\frac{1}{t_{total}}$  for the likelihood is also well motivated. The anthropic principle

essentially states that  $t_{past}$  should be viewed as having been randomly sampled from  $t_{total}$ . Assuming a uniform distribution across values of  $t_{past}$ , the probability of any particular  $t_{past}$  will be  $\frac{1}{t_{total}}$ .

Crucially, both the priors and the likelihoods of this Bayesian framework can be modified to suit the situation at hand. A simple, flexible set of priors is provided by the Erlang distribution

$$P(t_{total}) = \frac{t_{total}e^{-t_{total}/\beta}}{2\beta^2} \quad (5)$$

where  $\beta$  is a free parameter. This distribution has a broad peak at  $t_{total} = \beta$ , and decays to zero at 0 and  $\infty$ . This parameterized peaked distribution provides a simple means to summarize many of the kinds of distributions that might be encountered across temporal domains. Similarly, the effect of multiple examples of  $t_{past}$  can be introduced by modifying the likelihoods. Extending the anthropic principle, we can assume that each example is drawn independently from  $t_{total}$ . The probability of observing a set of  $n$  examples will then be  $(\frac{1}{t_{total}})^n$ .

## A Bayesian model

The specification of a Bayesian model requires identifying the distributions governing the prior probabilities and the likelihoods. For the case of predicting the future, Gott's (1993) method provides a good starting point. For each model considered in this section, we will assume that people's responses reflect the point  $t$ , such that  $P(t < t_{total}) = 0.5$ . This is essentially assuming that people adopt an unbiased criterion in making their judgments. Furthermore, we will examine the predictions of each model when provided with one, three, or ten examples of  $t_{past}$  ( $n = 1, 3, 10$  respectively).

The simplest case of Gott's (1993) method is just the delta  $t$  argument, as presented in Equation 1. The predictions of the delta  $t$  argument are shown in the leftmost panel of Figure 1(a). The predictions are unaffected by  $n$ , and are thus constant at  $t = 2t_{past}$ . This seems to defy intuition, and is the weakest of the models we will consider.

At the next level of complexity is the introduction of  $(\frac{1}{t_{total}})^n$  for the likelihood of a set of  $n$  examples and the uninformative prior  $P(t_{total}) \propto \frac{1}{t_{total}}$ , yielding the closed form prediction  $t = 2^{1/n}t_{past}$ . As shown in the second panel of Figure 1(a), the model shows an effect of the number of examples. The main problem with this model is that the prior does not make use of the flexibility provided by the inclusion of prior knowledge in inference. In particular, the uninformative prior makes scale invariant predictions about generalization, which means that  $t_{future}$  will be a constant proportion of  $t_{past}$ , whether predicting the future of the human race or a 78-year old man.

Substituting the Erlang distribution for the uninformative prior renders Equation 4 into

$$P(t_{total}|T) = \frac{(t_{total})^{1-n}e^{-t_{total}/\beta}}{\int_{t_{past}}^{\infty} (t_{total})^{1-n}e^{-t_{total}/\beta} dt_{total}} \quad (6)$$

where  $T$  is the set of  $n$  examples of  $t_{past}$ . The third panel of Figure 1(a) shows the predictions of this model for  $n = 1, 3, 10$  with  $\beta$  ranging from 0.5 to 4.5 in unit increments. The model shows reduced predictions with more examples, and increased predictions with larger values of  $\beta$ .

This third model is what will be tested against people’s predictions about the future. The use of a parameterized prior and specification of the likelihoods makes it appear somewhat more complex than Gott’s (1993) original prescription. As parsimony contributed to the psychological plausibility of the approach, this additional complexity needs to be justified. One source of justification is the success of the Bayesian framework in describing behavior in other domains. Tenenbaum (1999) discussed several examples of tasks that map naturally onto the temporal problems addressed by Gott.

The tasks considered by Tenenbaum (1999) were cases of inductive concept learning, where people learn about a concept through the provision of positive examples. An example of this kind of task would be predicting healthy levels of imaginary toxins. People would be given a number that they are told is a healthy level of a particular toxin, then asked to guess the highest level of the toxin that would be considered healthy. This situation is exactly analogous to predicting the total duration of an event from the amount of elapsed time since the start of the event. In both cases, a person is given a number that is assumed to be randomly sampled from the set of all numbers satisfying a particular criterion, and asked to judge the nature of this criterion. Since both duration and toxin levels are numbers required to be between 0 and some maximum number, this judgement requires the estimation of the maximum number ( $t_{total}$  in the case of predicting the future). Tenenbaum (1999) found that a Bayesian framework gave a good account of people’s performance on this kind of task.

## Model predictions

The Bayesian model outlined above, and depicted in the third panel of Figure 1(a), has some obvious implications. Most central is how the provision of further information should affect predictions. The tightening of the range of acceptable values of  $t_{total}$  corresponds to an important component of Tenenbaum’s (1999) account of concept learning. As people are given more examples of a concept, they become less inclined to generalize beyond the properties of those examples. In the case of predicting the future, where all hypotheses have a value of  $t_{total}$  as their sole property, this manifests as a tendency to accept the smallest possible value of  $t_{total}$  that includes all observed values of  $t_{past}$ .

The Bayesian model also clearly defines the phenomena of prediction to be symmetric. In forming a judgment about  $t_{total}$ , knowledge of  $t_{past}$  and knowledge of  $t_{future}$  are equally informative. Given one of these pieces of information, it is possible to calculate a range of acceptable values for  $t_{total}$ . If people apply similar methods in making judgments about time the effects should be maintained regardless of which of  $t_{past}$  and  $t_{future}$  are

provided. This is not an absolute symmetry, however: if scenarios in which  $t_{future}$  and  $t_{past}$  are provided differ in the distribution of prior probability, then the predicted values of  $t_{total}$  may also differ.

One further implication of the model is that manipulating the prior probability distribution across the hypothesis space will produce a general change in predictions, at least until the effect of the priors is overwhelmed by the likelihoods. In particular, inducing a prior preference for a relatively high value will bias inferences towards hypotheses around that value. If people employ approximately Bayesian methods in forming their judgments, introducing information that biases the prior probability distribution in this way should result in higher predictions, especially when those predictions are based upon few observations.

The Bayesian framework for predicting the future has three clear implications for the kind of judgments that it will produce. The effects of further information, symmetry of predictions, and effects of prior probabilities are all important properties of how these judgments are made. Experiments 1, 2, and 3 examine these predictions in turn.

## Experiment 1: New information

### Method

**Participants** Participants were 81 undergraduates from Stanford University, participating for partial course credit. The participants were randomly assigned to four groups.

**Materials** Four simple scenarios were developed for exploring the predictions of the Bayesian framework. The first scenario described a coffee shop that had recently started selling teacakes. This scenario is given below. Participants were shown  $t_{past}$ , and asked to predict  $t_{total}$ . The second scenario told participants that they were visiting a foreign country in which trains ran precisely to schedule. The schedule was set up so that exactly the same amount of time passed between successive trains. On the platform was a clock showing how long it had been since the last train arrived. Participants were told the value on the clock when they reached the station,  $t_{future}$ , and asked to predict  $t_{total}$ .

These scenarios were compared with two analogous situations that made no reference to time. One of the comparison scenarios was the healthy levels of toxin experiment described above. The second was a version of the Jeffreys (1961) tramcar problem: participants were told the serial number of a taxicab (as well as being given the information that all cabs are given a unique number between 1 and the total number of cabs in the company) and asked to guess the number of cabs in the company.

Each scenario had three sections. The first section outlined the situation and give a single number on which judgments were to be based. The second and third sections added further information, giving a total of three numbers and ten numbers respectively. The first number given was the largest, meaning that further observations would only tighten the range of generalization. The sets

of numbers given were identical for the teacake and toxin scenarios and the train and taxicab scenarios, and were approximately uniformly distributed. The largest example was 34 minutes (ng/mL) for the teacake (toxin) scenario, and 103 seconds (cabs) for the train (taxicab) scenario.

For example, the first section of the teacake scenario was

Each day, on your way to class, you walk past a coffee shop. The shop has recently started a new advertising campaign: they bake fresh teacakes regularly throughout the day, and have a clock outside that shows how long it has been since the teacakes were taken out of the oven. You are interested in buying a teacake as soon as it is removed from the oven, and wonder how often batches of teacakes are baked. Today, the clock shows that it has been 34 minutes since the last batch of teacakes was removed from the oven.

Please write down your best guess of how much time elapses between batches of teacakes, in minutes. Try to make a guess, even if you feel like you don't have enough information to make a decision - just go with your gut feeling. You may assume that the batches of teacakes are always separated by the same amount of time.

The second section gave the additional times of 21 and 8 minutes, and the third section gave further times of 18, 2, 5, 27, 22, 10 and 14 minutes.

**Procedure** The procedure used in all three experiments was identical: Each participant received a sheet providing general instructions about the task, and a questionnaire of one of four kinds.

## Results and Discussion

Plausible responses to these problems are constrained to be greater than the largest example provided. Responses were transformed such that  $t = \frac{x}{t_{past}}$ , where  $x$  is the raw score and  $t_{past}$  is the largest example, and participants who gave  $t < 1$  were excluded from the analysis, eliminating approximately 15% of the participants in each experiment. Responses more than three standard deviations from the mean were considered outliers, and were also excluded. Only one outlier was identified in the course of all three experiments.

A one-way within-subjects ANOVA showed a statistically significant effect of the number of examples for each scenario ( $F(2, 30) = 9.71$ ,  $F(2, 32) = 18.00$ ,  $F(2, 44) = 9.57$ ,  $F(2, 30) = 15.05$ , for the teacake, train, taxicab, and toxin respectively, all  $p < .001$ ). Means and standard errors are shown in Figure 1(b), which demonstrate that the two temporal tasks show a similar effect of new information to the tasks of analogous statistical structure. The Figure also shows predictions generated by the Bayesian model, using the Erlang prior with  $\beta$  set independently for each scenario. The parameterization of the distribution reflects the different priors that might exist across different scenarios, relative to the scale of the

examples selected. The values of  $\beta$  for the teacake, train, toxin and taxicab scenarios were 1.6, 5.4, 0.7 and 4.4 respectively. The peak of the Erlang prior is at  $t_{total} = \beta$ , yielding values of 54 minutes between batches of teacakes, 9 minutes 21 seconds between trains, 24 ng/mL of toxin, and 460 taxicabs, all of which seem appropriate.

## Experiment 2: Symmetry of effects

### Method

**Participants** Participants were another 77 undergraduates from Stanford University, participating for partial course credit. The participants were randomly assigned to four groups.

**Materials** Again, four scenarios were used. The teacake and train scenarios from Experiment 1 were applied to a different set of participants, and modified versions of these scenarios were generated to make it possible to test the symmetry of predictions. The new scenarios were identical to the original teacake and train scenarios, except for the provision of  $t_{future}$  instead of  $t_{past}$  - the participants were informed how long it would be before the next batch of teacakes were removed from the oven, or the next train would arrive, and were asked to predict the period that went between these events. All scenarios asked for predictions with 1, 3, and 10 examples, replicating Experiment 1.

### Results and Discussion

Responses were screened using the same procedure as in Experiment 1. Each pair of putatively symmetric scenarios was subjected to a two-way within-between ANOVA, examining the effects of number of examples and temporal direction. The train scenario showed a statistically significant effect of number of examples ( $F(2, 56) = 17.40$ ,  $p < .001$ ), and no evidence of an effect of temporal direction ( $F(1, 28) = 0.50$ ,  $p = 0.49$ ) or an interaction between the factors ( $F(2, 56) = 0.08$ ,  $p = 0.93$ ). The effect of new information reproduces that of Experiment 1, and the non-significant result for temporal direction implies that any asymmetry in prediction is too weak to be detected by the present experiment.

The teacake scenario likewise showed a statistically significant effect of number of examples ( $F(2, 60) = 22.18$ ,  $p < .001$ ). However, the results also indicated a significant effect of temporal direction ( $F(1, 30) = 4.46$ ,  $p < .05$ ) and an interaction between the two factors ( $F(2, 60) = 3.667$ ,  $p < .05$ ). This difference between the two scenarios may be a result of the way that the asymmetry introduces new information about the teacakes. Changing "the clock shows that it has been 34 minutes since the last batch of teacakes was removed from the oven" to "the clock shows that it will be 34 minutes until the next batch of teacakes is removed from the oven" provides the implication that the next batch of teacakes is currently in the oven. The time between batches of teacakes can be divided into time in the oven and time left waiting. Of these, the time the teacakes spend in the oven is less flexible. Applying the anthropic principle, the observation that the teacakes are currently in the

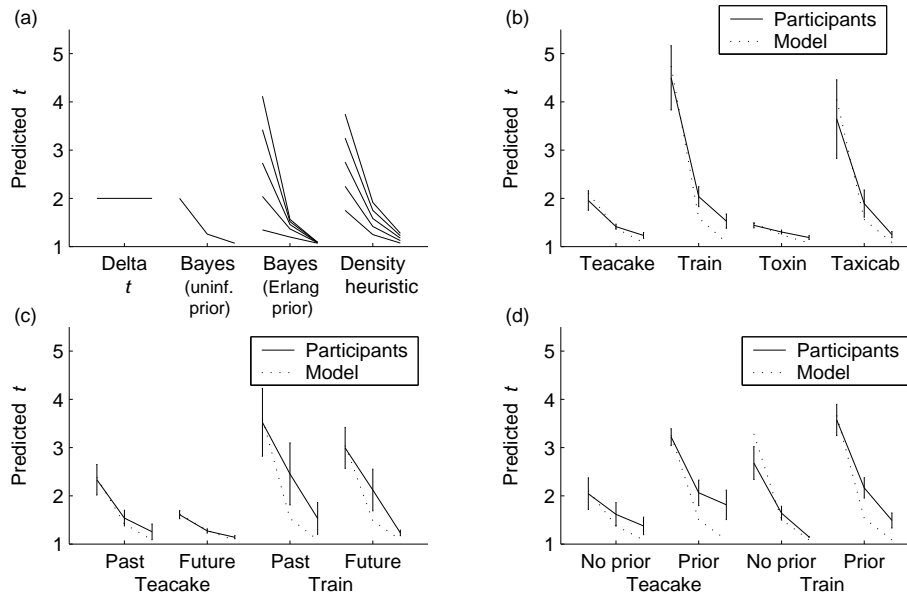


Figure 1: (a) Predictions of the various models, depicting point at which  $P(t < t_{total}) = 0.5$  for 1, 3 and 10 examples. On all graphs, the vertical axis shows the predicted value of  $t$  in proportion to  $t_{past}$ . (b) Results for Experiment 1. The solid line shows means (with one standard error). The dotted line shows the predictions of the Bayesian model with an Erlang prior. (c) Results for Experiment 2. (d) Results for Experiment 3.

oven suggests that the teacakes spend more time in the oven than waiting to be purchased. This correspondingly reduces the total amount of time that might be expected to pass between successive batches of teacakes.

Figure 1(c) shows the means and standard errors, which are reminiscent of those found in Experiment 1. The predictions of the model were made with  $\beta$  values of 2, 0.91, 3.75, and 2.95 for teacake-past, teacake-future, train-past, and train-future respectively. These values approximate those used in fitting the results of Experiment 1.

### Experiment 3: Manipulating priors

#### Method

**Participants** Participants were another 78 undergraduates from Stanford University, participating for partial course credit. The participants were randomly assigned to four groups.

**Materials** The teacake scenario from Experiment 1 and the train scenario from Experiment 2 were used, together with two new scenarios. The new scenarios gave participants information that was designed to alter their prior before they were given specific numbers upon which a prediction could be based. The sentence “A friend who you are walking with says that he worked in a coffee shop in the same chain last year, and that shops usually bake every two hours, although it varies from shop to shop” was added to the teacake scenario, and “In the course of your travels, you have noticed that most subway trains in this country run approximately every seven minutes, although it varies from place to place” to the train sce-

nario.

All scenarios asked for predictions with 1, 3, and 10 examples, replicating Experiment 1.

#### Results and Discussion

Responses were screened using the same procedure as in Experiment 1. The scenarios were grouped into teacakes and trains, and examined for the effect of number of examples and manipulating priors using two-way within-between ANOVAs. The teacake scenarios showed an effect of the number of examples ( $F = 25.86, p < .001$ ) and manipulating priors ( $F = 4.70, p < .05$ ), as well as an interaction between the two ( $F = 3.80, p < .05$ ). Similar results were shown for the train scenarios. There was a statistically significant effect of the number of examples ( $F = 50.31, p < .001$ ), as well as an effect of manipulating priors ( $F = 5.85, p < .05$ ). In both groups, the effect of the number of examples replicates the results of Experiment 1, and the higher means for the group given the raised prior is consistent with the predictions of the Bayesian model.

Means and standard errors are shown in Figure 1 (d), together with the model predictions. The  $\beta$  values used in fitting the data were 1.6, 3.3, 3.25, and 3.85 for the teacake, teacake-prior, train, and train-prior conditions respectively. The  $\beta$  values are greater in the conditions where the prior was raised, and give peak values of 1 hour, 52 minutes for the teacakes and 6 minutes 40 seconds for the trains. It is notable that these values are within 10% of the priors supplied in the experimental materials, supporting the efficacy of the manipulation and the appropriateness of the model.

## General Discussion

Predicting the future is a difficult task, particularly when the predictions are formed on the basis of very little information. Gott (1993) suggested a simple method for making predictions about the future. Gott's method lacks the flexibility to be useful in a wide range of real-world situations, but the same principles allow the construction of a more general Bayesian model. This model shows many similarities to Tenenbaum's (1999) Bayesian framework for inductive concept learning, which has proven successful in other domains. Experiments 1, 2, and 3 explored whether the Bayesian model provided a reasonable account of the effects of new information, the symmetry of predictions, and the effects of prior probabilities upon people's judgments about the future.

Experiment 1 showed that the effect of providing further examples conformed to the predictions of the Bayesian model: more examples promoted a reduction in the scope of generalization, with predictions becoming closer to the largest example provided. Experiment 2 showed that these predictions were symmetric in time, taking the same form regardless of whether the judgment concerned the past or the future. Experiment 3 showed that people's predictions could be affected by the manipulation of their prior expectations, and that this effect was consistent with the interaction of priors and likelihoods in Bayesian inference.

### Psychological plausibility

It seems unlikely that the participants in this experiment were consciously performing Bayesian inference. A more probable explanation is that the problem can be solved by the application of a simple heuristic, which more closely resembles the cognitive process by which answers can be reached (cf. Gigerenzer, 1999).

One simple heuristic that produces results consistent with both the data and the normative Bayesian model is the rule "The distance to extrapolate is the range of scores divided by the number of examples", which is an unbiased frequentist estimator of  $t_{total}$ . The predictions of this 'density' heuristic are shown in the fourth panel of Figure 1(a). Priors are implemented by taking the average of  $t_{past}$  and  $\beta$  before dividing by  $n$ , and the curves shown illustrate  $\beta$  ranging between 0.5 and 4.5 in unit increments. Note that the extent of generalization decreases with the number of examples, as in the present experiments.

The existence of such a heuristic does not affect the claim that people's predictions about temporal events are consistent with a Bayesian framework. The algorithmic properties of the Bayesian model and the heuristic may differ, but their computational properties are similar. In fact, producing the response  $t = t_{past} + \frac{t_{past}}{n}$ , where  $n$  is the number of examples, serves as a first order approximation of the Bayesian model with prior  $P(t_{total}) \propto \frac{1}{t_{total}}$ .

### Predicting future research

The present results provide support for the claim that a process consistent with Bayesian inference underlies peo-

ple's judgments on these tasks. However, the generality of the findings needs to be extended. In particular further scenarios need to be investigated. Demonstrating the consistency of the results across a range of contexts and set of numerical examples will increase the strength of the findings.

The model-fitting presented in the preceding experiments helps to show that the results are consistent with a Bayesian model, but one of the most important outcomes of the model-fitting is that the peak values of the prior distributions are in appropriate ranges. This provides a further avenue for future research: empirically estimating the parameters of the prior distribution, and using these results to predict the effects of providing examples. This would involve administering the scenarios without giving a time displayed on the clock, and asking people to estimate  $t_{total}$ . The resulting distribution will give a prior,  $P(t_{total})$ , that can be used to predict further responses.

Finally, it is interesting to note that the ability to predict the future may be important to domains other than conscious planning. For instance, Anderson (1990) argued that memory may display similar temporal sensitivities. The major challenge for any memory system is to index entries in a fashion whereby those that are needed will be readily available. This requires predicting the future: given an event, the expected future occurrence of the event must be inferred if it is to be indexed appropriately. Anderson (1990) suggests that these inferences occur unconsciously, and are an important part of the human memory. One attractive component of future research is thus exploring the extent to which unconscious temporal judgments reflect Bayesian principles.

### Acknowledgements

This research was supported in part by a grant from MERL, and a Hackett Studentship to the first author. For valuable discussions, we thank Mira Bernstein, Roger Shepard and David Somers.

### References

- Anderson, J. R. (1990). *The Adaptive Character of Thought*. Lawrence Erlbaum, Hillsdale, NJ.
- Buch, P. (1994). Future prospects discussed. *Nature*, 368:107-108.
- Gigerenzer, G. (1999). *Simple heuristics that make us smart*. Oxford University Press, Oxford.
- J. R. Gott III (1993). Implications of the Copernican principle for our future prospects. *Nature*, 363:315-319.
- J. R. Gott III (1994). Future prospects discussed. *Nature*, 368:108.
- Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford.
- Landsberg, P., Dewynne, J., and Please, C. (1993). Rise and fall. *Nature*, 365:384.
- Press, S. J. (1989). *Bayesian statistics: Principles, Models, and Applications*. Wiley, New York.
- Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In Kearns, M. S., Solla, S. A., and Cohn, D. A., editors, *Advances in Neural Information Processing Systems 11*, pages 59-65. MIT Press, Cambridge, MA.

# Vagueness in Context

Steven Gross (gross2@phil.upenn.edu)

Department of Philosophy; 433 Logan Hall; University of Pennsylvania  
Philadelphia, PA 19104 USA

## Abstract

This paper argues that whether an utterance of a vague term makes any contribution to propositional content is context-sensitive and that attention to this fact allows for an attractive solution to the sorites paradox.

## Introduction

A predicate is vague if it permits borderline cases such that it's neither clear that the predicate does apply nor that it doesn't.<sup>1</sup> What, for example, is the least number of hairs a man must have in order not to be bald? Vagueness is a pervasive feature of natural languages, but it has proven rather resistant to theoretical delineation. For any attempt to characterize the *semantics* of vague terms (what they *mean*) and their *logic* (what reasoning involving them is *valid*) must yield a compelling dissolution of the ancient sorites paradox, which is no easy matter. Here's a version of the paradox using the predicate 'bald':<sup>2</sup> Someone with no hairs is bald. But one hair can't make the difference between being bald and not being bald. (That is, for any number  $n$ , if someone with  $n$  hairs is bald, then someone with  $n+1$  hairs is bald.) So, someone with 1,000,000 hairs is bald. Well-nigh unobjectionable premises seem to lead by well-nigh unobjectionable reasoning to an absurd conclusion.<sup>3</sup> What has gone wrong?

The enormous literature logicians, philosophers, and semanticists have produced on vague language over the last few decades has

generated a plethora of competing possible positions, each with its own well-known problems.<sup>4</sup> These discussions have succeeded in shedding much light on the phenomenon but have not generated anything remotely approaching a consensus on the relevant issues. So, let us proceed once more into the breach.

In this paper, I defend an approach to vague language and the associated sorites paradox that emphasizes the context-sensitivity of vague terms. A term is *context-sensitive* if and only if its contribution to propositional content can vary across occasions of use without any change in the term's standing meaning in the language. (Indexical expressions, such as the pronoun 'I', provide standard examples.)<sup>5</sup> Previous approaches have adverted to context-sensitivity in attempting to defang the sorites,<sup>6</sup> but they have assumed that vague terms always make *some* contribution to propositional content, whereas I stress how considerations of conversational coherence can affect whether, in a given context, the use of a vague term succeeds in contributing to content at all.

My paper has three parts. I first put forward a necessary condition on the expression of a proposition and suggest that utterances of sentences containing vague predicates only sometimes satisfy it. Then, I argue that, in particular, the consideration of a sorites paradox can result in the violation of this condition. Finally, I briefly indicate some of the virtues embodied in this approach to vagueness.

## A Condition on the Expression of a Proposition

<sup>1</sup> This characterization is somewhat rough, but (as it's frequently noted in the literature) it's difficult to characterize vagueness in a non-theory-laden manner.

<sup>2</sup> It should be assumed throughout that the hairs on the person's head are arranged in a manner optimal for non-baldness.

<sup>3</sup> The reasoning used is simply Universal Instantiation and *Modus Ponens*. The latter alone suffices, if one replaces the universally quantified second premise with the appropriate conditionals. Note, in particular, that the principle of mathematical induction is *not* employed, though it (or some other sufficiently strong principle) would indeed be required to reach the conclusion that *no* one is bald.

<sup>4</sup> Keefe and Smith (1997) is an excellent reader, the editors' introduction providing a valuable survey of the field. Williamson (1994) is an indispensable monograph. The literature on vague *concepts* has been less well-developed and was until recently dominated by degree-theoretic approaches. But see Kamp and Partee (1995).

<sup>5</sup> This use of the term 'context sensitive'—common in semantics and pragmatics—should not be confused with its use in syntax to describe rules insensitive to surrounding syntactic context.

<sup>6</sup> See, e.g., Kamp (1981), Bosch (1983), Raffman (1994), and van Deemter (1996).

Suppose a jar contains 100 color chips, spanning red to orange in imperceptible steps, and I ask you to grab a red one for me. Surely, you can satisfy my request; and when you say “Here’s a red one,” you express a true proposition. If so, then it is *not* a necessary condition on the expression of a proposition that

- (1) the uttered predicates, as used, partition the contextually relevant domain of discourse.

For there’s no reason to think that the predicate ‘is red,’ as used on that occasion, partitions the chips in the jar.

But now suppose I ask you to sort the chips according to whether they are red or orange. You hesitate—at least, once you recognize that they form a sorites series.<sup>7</sup> I call across the room: “I’m curious—is there an even number of red ones?” You are nonplused; you’re not sure what should count here as being red or being orange. Perhaps you request some clarification. If I have none to offer (we can assume I didn’t realize that the jar contained borderline cases), you will be unable to satisfy my request. Likewise, you will be unable to judge whether the sentence ‘The number of red chips is even’ would express something true or false. *Not*, I claim, because you are ignorant of the matter, and unable to rectify your lack—but because in *this* case the lack of partition results in the failure of the predicate to express a property: an utterance of the sentence would thus fail to express a proposition (would fail, that is, to issue in something assessable for truth or falsity).<sup>8</sup> Of course, as competent speakers of English, we understand the *sentence*, we know its standing meaning in the language—so an utterance of it wouldn’t amount to gibberish; and we would know as well much else relevant to interpreting the utterance, such as which chips were in question. But here sentence-meaning and the available contextual information would not suffice to enable the expression of a proposition. Just as with reference-failure (when a putatively referring expression fails on some occasion of use to refer to anything), property-expression-failure thwarts the expression of a proposition; and the failure to partition can

induce property-expression-failure. Leaving out the connecting step, we have:

- (2) a failure of a predicate to partition the contextually relevant domain of discourse can result in the failure to express a proposition.

No doubt these glosses are prejudicially theory-laden. But I needn’t claim that mundane cases *force* us to accept (1) and (2)—only that they suggest them. It’s *prima facie* reasonable to accept (1) and (2) in light of such cases, and to that extent they are *motivated*. Let’s see where they lead: the proof of the pudding is in the eating.

If we accept (1) and (2), we will want to ask *when* the failure to partition can result in the failure to express a proposition: what distinguishes the case in which I ask you to grab a red chip, from the case in which I ask you to sort all the chips in the jar? A natural thought is that, in the first case, the lack of partition *just doesn’t matter*: we may proceed *as if* ‘red’ partitions the chips, because we may *ignore* the borderline cases as *irrelevant* to our purposes. If something along these lines is correct, then it is a condition on the expression of a proposition that

- (3) the speaker may proceed as if the uttered predicates, as used, partition the contextually relevant domain of discourse.

It is clearly a crucial question whether and to what extent this idea can be clarified. Some light is shed if we recast the condition as a constraint on *pragmatic presupposition*—that is, as a constraint on the propositions presumed mutually taken for granted in a given conversational context.<sup>9</sup> The idea is that speakers, in using predicates, act as if, or presuppose, that the predicates, as used, partition the domain. When they must also presuppose, however, that a predicate, as used, does *not* partition the domain, when the failure to partition becomes contextually salient, then the resulting set of presuppositions is obviously inconsistent and thus incoherent. The recast condition thus reads:

- (C) It is a coherent presupposition that the predicates, as used, partition the domain.

<sup>7</sup> I use the phrase ‘sorites series’ for any series from which we can construct a *prima facie* paradoxical sorites argument.

<sup>8</sup> I am thus assuming that the “epistemic” view of vagueness, according to which borderline cases reflect our often in principle ignorance, is false. Williamson (1994) defends such a position.

<sup>9</sup> For this conception of context, see Stalnaker (1974) as well as other papers now collected in Stalnaker (1999).



There is obviously much more to be said here, but again this suffices to motivate (C). What I have to say next will exhibit its attractions.

### Application to the Sorites

I have been suggesting that *whether* sentences containing vague predicates express propositions is a context-sensitive matter. Many mundane utterances of such sentences succeed in expressing propositions, but some don't—in particular, those entered in contexts in which (C) is violated. What I'll argue now is that contexts in which one considers a would-be sorites argument fall into this latter class—save when one of the argument's premises is false or its conclusion true.

So, consider a standard version of the sorites—supposing the correctness of (C). Say, for example, we have objects numbered 1, 2, 3, ..., 5,000,000 such that each is F-er than its successor, and we are presented with the following would-be argument:

F(1)  
For all x, if F(x), then F(x+1)  
 F(5,000,000),

an argument which is paradoxical if we're inclined to consider 1 clearly F, 5,000,000 clearly *not* F, and the difference among neighbors too small to make a difference.<sup>10</sup>

What can we say on my approach? If we eschew logical deviancy, to attempt to assess the crucial sorites premise is to attempt to assess as well its negation 'There exists an x such that F(x) but not F(x+1).' But, in a setting in which the objects are ordered as above, to consider *that* just is to consider what would be the assertion of a partition: one is asking whether there's a *last* x that to which 'F' applies. Whether there is in fact a partition thus matters here; the lack of one, if such there be, cannot be ignored. Considering a sorites, that is, renders it salient whether there is a partition. So, when there is *not* one, the speaker cannot simply proceed as if there were. The condition on the expression of a proposition is thus violated; attempting to use the offending predicate in this way fails to issue in an argument at all. Indeed, no use of the predicate in this context will contribute to the expression of a proposition.

That's what happens when there's *not* a partition. But what about when there *is* one? Then, of course, one of the premises will be false

or the conclusion true. If either the predicate's extension or its complement is empty, there's trivially a partition: if the former (if *no* x is F), then the first premise is false, and if the latter (if *all* x are F), the conclusion is true. If neither is empty—if it's *not* the case that, for either one side or the other, *everything* in the domain falls in it—there is then a non-trivial partition, but then the sorites premise is false: there *is* an x such that F(x), but not F(x+1).

In sum, when the sorites predicate does *not* partition the domain, sentences containing it cannot express propositions and so no argument is presented; an argument is indeed presented when the predicate *does* partition the domain, but then only one that is straightforwardly unsound. The dissolution of the sorites thus follows fairly directly from acceptance of (C).

Indeed, the dissolution is so neat that one might reasonably wonder whether I can explain the force, however illusory, the paradoxical argument seems to possess. Let me try.

So, if the sorites is correctly dissolved along these lines, why do we nonetheless feel the *force* of the argument? A first point to note is that there is a side to us (or at least to many of us) that does *not* feel the force, at least not always. We are, I suggested, nonplussed in situations where (C) is violated; we feel that something is awry. The thought that vagueness *usually* just doesn't matter—and that the puzzles to which it putatively gives rise somehow fall into the category of "don't-cares"—is quite natural: at least it's certainly one I frequently encounter. This, I think, is an important datum—and it's well-accounted for by the present approach. The view is thus consistent with—and indeed perhaps explains—an aspect of the phenomenology, if you will, of vague language use.

But yet those premises seem true and that reasoning valid. Especially those of us whom the sorites has "intermittently obsessed for years"<sup>11</sup> will want to know how a non-argument could have kept us awake so many nights. There are really two facts that need explaining: first, why the sorites seems to have force prior to exposure to my diagnosis, and, second, why this force *persists* even if one does adopt this approach.

Well, the reasoning *is* valid in the following sense: the argument has the syntactic form of a valid inference schema, one such that *if* its premises express true propositions and *if*, in the course of the reasoning, there is no equivocation-inducing context-shift, then the reasoning

<sup>10</sup> I label the second premise the 'sorites' premise and shall refer to 'F' as the 'sorites' predicate.

<sup>11</sup> I borrow this excellent description of philosophical pathology from an unpublished paper on scepticism by Rogers Albritton.

preserves truth. But what we have just seen is that there *is no one* context in which the premises express truths and the conclusion a falsehood. If it can seem otherwise, this is in part because, in our attempt charitably to interpret these sentences—to project them into an appropriate context—it is all too easy to conflate the sorites premise, in which the predicate is *used*, and the *meta-linguistic* claim that the predicate fails to partition the domain.<sup>12</sup>

Note further that this urge to construct an appropriate context is to an extent beyond our conscious control. We just can't help trying to make sense of what we perceive as linguistic tokens; we typically exercise our interpretive capacities automatically, almost as a reflex.<sup>13</sup> In particular, this is true of our ability to track contextual features relevant to the understanding of utterances of context-sensitive sentences. Indeed, this is a necessity imposed by nature's design constraints: we simply would not be able to speak, and cognize generally, with the real-time facility we do possess, if the exercise of the requisite capacities required more conscious reflection. Lacking reflective awareness of the full extent of our reliance on contextual cues, we are nonplussed when our reflexive attempts to project a sentence into an appropriate context founders.<sup>14</sup> And even if we consciously conclude that our inability to identify an appropriate context for the sorites is owing to the absence of such a context, not to our ignorance or inadequacy, this doesn't mean that the would-be argument loses all force, psychologically speaking: again, the automatic nature of our interpretive skills places them to a certain extent beyond our control. We might thus compare the *persistence* of the sorites with that of the Müller-Lyer optical illusion.<sup>15</sup>

---

<sup>12</sup>Further interference is caused by the similarity to the claim that for all  $x$ , if  $F(x)$ , then *there is prima facie reason to believe* that  $F(x+1)$ . Cf. Travis (1985).

<sup>13</sup>Cf. Fodor (1983)'s dedication.

<sup>14</sup>Note that this failure does not so readily flummox us when the dependency upon context is more obvious—as when a deranged person yells at a 'you' who clearly isn't there.

<sup>15</sup>The comparison needn't be pushed too far in order to make its point. In particular, I don't mean to imply that there exists something like linguistic experience, analogous to visual or auditory experience—though, of course, the comprehension of what is said by particular utterances requires sensory experience of some sort. Another possible point of disanalogy is that whereas the Müller-Lyer illusion arises from contingent features of our visual system, it is perhaps arguable that the kind of cognitive design constraints that power the sorites are not specific to our species, or even to

## Some Virtues of this Approach

We now have a first reason for finding our condition attractive: it yields an attractive solution to the sorites. I'll use the space remaining to indicate briefly a few further virtues of the present approach.

A. It is not uncommon for responses to paradoxes (and not just responses to the sorites) to call forth the complaint that they are unmotivated and unilluminating, mere *ad hoc* tricks tailored to finesse a local problem. The present approach, however, is not open to this charge. I have already claimed that our condition on the expression of a proposition is *motivated*. If this is right, then our response to the sorites is to that extent motivated as well. I have *also* already shown how one can locate our condition in the broader theoretical framework that identifies a context of utterance with a set of pragmatic presuppositions. This effects a unification of otherwise disparate phenomena and enables a perspicuous description of their interaction. I'll now indicate two further ways in which this approach finds place in a more general perspective and hence helps *illuminate* the phenomena in question.

First, the approach readily generalizes to various other, *prima facie* related puzzles. This is because the presence of vagueness is not the only reason why a predicate may fail to partition a domain. There are, for example, predicates whose application may depend upon a contextually varying combination of conditions (with contextually varying weights). Hard cases, in which these conditions of application seem insufficient (is coffee food?), may likewise be seen as violating our condition.<sup>16</sup> What's more, we may see such puzzles as the problem of the many and those surrounding vague identity as involving referential indeterminacy closely correlated to the failure of certain predicates to partition the domain. It seems undeniable, for instance, that a competent speaker can, on some occasion, refer to a desk; but it can seem impossible to say which of the many candidate collections of molecules is, or constitutes, the desk to which she refers. But perhaps this is a bad question: that the predicate 'is a part of the desk' fails to partition the domain, though irrelevant normally, can block the expression of a

---

those similarly constituted or organized, but rather apply to all (finite) rational agents.

<sup>16</sup> The relevant phenomenon goes by many names. I borrow the coffee example from Sorenson (1991) who labels it 'conflict' vagueness.

proposition in cases where it matters, and this is arguably correlated with, if not explanatory of, 'the desk's' contextually varying ability to refer.<sup>17</sup> Perhaps these brief remarks are insufficient to convince, but they do at least indicate how we may exploit the fact that our condition adverts to matters broader than vagueness to illuminate a variety of puzzling phenomena. The solution to the sorites would thus follow as but one consequence of a more general framework.

Second, recognizing our condition on the expression of a proposition helps illuminate what we might label the *dynamics* of vague language use. I don't have space to go much into these matters here, but the basic point is that the violation of (C) puts pressure on speakers to adjust their use of the relevant predicate so as to *restore* (C)'s satisfaction. Among the more obvious options is to *sharpen* the offending vague predicate. As your boss, for example, I might settle borderline cases by simply *stipulating* that chips shall count as red, for the purposes at hand, only if they exhibit such-and-such precisely characterized reflectance patterns. Now, the *amenability* of vague terms to such sharpening is an aspect of their standing linguistic meaning: vague terms are context-sensitive in that they may express different properties on different occasions of use, depending on the standards of precision in play. (For example, the contextually relevant standards of precision for being too young can shift, expand, and contract depending on whether we're discussing whether she may read from the Torah, drive a car, or stay up to watch the final election returns.) Sentences containing vague predicates are thus not only context-sensitive as to *whether* they express a proposition (as I urge above), they are of course also context-sensitive as to *what* proposition they express, when they express a proposition at all. And these two facets of their context-sensitivity interact, in that it is *because* of the latter that a speaker can *adjust* the context so as to avoid the failure to express a proposition allowed by the former. Adjusting a context to sharpen a predicate is clearly subject not only to semantic constraints but more generally to constraints of reasonableness. Just what these constraints are is a complex matter—but it is only the recognition of (C) as a condition on the expression of a proposition that allows us a purchase here at all.

**B.** Another common pitfall responses to the sorites must avoid is the problem of higher-order

vagueness. On my proposal, it is the salience of the failure to partition that forces us into an incoherent context: but is there, in a given sorites series, a first object the salience of which effects this context-shift? In effect, we are putting forward a meta-linguistic sorites: consider utterances of the sentences 'One grain does not constitute a heap,' 'Two grains do not constitute a heap,' 'Three grains do not constitute a heap,' . . . —which is the first utterance that fails to express a proposition?

My view, however, yields a natural answer to such questions. The predicate 'expresses a proposition' is itself vague, and so, as with all vague terms, sentences containing it will fail to express propositions when the predicate's failure to partition the domain cannot be ignored. Higher-order vagueness is thus reflected on this approach in the vagueness of the terms used to describe language use generally (and thus used to describe vagueness in particular).

Of course, it should only be expected that there be vagueness here too: why should the language used to describe language be immune to the vagueness that infects practically all empirical terms? Indeed it would be extremely surprising, if things were otherwise; the precision of this one region of language would cry out for explanation. But, in fact, as the meta-linguistic sorites itself demonstrates, there *are* borderline cases of expressing a proposition: a realistic view must therefore find proper place for them, rather than positing answers where none are to be had. Given that this region of language *does* contain vagueness, it is thus a virtue of my view that it covers these cases as part of a uniform treatment. (We also have here a further example of illumination: it is *instructive* to see how first-order vagueness among terms generally is among the sources of that vagueness to which terms used to describe language use in particular are prone.)

**C.** I have space to mention but one more virtue my approach possesses—viz., the fact that it avoids those offenses to common sense characteristic of much discussion of the sorites.

If we may measure a puzzle's difficulty by the *prima facie* absurdity of the sincerely and ably defended responses it elicits, then it is clear that the sorites ranks frustratingly high among its philosophical peers. Nihilism (the view that vague predicates are empty) provides the most extreme example, but there are also, for instance, the claims that contradictions are half-true, typically endorsed by degree-theoretic approaches; that vagueness is but an epistemic phenomenon, reflecting our (often in principle) ignorance of borderlines; and that vagueness does

---

<sup>17</sup> For the problem of the many, see Unger (1980). On vague identity, see Evans (1978).

not exist at higher-orders—there is always a sharp line between the clear and gray ranges of a predicate’s application—to which at least simpler versions of supervaluationism are committed.<sup>18</sup> Indeed, it is a common sentiment among writers on vagueness that *any* position will exact a price—so formidable is the puzzle. But the suggestion I have explored, as far as I have been able to determine, is an exception. If I am right that my view better avoids offending common sense than its competitors, it obviously possesses in that respect an enormous advantage.

I hasten to add that I do believe that my approach brings in tow some surprising *theoretical* commitments. One, which is obvious, is that one can express a proposition without the uttered predicates being associated with a determinate extension. So, propositions can’t be what many people take them to be. Another, which I did not have space to discuss here, is that, on my view, the phenomenon of vagueness imposes limits on our ability to survey our semantic competence: it restricts the propositions expressible within any given context and thus the propositions available for the construction of truth-conditions, and it likewise inhibits our ability to isolate, on the one hand, the contribution to content of linguistic meaning, from, on the other, the contribution of context. A desire to avoid these consequences would no doubt constitute a reason to resist my approach. But to question certain highly *theoretical* claims is not to maintain a *prima facie* absurd view. I would thus turn matters around: if it is only those assumptions that sustain the sorites paradox, then we have an argument for why those assumptions have to go. I won’t go so far as to claim this upshot as a further *virtue* of my view, but it certainly adds to its interest.<sup>19</sup>

### Acknowledgements

I thank the many people who have helped me think about this material, especially Warren Goldfarb, Richard Heck, and Ian Proops.

### References

- Ballmer, T. T. and M. Pinkal, eds. (1983). *Approaching Vagueness* (North-Holland).  
 Bosch, Peter (1983). “‘Vagueness’ is Context-Dependence. A Solution to the Sorites Paradox,” in Ballmer and Pinkal (1983).  
 Evans, Gareth (1978). “Can There Be Vague Objects?” *Analysis* 38.

- Fodor, Jerry (1983). *The Modularity of Mind* (MIT).  
 French, P., T. Uehling, and H. Wettstein (1980). *Midwest Studies in Philosophy: Studies in Metaphysics* 4 (Minnesota).  
 Gross, Steven (1998). *Essays on Linguistic Context-Sensitivity and its Philosophical Significance* (Doctoral Dissertation, Department of Philosophy, Harvard University).  
 Kamp, Hans (1981). “The Paradox of the Heap,” in Mönnich (1981).  
 Kamp, Hans and Barbara Partee (1995). “Prototype Theory and Compositionality,” *Cognition* 57.  
 Kanawaza, M., C. Piñon, and H. Swart, eds. (1996). *Quantifiers, Deduction, and Context* (CSLI).  
 Keefe, Rosanna and Peter Smith, eds. (1997). *Vagueness: A Reader* (MIT).  
 Mönnich, U., ed. (1981). *Aspects of Philosophical Logic* (Reidel).  
 Munitz, M. and P. Unger, eds. (1974). *Semantics and Philosophy* (NYU).  
 Raffman, Diana (1994). “Vagueness without Paradox,” *Philosophical Review* 103.  
 Sorenson, Roy (1991). “Vagueness within the Language of Thought,” *The Philosophical Quarterly* 41.  
 Stalnaker, Robert (1974). “Pragmatic Presuppositions,” in Munitz and Unger (1974).  
 Stalnaker, Robert (1999). *Context and Content* (Oxford).  
 Travis, Charles (1985). “Vagueness, Observation, and the Sorites,” *Mind* XCIV.  
 Unger, Peter (1980). “The Problem of the Many,” in French *et al.* (1980).  
 van Deemter, Kees (1996). “The Sorites Fallacy and the Context-Dependence of Vague Predicates,” in Kanawaza *et al.* (1996).  
 Williamson, Timothy (1994). *Vagueness* (Routledge).

<sup>18</sup> For details, see the works cited in footnote 1.

<sup>19</sup> I discuss many of the issues raised above at greater length in Gross (1998, chap. IV).

# Simulating Causal Models: The Way to Structural Sensitivity

**York Hagmayer** (york.hagmayer@bio.uni-goettingen.de)  
Department of Psychology, University of Göttingen,  
Gosslerstr. 14, 37073 Göttingen, Germany

**Michael R. Waldmann** (michael.waldmann@bio.uni-goettingen.de)  
Department of Psychology, University of Göttingen,  
Gosslerstr. 14, 37073 Göttingen, Germany

## Abstract

The majority of psychological studies on causality have focused on simple cause-effect relations. Little is known about how people approach more realistic, complex causal networks. Two experiments are presented that investigate how participants integrate causal knowledge that was acquired in separate learning tasks into a coherent causal model. To accomplish this task it is necessary to bring to bear knowledge about the structural implications of causal models. For example, whereas common-cause models imply a covariation among the different effects of a common cause, no such covariation between the different causes of a joint effect is implied by a common-effect model. The experiments show that participants have virtually no explicit knowledge of these relations, and therefore tend to misrepresent the structural implications of causal models in their explicit judgments. However, an implicit task that only required predictions of singular events showed surprisingly accurate sensitivity to the structural implications of causal models. This dissociation supports the view that people's sensitivity to structural implications is mediated by running simulations on mental analogs of the causal situations.

## Introduction

In everyday life as well as in scientific research we rarely observe the behavior of complex causal networks at once. A more typical scenario is that we learn about single causal relations separately, and later try to integrate the different observed relations into a more complex interconnected causal model. For example, we might first learn that aspirin relieves headache. Later we may observe that aspirin unfortunately also creates stomach problems. Now we are in the position of putting these two pieces of knowledge together. The question is how? How are different fragments of causal knowledge integrated into coherent complex structures?

## Bayesian Causal Models

One recent approach to this problem that has become increasingly popular in the past few years postulates *Bayesian network models* for representing causal knowledge (see Pearl, 1988, 2000; Glymour & Cooper, 1999). Bayesian network models provide compact, parsimonious representations of causal relations. For example, Figure 1 displays a causal model that connects five events,  $X_1$ ,  $X_2$ ,  $X_3$ ,  $X_4$ ,  $X_5$ . One way to represent this domain is to list the

32 probabilities of the joint probability distribution,  $P(X_1, X_2, X_3, X_4, X_5)$ , by considering every combination of present and absent events. Another possible strategy is to encode the base rates and all covariations that can be computed between five events. However, even with modestly complex structures the number of covariations becomes very large, especially when more complex higher-order covariations between multiple events also are considered. Bayesian network models reduce the complexity of representing causal knowledge by distinguishing between direct causal relations (the arrows in Fig. 1), and covariations that can be derived by using information encoded in the *structure* of the causal models. The structure of causal models primarily expresses information about conditional independence between events. For example, in Figure 1 event  $X_4$  is coded as being independent of event  $X_5$  conditional upon event  $X_3$ . Conditional independence greatly simplifies computations by allowing the derivation of the indirect relations from products of the relevant components (see Pearl, 1988; Glymour & Cooper, 1999). In Figure 1 the joint probability distribution can be factorized into the product of a small number of unconditional and conditional probabilities,

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_5 | X_3) \cdot P(X_4 | X_2, X_3) \cdot P(X_3 | X_1) \cdot P(X_2) \cdot P(X_1).$$

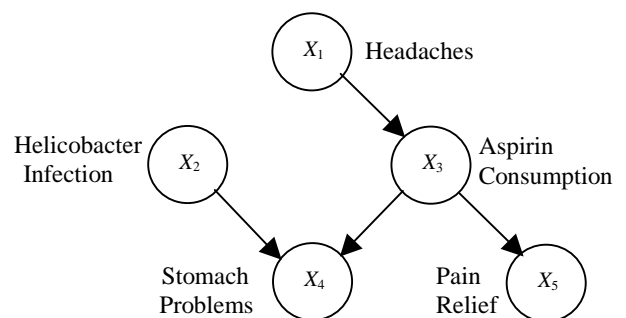


Figure 1: Example of a Bayesian Network

The distinction between direct causal relations and indirect relations can also be used for the *integration* of separate pieces of causal knowledge. Combining the information that aspirin relieves headache with the information that it additionally causes stomach problems yields a

common-cause model with aspirin playing the role of the common cause of two independent effects, relief of headache and stomach problems (see Fig. 2, left). By contrast, integrating the two causal relations “Aspirin causes stomach problems” with “Helicobacter pylorus causes stomach problems” would yield a different structure, a common-effect model, in which two independent causes converge on a joint effect (see Fig. 2, right). In both examples, two independent causal relations are being integrated. However, the outcome of the integration process is different. The two different causal models entail different implications for the indirect relations between events.

### Structural Implications of Causal Models

The basis for the possibility of integrating different causal links into coherent wholes are the structural implications of causal models. In our experiments we focused on two simple models, a common-cause and a common-effect model. Both models integrate two causal links but entail distinctly different implications for the non-causal relations.

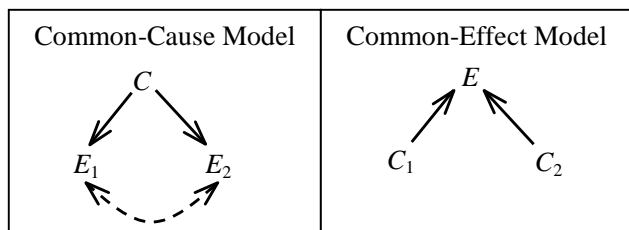


Figure 2: Implications of Different Causal Models

Figure 2 (left) depicts a common-cause model with a common cause  $C$  producing two independent effects  $E_1$  and  $E_2$ . Common-cause models of this kind entail a (spurious) covariation among the effects. Provided the common cause independently generates the two effects, the joint probability of the effects,  $P(E_1, E_2)$ , can be calculated by taking the product of the base rate of the cause,  $P(C)$ , and the two conditional probabilities,  $P(E_1|C)$  and  $P(E_2|C)$  (see also Appendix). Thus, although the two effects may never have been observed together, the causal model still allows it to derive a prediction for the patterns that should be expected. Common-cause models clearly differ from common-effect models. Figure 2 (right) shows an example in which two causes,  $C_1$  and  $C_2$ , are linked with a joint effect  $E$ . Common-effect models do not imply covariations among the different causes of the joint effect. The causes may covary in a specific learning situation but this covariation is not implied by the model, it is something that has to be explicitly encoded. This is the reason why in the example shown in Figure 1 common effects were conditionalized on patterns of its direct causes (e.g.,  $P(X_4|X_2, X_3)$ ). However, this is only possible when all the relevant events have been observed together, and when the number of relevant patterns is small enough not to surpass information processing limitations. In more complex cases and in situations in which causal knowledge has to be generated from different learning experiences, *causal schemas* have been postulated in the literature (Pearl,

1988). For common-effect models, the *noisy-or schema* has been proposed as a plausible integration schema (see also Waldmann & Martignon, 1998). According to this schema  $P(X_4|X_2, X_3)$  can be reduced to  $[1-(1-P(X_4|X_2)) \cdot (1-P(X_4|X_3))]$ , an expression that only contains probabilities referring to direct causal relations. The noisy-or schema assumes that different causes have independent and additive influences on the common effect. Given that common-effect models do not imply covariations among the causes a further reasonable default assumption is that they occur independently. A number of psychological experiments have shown that learners indeed tend to initially assume independence (see Waldmann, Holyoak, & Fratianne, 1995).

### Sensitivity to Structural Implications: Computation vs. Causal Simulation

Previous research has demonstrated sensitivity to structural implications of causal models in causal learning (Waldmann & Holyoak, 1992; Waldmann, 2000), causal reasoning (Waldmann & Hagmayer, 1998), and categorization (Waldmann et al., 1995). The processes underlying this sensitivity are unclear, however. The standard approach within the area of Bayesian modeling is to explicitly derive the predicted event patterns or covariations and test these predictions against the data at hand. It appears unlikely that this strategy could be followed in intuitive everyday reasoning. Despite the fact that Bayesian models provide a parsimonious way of representing domain knowledge it is also clear that the explicit derivation of indirect relations is often complex and computationally demanding (Glymour & Cooper, 1999). In fact, one reason for the increasing number of automated statistical tools that are currently offered to researchers lies with the fact that the task surpasses the capacity limitations of intuitive reasoning.

However, there is an alternative, more implicit strategy. Instead of explicitly computing covariations we may form mental representations of causal structures that are analogous to the graphical structures used in Bayesian network modeling (e.g., Fig. 1). Similar to toy models, these causal models can then be used to run mental simulations (see also Barsalou, 1999). For example, instead of calculating the probability of patterns within a common-cause model with one cause and two effects we could mentally imagine the presence or absence of the cause, and then generate predictions for each individual effect based on the observed covariations between the cause and either effect. Since these predictions are triggered by a common event within a mental common-cause model the predicted patterns should show the covariations that are implied by the structure of the mental model. These covariations are not the consequence of an explicit computation, they rather are a side effect of the structure of the causal model used to simulate the causal situation in the real world. Therefore it may well be that the predicted patterns exhibit covariations of which the learners are not aware. For the learner it is only necessary to focus on the direct causal relations. All the indirect relations are taken

care of by running simulations on mental analogs of the objective causal situation.

Two experiments will be presented in which participants acquired partial knowledge about separate fragments of common-cause or common-effect models. To test whether they were sensitive to the additional covariations implied by the different causal models, two types of measures were collected. *Explicit knowledge* was assessed by means of probability estimates in which participants were requested to estimate the strength of the indirect, not directly observed relation. Based on the assumption that explicit computations of the answers to these questions are hard we expected poor performance with this task. However, the second task was designed to tap into *implicit knowledge* generated by causal simulations. In this task, participants were requested to predict the pattern of events they expected to see. For example, in a common-cause condition (see Fig. 2) the experimenter instructed participants to imagine that the cause was present and to make a prediction about the two effects. A typical finding with this type of task is that participants tend to match the probabilities they have seen in the learning situation. Since in the present task the two effects never have been seen together, direct experience with the patterns is not available. However, it is possible that participants match the probabilities for each relation independently within a mental analog of a common-cause model. The model itself generates covariations that have never been observed directly. The crucial measure in this task is the covariation between the predicted effects that can be derived from participants' responses. The causal-simulation account predicts that these patterns should display the covariations implied by the causal models even when no explicit knowledge could be detected in the explicit task.

## Experiment 1

The goal of this experiment was to investigate whether learners who have acquired partial knowledge about fragments of causal models are sensitive to the structural implications of these models. Participants were given the task to learn about the causal relations between the mutation of a gene and the prevalence of two (fictitious) substances (enzyme BST and brasus protein). We used a trial-by-trial learning procedure in which participants worked through a stack of index cards with information on the front side about whether a mutation of the gene occurred or not. By turning around the individual cards participants received information about the presence or absence of either the enzyme BST or the brasus protein. To ensure that no covariation between the enzyme and the protein could be observed the cards were divided into two different stacks, one for each substance. Participants were instructed to alternate between the stacks in the course of the learning phase. In the initial instruction the separate stacks were characterized as displaying the raw data of two different research projects located at different universities. The task and the presentation of the data were identical for all participants. They first received information about the mutation of the gene on the front side of the cards, and then were shown information about the occur-

rence of either the enzyme or the protein on the backside. The learning phase consisted of 80 cards, 40 for each substance.

Two factors, type of causal model and degree of covariation, were manipulated yielding four experimental conditions. The first factor contrasted two different causal models. One group of participants read in the initial instructions that the researchers were interested in finding out whether the mutation causally influences the two substances (common-cause model)(see Fig. 2, left). In contrast, for the second group the two substances were described as potential causes of the mutation (common-effect model)(see Fig. 2, right). The second factor manipulated the strength of the relation between mutation and the two substances. The strength was always equal for both substances and either weak or strong. Table 1 displays the absolute frequencies used in this experiment. Thus, for example, participants in the condition with strong connections saw 16 cases for each substance in which the presence of a mutation of the gene was paired with the presence of the substance.

Table 1: Frequencies in Experiment 1

	Strong Condition		Weak Condition	
	Substance	No Substance	Substance	No Substance
Mutation	16	4	10	10
No Mutation	0	20	6	14

Apart from the different initial instructions about the underlying causal model the learning phases and the test phases were identical within the conditions with strong or weak relations. Regardless of whether the mutation of the gene was introduced as a cause or as an effect, information about its presence or absence was delivered before information about the substances was given.

The learning phase was followed by a test phase in which participants' assumptions about the covariation between the two substances was assessed. This covariation had to be inferred because the two substances had never been seen together. To test whether participants were sensitive to the different implications of the two causal models we compared an implicit with an explicit measure of knowledge. In the *implicit* test procedure participants received 20 new index cards in a random order, half of them indicating that in this particular case a mutation had occurred. The rest of the index cards described cases in which no mutation had occurred. Participants' task was to predict for each case individually whether either of the two substances was present or absent. No feedback about the substances was provided during this test phase. Since patterns of substances had to be predicted it was possible to analyze the amount of covariation between the substances in the responses of the participants. We used the *phi* correlation coefficient as a measure of the degree of the implicitly predicted covariation (see Appendix). In a second task that followed the

implicit task, we investigated participants' *explicit* expectations. In this task they had to estimate the probability that the second substance is present conditional upon the first being present ( $P(\text{substance}_2|\text{substance}_1)$ ) and being absent ( $P(\text{substance}_2|\sim\text{substance}_1)$ ). As with the implicit measures the explicit estimates were transformed into *phi* correlations that allowed us to directly compare the implicit with the explicit measure.

What are the normative Bayesian predictions for the presented data? When a common-cause model is assumed it is appropriate to encode the conditional probabilities directed from the cause (i.e., mutation) to its effects (i.e., substances). In this direction, the data display a conditional probability of either substance in the presence of the mutation (i.e.,  $P(\text{substance}|\text{mutation})$ ) of .80 in the strong and .50 in the weak condition. The corresponding values in the absence of a mutation ( $P(\text{substance}|\sim\text{mutation})$ ) are 0 in the strong versus .30 in the weak condition. Taking the difference of these numbers yields the widely used *contingency* ( $\Delta P$ ) measure of statistical strength (Eells, 1991). Accordingly, the contingency is  $\Delta P = .80$  in the strong and  $\Delta P = .20$  in the weak condition.

Within the framework of a common-effect model the same data should again be analyzed from causes to effects. In this condition the substances play the role of the causes. Thus, it is appropriate to compare  $P(\text{mutation}|\text{substance})$  with  $P(\text{mutation}|\sim\text{substance})$ . The data yield a probability of the mutation in the presence of the substance of 1 in the strong and of .63 in the weak condition. The corresponding values in the absence of the substance are .17 in the strong and .42 in the weak condition. These numbers imply almost the same contingencies as in the common-cause condition of  $\Delta P = 0.83$  (strong condition) and  $\Delta P = .21$  (weak condition).

On the basis of structural information from the causal models these numbers can be used to derive the predicted covariation between the substances. While the common-effect model does not imply a covariation, the common-cause model entails that the observed joint probability should correspond to the product of the base rate of the cause and the conditional probabilities observed for each causal link. These probabilities can be transformed into a *phi* coefficient of correlation (see Appendix). The data presented imply a *phi* correlation of  $r = .67$  between the substances in the strong condition and of  $r = .042$  in the weak condition.

## Results and Discussion

The results are based on 48 students from the University of Göttingen who were randomly assigned to one of the four learning conditions. Table 2 shows the means for both the explicit and the implicit measure obtained in the four conditions.

The correlations that the participants generated in the implicit prediction task resemble very closely the ones normatively implied by the causal models. Participants in the common-cause condition generated a high mean correlation of .62 between the substances when the causal connections were strong and a mean correlation of -.004 when they were weak. In contrast, in the common-effect

condition in which they received identical learning inputs as participants in the corresponding common-cause condition the prediction responses displayed generally low correlations in both conditions. An analysis of variance revealed a significant main effect for the factor causal model,  $F(1, 44) = 7.28$ ,  $p < .05$ ,  $MSE = .14$ , and a significant main effect for the factor strength of covariation,  $F(1, 44) = 18.4$ ,  $p < 0.01$ ,  $MSE = .14$ . The interaction failed to be significant,  $F(1, 44) = 2.33$ ,  $p = .13$ ,  $MSE = .14$ .

Table 2: Means of Implicit and Explicit Measures (Experiment 1)

	Implicit Measure: Generated Correlations		Explicit Measure: Estimated Correlations	
	Common-Cause Model	Common-Effect Model	Common-Cause Model	Common-Effect Model
Strong Relations	.622	.168	.286	.161
Weak Relations	-.004	-.130	-.109	.039

The explicitly estimated correlations clearly differed from the implicitly generated ones (see Table 2). There was no significant difference of the estimated correlations in the two contrasted causal models,  $F < 1$ . Only the difference between the conditions in which strength of covariation was manipulated proved significant,  $F(1, 44) = 8.05$ ,  $p < .01$ ,  $MSE = .10$ .

These results indicate that participants showed little sensitivity to the implications of causal models when the task required explicit estimates. They seemed to be aware of the fact that the inferred covariations somewhat depend on the strength of the causal links responsible for the covariations, but they did not explicitly grasp the structural difference between common-cause and common-effect models. By contrast, the implicit measure displayed surprisingly accurate predictions. In this task, participants clearly differentiated between common-cause and common-effect models despite identical learning inputs. In our view, this finding supports the prediction that sensitivity to structural implications can be achieved by running simulations on mental analogs of causal models.

## Experiment 2

In Experiment 1 participants first were informed about whether a mutation of the gene occurred or not, and then learned for each substance separately whether it was present or absent. This procedure served the goal of presenting identical learning inputs to participants in the different conditions. It raises the question, however, whether the observed asymmetries of sensitivity to implied covariations are due to the contrasted causal models or rather to differences in the direction of required inferences during learning. In the common-cause condition learning was directed from cause to effects (predictive learning), whereas in the common-effect conditions the very same



learning items implied that learning proceeded from effect to causes (diagnostic learning). Thus, it may be speculated that differences between predictive and diagnostic learning rather than differences in the underlying causal models may be the reason for the obtained results.

The goal of Experiment 2 was to replicate the results of Experiment 1 and to control for the direction of learning. Moreover, unlike in Experiment 1 the conditional probabilities and contingencies were equalized in the contrasted conditions. Material and procedure were taken from Experiment 1. All participants had the task to learn about the causal connection between mutation and the two substances. Again, as learning input they received index cards separated into two stacks which either provided information about the relation between the mutation and the enzyme BST or between the mutation and the brusus protein. Table 3 shows the frequencies of the different patterns that were presented during the learning phase.

Table 3: Frequencies in Experiment 2

	Substance	No Substance
Mutation	25	5
No Mutation	5	25

These frequencies implied conditional probabilities between the mutation and the substances that were completely symmetric ( $P(\text{mutation}|\text{substance}) = P(\text{substance}|\text{mutation}) = .8$ , and  $P(\text{mutation}|\sim\text{substance}) = P(\text{substance}|\sim\text{mutation}) = .2$ ). Thus, the contingencies were identical in both directions ( $\Delta P = .60$ ).

Two factors were manipulated in Experiment 2. The first factor manipulated the assumed causal model by means of differential initial instructions. As in Experiment 1, the mutation of the gene was either introduced as the cause of the two substances (common-cause model) or as their effect (common-effect model). The second factor manipulated the learning direction. Learning proceeded either from causes to effects (predictive learning) or from effects to causes (diagnostic learning). Thus, half of the participants received information about the mutation first before learning about the substances whereas the other half first read information about the presence or absence of one of the substances, and then received feedback about the mutation. In fact, the same index cards were used for all participants, the only difference was which side they saw first. Information about the mutation was shown first in the predictive version of the common-cause condition and in the diagnostic version of the common-effect condition. The reversed cards showing information about the substances first were given to participants in the predictive common-effect and the diagnostic common-cause conditions. Using the procedures described in the Appendix, a  $\phi$  correlation of  $r = .37$  between the substances can be derived for the common-cause model in which they played the role of effects. This is about half the size of the implied covariation in the condition with strong relations of Experiment 1. Thus, a smaller effect

size is to be expected in the present experiment. In contrast to the common-cause model, the common-effect model does not imply any covariation between the causes. These different structural implications are, of course, independent of the direction of learning.

As in Experiment 1, sensitivity to implied covariations was assessed by means of implicit and explicit measures. Regardless of the learning direction the implicit test always presented information about the mutation of the gene as the cue for the predictions. Participants were shown 20 new cases, half of which describing mutations, and had to predict for each case individually whether either of the substances was present or not. The explicit task in which participants estimated conditional probabilities followed the implicit one (see Experiment 1).

## Results and Discussion

64 students from the University of Göttingen were randomly assigned to one of the four conditions. The means of the  $\phi$  correlations that were either generated (implicit measure) or estimated (explicit measure) in the four different conditions are shown in Table 4.

Table 4: Means of Implicit and Explicit Measures (Experiment 2)

Learning Direction	Implicit Measure: Generated Correlations		Explicit Measure: Estimated Correlations	
	Common-Cause Model	Common-Effect Model	Common-Cause Model	Common-Effect Model
Predictive	.243	.013	.258	.239
Diagnostic	.186	-.001	.129	.210

As in Experiment 1, assumptions about the underlying causal model clearly influenced the implicit measure. The main effect for the factor causal model was significant for the generated correlations,  $F = 4.97$ ,  $p < .05$ ,  $MSE = .14$ . In general, participants generated higher correlations between the substances when they were viewed as effects (common-cause model) than when they had been characterized as causes of the mutation (common-effect model). In the common-effect condition the generated covariations between the two substances (i.e., the causes) were very close to 0 which supports our prediction that independence between causes is assumed in common-effect models. Neither the factor learning direction nor the interactions with this factor proved significant ( $F < 1$ ).

In contrast to the implicit measures, no sensitivity to the structural implications of causal models could be detected with the explicit measures. In general, participants tended towards correlations that clearly differed from 0 but showed no sensitivity to the assumed causal model. None of the effects approached significance in an analysis of variance in which type of causal model and learning direction entered as factors ( $F < 1$ ).

These results clearly support the conclusions of Ex-

periment 1 by demonstrating sensitivity to structural implications with an implicit but no sensitivity with an explicit measure. Consistent with the normative analysis, the implicit measures yielded higher covariations for the common-cause than for the common-effect model. The present experiment also shows that this pattern of results is not due to differences in the learning procedure (predictive vs. diagnostic) but rather is based on differences of the assumed causal models.

## Conclusions

Research on causality belongs to the truly interdisciplinary topics of cognitive science. There are differences in the research focus between disciplines, however. Whereas the majority of studies within cognitive psychology have focused on single cause-effect relations, researchers in the areas of computer science and philosophy have become increasingly interested in complex causal structures (e.g., Glymour & Cooper, 1999; Pearl, 2000). The goal of the present research is to bridge this gap without forgetting the inherent information processing limitations of humans. It is unlikely that untutored human learners are able to store and use the complex information embodied in even fairly simple causal structures. Therefore we have focused on a more realistic task in which participants learned about different fragments of a causal model separately, and later were confronted with the task to integrate the different pieces in order to predict unobserved covariations. To solve this task correctly, knowledge about structural implications of different causal models has to be activated. Research on Bayesian networks has shown that structural information greatly simplifies causal computations but it also has demonstrated that the task still remains complex. Consistent with this analysis both experiments have demonstrated that participants showed little explicit knowledge about differences between causal models, even when the models were extremely simple. Participants' explicit judgments did not distinguish between a condition in which the target events were two effects of a common cause and a condition in which these events represented two causes of a common effect. This result raises doubts as to humans' competence to correctly learn about causal structures in the world. However, a second, more implicit measure displayed surprisingly accurate inferences. When the task required predicting individual events, participants proved sensitive to the difference between common-cause and common-effect models. This dissociation between explicit and implicit measures is consistent with the view that mental simulations of causal models support the implicit task. Generating predictions by means of a mental simulation capitalizes on causal structure without requiring explicit knowledge. As long as the mental representation mirrors the causal features of the represented domain, simulations should display the same structural constraints. Therefore causal simulations allow us to generate correct predictions without requiring complex, explicit computational inferences.

## References

- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-660.
- Glymour, C. N., & Cooper, G. F. (1999). *Computation, causation, and discovery*. Cambridge: MIT Press.
- Eells, E. E. (1991). *Probabilistic causality*. Cambridge: Cambridge University Press.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco: Morgan Kaufmann.
- Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 53-76.
- Waldmann, M. R., & Hagmayer, Y. (1998). Estimating causal strength: The role of structural knowledge and processing effort. *Unpublished manuscript*.
- Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, 121, 222-236.
- Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General*, 124, 181-206.
- Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry, *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

## Appendix

The following derivation shows how joint probabilities and correlations can be derived for common-cause models. In the formulas,  $s_1$ ,  $s_2$  represent the two substances and  $m$  the mutation. “~” signifies the absence of an event. The joint probability of the two substances can be computed by

$$P(s_1.s_2) = P(s_1.s_2|m) \cdot P(m) + P(s_1.s_2|\sim m) \cdot P(\sim m) \quad (1)$$

Common-cause models assume that the effects are independent conditional upon the states of the common cause, that is:

$$P(s_1.s_2|m) = P(s_1|m) \cdot P(s_2|m)$$

Thus, Equation 1 can be simplified:

$$P(s_1.s_2) = P(s_1|m) \cdot P(s_2|m) \cdot P(m) + P(s_1|\sim m) \cdot P(s_2|\sim m) \cdot P(\sim m)$$

The joint probabilities for the other patterns (e.g.,  $P(s_1.\sim s_2)$ ) can be calculated in a similar fashion. These probabilities can be used to compute *phi* correlation coefficients based on the following formula:

$$r = \frac{P(s_1.s_2) \cdot P(\sim s_1.\sim s_2) - P(s_1.\sim s_2) \cdot P(\sim s_1.s_2)}{\sqrt{(P(s_1.s_2) + P(s_1.\sim s_2)) \cdot (P(s_1.s_2) + P(\sim s_1.s_2)) \cdot (P(s_1.\sim s_2) + P(\sim s_1.\sim s_2)) \cdot (P(\sim s_1.s_2) + P(\sim s_1.\sim s_2))}}$$

This procedure of computing *phi* correlations can be applied to the patterns predicted by the participants (implicit task) as well as to the estimated conditional probabilities (explicit task).

# The Problem with Logic in the Logical Problem of Language Acquisition

Petra Hendriks (P.Hendriks@let.rug.nl)

Department of Dutch / Cognitive Science & Engineering (TCW)

University of Groningen

Grote Kruisstraat 2/1, 9712 TS Groningen

The Netherlands

## Abstract

This paper discusses the motivation behind the nativist position with respect to linguistic knowledge. In particular, the discussion focusses on the argument from the “poverty of the stimulus”, which is generally assumed to be the most important argument in favor of a nativist position. On the basis of current views on human reasoning and learning, we will argue that the logical (i.e., non-empirical) part of the poverty of the stimulus argument is invalid. This result substantially weakens the nativist position, although it does not imply that the assumption that there must exist a certain amount of innate domain-specific knowledge has to be abandoned altogether.

## 1. The Logical Problem of Language Acquisition

A fundamental assumption within modern generative syntax is the assumption that knowledge of language is for a considerable part innate. This innate knowledge takes the form of universal principles and parameters that underly all human languages (cf. Chomsky, 1986a; Chomsky, 1995). The most important argument in favor of the nativist position with respect to linguistic knowledge is the so-called poverty of the stimulus argument. The poverty of the stimulus argument yields, in Wexler’s (1991) terms, “Chomsky’s most unique argument” and “the most powerful theoretical tool that we have available to us” (see also Wexler, 1999).

This argument forms the basis of the logical problem of language acquisition, which is essentially an induction problem. A child only hears a finite number of sentences but has to generalize from this input to an infinite set of sentences that includes the input sample. This infinite set is the set of sentences generated by the language the child has to learn, which will be referred to here as the target language. What makes this induction task an extremely difficult one, however, is that an in principle infinite number of hypothetical languages is consistent with the finite input sample. The child has to single out the correct set corresponding to the target language and reject all other sets, which are incorrect hypotheses about the target language. Because every child eventually ends up speaking her mother tongue, children must be guided by constraints that allow them to reject the incorrect hypotheses. Just presenting the child with more sentences of the language she is learning (i.e., providing her with more positive evidence) does not reduce the set of hypothetical languages to the

correct one in all cases (cf. Gold, 1967). If the target language is a subset of the hypothetical language the child entertains, no amount of positive evidence will lead the child to conclude that the adopted hypothesis is incorrect. In this case, only negative evidence will suffice to reject the larger set in favor of the smaller set. However, negative evidence does not seem to occur very frequently in the language input of a child (Brown & Hanlon, 1970), and if it does occur, it usually is not very effective.

The conclusion must be that the language input of a child is insufficient (or, in other words, the “stimulus” is too “poor”) to be able to conclude to the target language. So how are children able to learn their mother tongue, if the information available to them is not sufficient to draw logically valid conclusions from it? Because this is a variant of the question Plato asked himself with respect to knowledge in general, the logical problem of language acquisition is also referred to as Plato’s problem (Chomsky, 1986b).

## 2. The Defective Nature of the Language Input

Another aspect that is sometimes mentioned in relation to the poverty of the stimulus argument is the qualitatively and quantitatively defective nature of the language input the child receives. That is, children frequently hear ungrammatical sentences from their parents and other people. Moreover, the utterances they encounter form only a small fragment of the language they are learning. These two characteristics of the language input have been argued to make language learning extremely difficult, if not impossible, without prior knowledge. The presence of many ungrammatical sentences in the language input is highly problematic because these ungrammatical sentences do not come labelled as ungrammatical. Since the set of utterances the child encounters is relatively small, relevant examples of certain grammatical constructions might not be encountered during the language-learning years. However, both the claim about the qualitatively defective nature of the language input and the claim about the quantitatively defective nature of the language input have been questioned (e.g., Pullum, 1996; Sampson, 1997).

According to Newport, Gleitman and Gleitman (1977), “the speech of mothers to children is unswervingly well formed. Only one utterance out of 1500 spoken to the children was a disfluency”. On the basis of this evidence, it

cannot be maintained that the language input to the child is qualitatively defective, or “degenerate”.

The claim about the quantitatively defective nature of the language input, that is, the non-occurrence of relevant grammatical constructions in the child’s input language, has been refuted by empirical evidence as well. The standard example Chomsky and many others use to illustrate the poverty of the stimulus argument is the formation of yes/no questions (e.g., Chomsky, 1980; Chomsky, 1988). The formation of yes/no questions is dependent on the abstract property of structure-dependency, in particular on distinguishing the main clause from embedded clauses. In order to form a correct yes/no question, the finite verb of the main clause has to be moved to the front of the sentence. To refute the simple but structure-independent and thus false hypothesis that it is the first verb in the sentence that must be moved, the child needs to encounter questions involving an embedded clause which precedes the main verb (for example, “will those who are coming raise their hands?”). Although it is claimed by Chomsky and others (without providing any empirical motivation for this claim) that these examples are very rare, Pullum (1996) found that about 12% of the yes/no questions in the corpus he searched were crucial examples which refuted the incorrect hypothesis. So, relevant sentences for the acquisition of the formation of yes/no questions are expected to occur in the input language of the child. Of course, Pullum did not show this for all other examples that have been used to illustrate the poverty of the stimulus argument, but there is no evidence that it will be different for other examples. Thus, the language input to the child seems to be neither “degenerate” nor “meager”. For this reason, I will focus on the unavailability of negative evidence as the crucial and most uncontroversial aspect of the poverty of the stimulus.

### **3. The Nativist Solution**

The solution that most generative syntacticians have adopted for the logical problem of language acquisition is to assume that the core of the grammar is already present in the child before language learning starts off. This assumption has changed the agenda of research on language learning completely. Language learning is no longer viewed as the acquisition of knowledge on the basis of information present in the input data. Rather, children are born with a “language instinct” (Pinker, 1994). Under this nativist view, language learning merely is a matter of setting the parameter values of an innate universal grammar (UG) on the basis of specific triggering experience. A nativist position is also taken in the recently developed linguistic framework of Optimality Theory (Prince & Smolensky, 1993; Prince & Smolensky, 1997), although their solution to the logical problem of language acquisition differs from the generative solution (see Tesar & Smolensky, 1998).

Although it is seldomly recognized, the logical problem of language acquisition is not an unavoidable and theory-independent problem. Even if one agrees on the mentalist claim that the human brain contains a symbolical

representation of a mental grammar (an assumption which is refuted by radical connectionists) and that this mental grammar is at least as complex as a context-free grammar, the logical problem of language acquisition only arises as a result of two additional assumptions that are generally adopted within generative syntax.

The first assumption is the assumption that syntax forms an autonomous module of language. This assumption is fundamental to generative syntax. Hence, generative syntacticians like Lightfoot (1982) and Cook and Newson (1996) illustrate the logical problem of language acquisition by putting it on a par with trying to learn chess or snooker by watching people play the game. Crucial here is the fact that chess and snooker are systems of purely formal rules that do not refer to anything outside the system. The nature of the input and output of the process of learning has implications for the type of information available to the learner. As Grimshaw (1981) puts it: because of the autonomy of syntax, “UG does not permit deduction of a syntactic analysis from an analysis of the semantics of a phrase”. In other words, because syntax is autonomous, the formal properties of the grammar must be learnt from the formal properties of the input and cannot be inferred from its meaning.

The second assumption that is crucial to the view that there exists a logical problem of language acquisition is the identification of learning a language with finding the correct hypothesis through a process of hypothesis formulation and refutation (cf. Pinker, 1989; Wexler & Culicover, 1980). Note that this view of language acquisition as hypothesis testing is not present anymore in current nativist theories of language acquisition. I will return to this point in section 5. In the next section, I will demonstrate the dependence of the logical problem of language acquisition on the assumption that language acquisition is a process of hypothesis testing and logical deduction.

### **4. Language Acquisition as Logical Deduction**

As was noted in the previous section, an assumption underlying the logical problem of language acquisition is the assumption that the child learns her mother tongue through hypothesis formulation and refutation. Now if syntax is assumed to be autonomous, this process of hypothesis testing must be a process of logical reasoning. In particular, the process of hypothesis testing must involve logical deduction. Deductive reasoning involves deriving a conclusion from given information by using a set of formal (i.e., based on the form of the input) mental operations, without adding new information. This contrasts with inductive reasoning, which involves extrapolating a rule based on limited information. If it is not assumed that children employ deductive reasoning in hypothesis testing, there would be no logical problem of language acquisition at all, since nothing prevents children from concluding to the target language in the absence of negative evidence, except for the rules of logical deduction. Of course, it then remains to be explained how children arrive at exactly the same

grammar, but note that it is an empirical issue whether children indeed do.

Very few linguists actually discuss the mechanism that is supposed to lead children to conclude to the target language in the situation sketched by the logical problem of language acquisition. Linguists who use the poverty of the stimulus argument to support their theoretical point of view but do not discuss the learning mechanism involved, sometimes have been criticized for neglecting to take into account this mechanism. A common reaction to this criticism is that the actual mechanism does not really matter because Gold's (1967) proof, that positive evidence is not sufficient to learn a context-free language, is a formal proof. Such a formal proof is argued to be independent of the learning mechanism involved. However, implicit in Gold's proof is the assumption that learning a language can be identified with logically deducing the correct hypothesis on the basis of relevant evidence. Although it has been noted that there are some problems with Gold's model of language acquisition (Elman et al., 1996; Quartz & Sejnowski, 1997), this particular aspect of Gold's model has not been mentioned before as yielding a problem.

To motivate the claim that a process of hypothesis testing forms the basis of the logical problem of language acquisition, here are a number of quotations from the literature. According to Pinker (1989), for example, "[e]xplaining successful learning basically consists of showing that the learner can entertain and stick with a correct hypothesis and can falsify any incorrect ones" (p. 6). Chomsky (1988) likens the problem of language acquisition to the endeavour of a Martian scientist trying to understand Spanish, "pursuing the methods of the sciences, the methods of rational inquiry [...] His problem is to construct a hypothesis as to what the rule is and to test it by looking at more complex examples" (pp. 41-42). Perhaps children proceed exactly as this Martian scientist did in his inquiry, Chomsky continues. But this cannot be correct, since no negative evidence is available to children. Therefore, Chomsky concludes that innate principles must guide language acquisition. The motivation for assuming that certain properties of human language must be innately determined is explained by Crain (1991) as follows: "every child comes to know facts about language for which there is no decisive evidence from the environment. In some cases, there appears to be no evidence at all; in others the evidence is compatible with a number of alternative hypotheses (including false ones)" (p. 598). Jackendoff (1994) suggests that the unconscious task of a language-learning child can be compared with the conscious task of a linguist trying to discover the basic principles of human language: "they [i.e., children] must (unconsciously) discover for themselves the patterns that permit them both to understand these sentences and to construct new sentences for other people to respond to. Whether this process of discovery goes on unconsciously in the child or consciously in the linguist, the very same problems have to be solved" (p. 27). About the only way it can be explained that children are able to learn their

language is to assume that "children have a head start on linguists: children's unconscious strategies for language learning include some substantial hints about how a mental grammar ought to be constructed".

Summarizing, the basic idea of these authors is that a strategy of hypothesis testing is not sufficient for learning a natural language in the absence of negative evidence. This rejected strategy of hypothesis testing assumes children to behave like scientists and gather evidence in order to falsify incorrect hypotheses and employ hypothetico-deductive reasoning to draw the correct conclusions. However, the strategy of hypothesis testing by hypothetico-deductive reasoning seems to be based on implausible assumptions about human reasoning, as will be argued in section 6.

## 5. Language Acquisition as Parameter Setting

Many generative syntacticians will respond to the conclusion of the previous section by claiming that this is not a correct characterization of the current view on learning within the field of language acquisition. Rather, they will argue, language acquisition is currently viewed as a (blind) process of changing parameter values on the basis of specific triggering experience (cf. Gibson & Wexler, 1994). This is indeed true for the nativist framework of Principles and Parameters Theory and the Minimalist Program. However, the notion of parameter setting was introduced (along with the concept of an innate universal grammar) as a solution to the logical problem of language acquisition. Thus, first there was the logical problem of language acquisition, which made implicit use of the assumption that children employ logical deduction. This problem was subsequently solved by assuming an innate UG, which is accompanied by its own learning mechanism: parameter setting on the basis of triggering. Parameter setting therefore is the result of an argumentation that started out with the assumption that children employ logical deduction. If it would not have been assumed that children employ logical deduction, there would be no logical problem of language acquisition to be solved, and hence there would be no motivation for innate principles and for parameters that have to be set.

Discussions about learning mechanisms should be careful to distinguish between learning mechanisms assumed prior to the acceptance of the logical problem of language acquisition, and learning mechanisms assumed as a solution to this problem. This is not a trivial warning. When Lightfoot (1998) criticizes Elman et al. (1996) for their seeming lack of interest in the linguistic motivation for the innateness claim, in particular the poverty of the stimulus argument, and contrasts this with linguists, who seem to be interested in learning issues, Lightfoot is in fact already one step too far: "[l]inguists are actively interested in questions about learning algorithms. For example, an interesting debate is emerging about a trading relation between properties of UG and learning algorithms". Since the linguists Lightfoot refers to proceed from the conclusions drawn from the poverty of the stimulus argument, their

work does not bear on the innateness debate tackled in Elman et al. Rather, these linguists have already taken sides in the innateness debate, which makes it impossible to apply their results to the same debate again.

## 6. Do People Reason Logically?

One of the central themes within cognitive science is the question pertaining to the mechanisms underlying human reasoning. To investigate the role of logic and formal rules in the process of human reasoning, Wason (1966) and Griggs and Cox (1982), among others, carried out a series of well-known selection task experiments.

In Wason's experiment (Wason, 1966; Wason, 1968; Johnson-Laird & Wason, 1977), subjects were presented with an array of cards and told that every card had a letter on one side and a number on the other side. In addition, the following rule was given: "If a card has a vowel on one side, then it has an even number on the other side". Subjects were asked to select those cards that definitely have to be turned over to find out whether or not they violate the rule. Note that this rule has the form of a logical implication: *if p then q*. In propositional logic, such a rule is false if *p* is true and at the same time *q* is false. If the subjects in Wason's experiment would reason according to the rules of deductive logic, they would choose the cards with a vowel and the cards with an odd number. All other cards are irrelevant from a logical perspective. Indeed, most subjects chose the card with the vowel. On top of that, many subjects (46%) also chose the card with the even number, although it does not matter for the validity of the rule whether the other side of this card carries a vowel or a consonant. On the other hand, a card that was overlooked by almost all subjects was the card with the odd number. If there is a vowel on the other side of this card, the rule is violated. The correct answer, namely the card with the vowel and the card with the odd number, was given by only 4% of the subjects.

Griggs and Cox (1982) presented subjects with a task that was identical in form to Wason's task, but in which the abstract problem and the abstract rule had been replaced by a concrete problem and a concrete rule: "if a person is drinking beer, then he or she must be over 19 years of age". One side of the card had information about a person's age (16 or 19 years old) and the other side had information about what this person was drinking (a beer or a coke). If human reasoning takes place purely on the basis of the form of a problem, the results of the two experiments should have been identical. However, they were not. Like in Wason's experiment, all subjects turned over the card that affirmed the antecedent of the conditional clause (i.e., the card with "drinking a beer"). In contrast with Wason's experiment, however, many subjects (74%) also turned the card with "16 years of age", whereas almost none of Wason's subjects turned the card with the odd number. So, in Griggs and Cox's experiment most subjects gave the correct answer, namely the cards with "drinking a beer" and "16 years of age".

Apparently, then, the subjects in the second experiment used semantic information to solve the problem. In the first experiment, in which only information about the form of the problem was available, only few of the subjects managed to solve the problem. This suggests that human reasoning does not rely on a kind of mental logic. Logical rules are formal rules, which only take into account the form of the basic elements, not their meanings. If only information about the form of a problem is available, as in Wason's experiment, people make mistakes. Whenever they can, they use information about the meaning of the problem and about its context. In fact, people not only make mistakes in conditional reasoning tasks like the ones discussed above, they also make mistakes in other reasoning tasks requiring logical reasoning, for example in syllogisms. Conclusions that are consistent with beliefs or desires are more likely to be accepted as valid than conclusions that are not (e.g., Janis & Frick, 1943; Mayer, 1983). In general, people are not particularly good at solving problems correctly when the problems are presented to them in an abstract form. Concrete problems are solved by using all the knowledge that is available and might bear on the problem.

## 7. What Does Reasoning Tell Us about Language Learning?

The main conclusion that can be drawn from the reasoning experiments discussed in the previous section is that adults apparently are not naturally logical reasoners. But if adults do not reason according to some kind of mental logic, and if children do not differ from adults in this respect, then one of the two assumptions underlying the logical problem of language acquisition is not valid. As was argued in section 3 and 4, the logical problem of language acquisition only exists if it is assumed that syntax is autonomous and children reason logically. If one of these assumptions does not hold, then there is no logical problem of language acquisition. This does not imply that there is no empirical problem of language acquisition, though. But it implies that it is not *in principle* impossible that children learn their mother tongue from the language input they receive. Note that we cannot conclude from this that children definitely do not possess innate knowledge of language of some kind. But since the argument based on the logical problem of language acquisition appears to be invalid, the evidence for an innate UG is substantially weakened.

One could object that there is a difference between learning and reasoning and, therefore, that it is questionable whether results from the cognitive domain of reasoning apply to the domain of learning. However, human learning often involves deductive reasoning, in which one is able to discover or generate new knowledge based on beliefs one already holds. In addition, the logical problem of language acquisition is stated in such a way that it assumes children to reason about the hypotheses that are compatible with a given set of data and draw conclusions on the basis of the sentences they encounter. As an illustration, recall from section 4 Chomsky's comparison of a child learning her

language with a Martian scientist trying to understand Spanish. So, irrespective of whether learning and reasoning must be distinguished in practice, since learning involves reasoning in the logical problem of language acquisition, the results from the selection task experiments discussed in the previous section bear on the validity of the logical problem of language acquisition.

## 8. Conclusions

The logical problem of language acquisition is taken to be one of the strongest arguments in favor of the nativist view on language, since its validity is independent of specific empirical evidence. The basic idea behind this argument is that it is impossible in principle to acquire a language solely on the basis of the language input, irrespective of the presentation of the input data and the amount of positive feedback the child gets. In this paper, it was argued that there is no logical problem of language acquisition, since the combination of assumptions on which the formulation of the problem rests cannot be maintained in the light of current views on human reasoning and learning. In particular, people do not reason logically, which was the main assumption behind the logical problem of language acquisition. This does not imply that every aspect of language must be learnt from the input and that no innate linguistic knowledge or innate linguistic mechanisms can exist. But the evidence in favor of an innate UG must be based solely on empirical observations, now that the argument based on the logical problem of language acquisition has been shown to be invalid.

## References

- Brown, R., & Hanlon, C. (1970). Derivational Complexity and Order of Acquisition in Child Speech. In J.R. Hayes (Ed.), *Cognition and the Development of Language*. New York: Wiley.
- Chomsky, N. (1980). On Cognitive Structures and Their Development: A Reply to Piaget. In M. Piattelli-Palmarini (Ed.), *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. London: Routledge & Kegan Paul.
- Chomsky, N. (1986a). *Barriers*. Cambridge, MA: MIT Press.
- Chomsky, N. (1986b). *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.
- Chomsky, N. (1988). *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, MA: MIT Press.
- Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.
- Cook, V.J., & Newson, M. (1996). *Chomsky's Universal Grammar*. Oxford: Blackwell.
- Crain, S. (1991). Language Acquisition in the Absence of Experience. *Behavioural and Brain Sciences*, 14, 597-650
- Elman, J.L., Bates, E.A., Johnson, M.H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.
- Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, 25, 407-454.
- Gold, E.M. (1967). Language Identification in the Limit. *Information and Control*, 10, 447-474.
- Griggs, R.A., & Cox, J.R. (1982). The Elusive Thematic-Materials Effect in Wason's Selection Task. *British Journal of Psychology*, 73, 407-420.
- Grimshaw, J. (1981). Form, Function, and the Language Acquisition Device. In C.L. Baker & J.J. McCarthy (Eds.), *The Logical Problem of Language Acquisition*. Cambridge, MA: MIT Press.
- Jackendoff, R. (1994). *Patterns in the Mind: Language and Human Nature*. New York: BasicBooks.
- Janis, I.L., & Frick, F. (1943). The Relationship between Attitudes toward Conclusions and Errors in Judging Logical Validity of Syllogisms. *Journal of Experimental Psychology*, 33, 73-77.
- Johnson-Laird, P.N., & Wason, P.C. (1977). A Theoretical Analysis of Insight into a Reasoning Task. In P.N. Johnson-Laird & P.C. Wason (Eds.), *Thinking: Readings in Cognitive Science*. Cambridge, England: Cambridge University Press.
- Lightfoot, D. (1982). *The Language Lottery: Towards a Biology of Grammars*. Cambridge, MA: MIT Press.
- Lightfoot, D.W. (1998). Promises, Promises: General Learning Algorithms. *Mind & Language*, 13, 582-587.
- Mayer, R.E. (1983). *Thinking, Problem Solving, Cognition*. New York: W.H. Freeman and Company.
- Newport, E., Gleitman, H. & Gleitman, L.R. (1977). Mother, I'd rather do it myself: some effects and non-effects of maternal speech style. In C.E. Snow & C.A. Ferguson (Eds.), *Talking to Children: Language Input and Acquisition*. Cambridge: Cambridge University Press.
- Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, MA: MIT Press.
- Pinker, S. (1994). *The Language Instinct: The New Science of Language and Mind*. New York: William Morrow and Co.
- Prince, A., & Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. To appear in Linguistic Inquiry Monograph Series, Cambridge, MA: MIT Press.
- Prince, A., & Smolensky, P. (1997). Optimality: From Neural Networks to Universal Grammar. *Science*, 275, 1604-1610.
- Pullum, G. (1996). Learnability, hyperlearning, and the poverty of the stimulus. *Proceedings of the 22<sup>nd</sup> Annual Meeting: General Session and Parasession on the Role of Learnability in Grammatical Theory*, Berkeley Linguistic Society.
- Quartz, S.R., & Sejnowski, T.J. (1997). The Neural Basis of Cognitive Development: A Constructivist Manifesto. *Behavioral and Brain Sciences*, 20, 537-596.

- Sampson, G. (1997). *Educating Eve: The 'Language Instinct' Debate*. London: Cassell.
- Tesar, B., & Smolensky, P. (1998). Learnability in Optimality Theory. *Linguistic Inquiry*, 29, 229-268
- Wason, P.C. (1966). Reasoning. In B.M. Foss (Ed.), *New Horizons in Psychology*. Harmondsworth: Penguin Books.
- Wason, P.C. (1968). Reasoning about a Rule. *Quarterly Journal of Experimental Psychology*, 20, 273-281.
- Wexler, K. (1991). On the Argument from the Poverty of the Stimulus. In A. Kasher (Ed.), *The Chomskyan Turn*. Cambridge, MA: Blackwell.
- Wexler, K. (1999). Innateness of Language. In R.A. Wilson & F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA: MIT Press.
- Wexler, K., & Culicover, P.W. (1980). *Formal Principles of Language Acquisition*. Cambridge, MA: MIT Press.



# Memory-Based Problem Solving and Schema Induction in Go

Alex Heneveld, Alan Bundy, Michael Ramscar  
{heneveld,bundy,michael}@cogsci.ed.ac.uk  
Institute for Representation and Reasoning  
Division of Informatics  
University of Edinburgh  
Edinburgh, Scotland EH8 9LW

Julian Richardson  
julianr@cee.hw.ac.uk  
Department of Computing and  
Electrical Engineering  
Heriot-Watt University  
Edinburgh, Scotland EH14

## Abstract

This project presents a memory-based, analogical model of complex problem solving with a technique of schema formation. Cases in the game of Go are described in a predicate logic representation of spatial stone arrangements near recent moves on the board, and then structure-mapping (Gentner 1983) is used to suggest candidate moves in novel situations based on exemplar cases from expert games. The analogy process is also used to generalise across previous cases to form new schema cases. Problem solving using these prototype schemas is compared with the exemplar-only model. The exemplar run effectively found solutions to about 50% of the problems; schemas performed very similarly, taking half as long and identifying a few useful Go principles. This suggests to us that pure-exemplar models of memory-based processing can be made faster and more compact by introducing schemas. Analysing the model's weaknesses highlights the need for richer board representation and for a reminding stage to select relevant cases. Future work will also focus on using a move evaluation stage to filter spurious generalisations, and using both the evaluation and the generated schemas to enrich the representation.

## Introduction

This paper explores a model of memory-based problem solving to see whether analogical reasoning can be effective at suggesting solutions to complex problems and to see whether schema induction via analogy can serve as a basis for learning useful generalisations. We describe two machine learning experiments designed to test how well exemplars and abstracted schemas perform when used to suggest candidate moves in the game of Go, and review some computational and cognitive implications of this work.

In the memory-based psychological paradigm, experience cases stored in memory are the starting point for solving a target problem, roughly in a three step process.

- (1) *Reminding*: surface features prime cases
- (2) *Matching*: cases analysed to suggest solutions
- (3) *Evaluation*: solutions considered in context

Available features of the problem cue the retrieval of experiences containing similar features, in the first, reminding step. These potentially relevant experiences are then analysed — such as by looking for analogous propositional structure — and if it is a good match, a portion of the remembered experience may be transferred as a potential solution. (In iterative models, the matching step may uncover new surface features which cause new cases to be retrieved.) Potential solutions are evaluated relative to the context and the goals of the target problem, in the third step, and used to form the eventual solution. This approach is closely related to case-based reasoning, and is also apparent in exemplar models of categorisation (Nosofsky, 1984) and some recent work on natural language processing (Daelemans, van den Bosch, & Weijters, 1997).

A schema is a description of general experiences, often formed from a family of episodes with elements in common. They can supplement or organise a simple exemplar model by offering a concise source of the essential factors in many experiences without the incidental details present in episodic memories. Frames, scripts, and model-based reasoning are examples of their use in AI. The theory of pragmatic reasoning schemas (Cheng & Holyoak, 1984) is a clear account of how schemas can combine the best attributes of competing memory-based and rule-based views to model logical reasoning. In the categorisation literature, both prototype and theory (see Komatsu, 1992, for review) models can be considered as relying solely on a schema-definition of categories. Unfortunately in all these theories how to form these schemas is a difficult problem: the second set of experiments below tries a rudimentary method of inducing schemas from analogical matches performed during the memory-based process.

The strategy game Go was selected as a domain for our experiments for three main reasons: it is a plentiful source of difficult problems, many of which computers cannot currently solve; a vast amount of data is available on the Internet Go Server (IGS, 2000); and it does not involve much outside knowledge. The game is played by two players, one with black stones and the other with white, who take turns placing their stones in any unoccupied space on a 19x19 board, with the goal of amassing the greatest amount of territory. Players can capture their opponent's stones, individually or in orthogonally connected groups, by surrounding them in such a way that the captured group is not adjacent to any empty squares of the board. Captured groups are removed from the board, and the newly-unoccupied region typically becomes the capturing player's territory.

Many books can give more information on the rules, the strategy, and the history of the game (Bozulich, 1992, is good for this). Archives for research in Go-playing computer programs are on-line (Reiss, 2000), as are archives for psychological studies in Go (Burmeister, 2000).

One of these psychological studies (Saito & Yoshikawa, 1996) indicates that expert Go players quickly focus on their eventual move, (order of 200 ms in TsumeGo); and that they usually consider the outcome of only one or two possible moves. For these candidate moves, there is a lengthy lookahead evaluation which may go as far as 11 moves deep. Traditional Go-playing programs, even those which have been proposed as cognitive models, perform a broader search on a much greater number of candidate moves, necessarily to a lesser depth and with the result that no Go programs play any better than an amateur. It is fascinating that skilled players are able to focus rapidly — intuitively — on the best moves. In this research, we explore a memory-based model of how this might be done and also whether the abstraction of schemas can benefit performance.

### Experimental Design and Setup

In our view, the expert Go player relies on a large number of case experiences in memory, efficiently represented, and when a new problem is presented, retrieves a small number of possibly-relevant exemplars for fuller evaluation. Schemas may also be involved in representing common, recurring segments of exemplars as easily-accessible, general cases. The major issues involved in developing a computational cognitive model of this view are how the experiences are internally represented, how relevant experiences are selected from memory, and how schemas are formed.

### Representation

As in many machine models of complex problem solving, the representational format used here was a propositional description language recording a small number of perceptually basic, salient features. Specifically, for each problem, this encodes the colours of neighbouring stones around the two previous moves (X and O); the relative position between these two moves (in the format (rel X (rel-1 w 1)), meaning one position to the west of the last move); and, for cases including the expert’s solution, the relative position between the actual response (Q) and the two previous moves.

```

E1 (my-last-move O)
E2 (is-colour white O)
E3 (is-colour black (rel O (rel-1 w 1)))
E4 (is-colour black
    (rel O (rel-1 n 1)))
E5 (is-colour black
    (rel O (rel-1 s 1)))
E6 (is-colour white
    (rel O (rel-1 s 2)))
E7 (make-move Q)
E8 (is-colour white Q)
E9 (equal-position Q (rel O (rel-1 e 1)))

```



Figure 1: A fragment describing an atari opening east

This subset of information was selected because it corresponds generally to the initial observations that a player makes, guided by attentional cues to recent moves and nearby stones, and because as limited and easily-compiled as it is, it already contains enough information to begin drawing conclusions about where to play in certain circumstances. A more complete model of the expert’s initial representation might note a great many more features, symbolising more complicated concepts, but we hoped that this simplification would yet give promising results. The routines we developed to build the description from on-line game records and the routines to evaluate the descriptions as LISP code for visual output are designed to work with a family of vocabularies.

### Analogical Processing

The target problems were compared with cases in memory using the Structure-Mapping Theory of Analogy (Gentner, 1983). We chose this theory because it has been widely examined in the literature on the psychology and computer modelling of analogy, and because the Structure-Mapping Engine program (Falkenhainer, Forbus, & Gentner, 1989) is ideally suited to our representation. Routines in SME can easily read our descriptions, perform the analogy in keeping with a demonstrated psychological theory, score potential matches, and return inferences which correspond to candidate move suggestions.

In the implementation of our model, individual analogies are taken between a target problem and each of 1500 cases in memory. In practice, this is a slow, sequential process that takes between 10 and 30 minutes per problem on a Sun Ultra-10 workstation. In theory, however, each analogy is independent, and this might correspond to a very quick, parallel neural computation done by the brain. It could also be made significantly quicker by incorporating a more diverse description language with a “reminding” approach. A smaller number of cases which seem relevant can be primed by surface features in the target problem, with only those cases containing similar predicates used for the analogical matching (Forbus, Gentner, & Law, 1995). Whether these improvements could get the time per problem down to the order of 200 ms is unclear; up to 30 minutes per problem, though, is not prohibitive at this stage of the experimentation.

### Schema Induction

Once an analogical match has been made between a target problem and a source (base) case, the result can also be used to abstract the common substructure. Instead of looking only at the inferences, the process copies all the matching expressions into a new, schema case. This encodes the essential description elements from pairs of exemplar cases into more compact, abstracted descriptions which can then be used as base cases for solving future problems. Additionally, by repeating the induction on schema cases, the process can

also capture patterns recurring across many experiences in memory. This cognitive model of generalisation learning has been argued for on the basis of order effects by Kuehne, Forbus, & Gentner (1999). A similar algorithm, the Least General Generalisation (Plotkin 1971), has been widely used in AI to induce descriptions in logic, and a related approach has been applied to pattern learning in Go (Stoutamire, 1991). In our model, applying this type of schema induction after analogy has been used to generate candidate solutions has the advantage that the generalisation is returned with very little additional computational effort.

To perform the induction, we developed a LISP module within the SME package that transforms the result of a completed analogy into an abstracted schema. The schema uses the same description language as the parent cases, copying identical predicates exactly and inventing new tokens where the corresponding labels in the parents differ. The resulting schema can be used directly as a base case for analogical reasoning or output to a file. The second set of experiments reviewed below investigates the utility of this induction algorithm in solving complex problems and learning patterns.

### Problem Solving from Examples

The first experiment tried to solve 100 random Go problems by analogy to a library of exemplar cases. We take *solve* to mean that the move chosen by the expert player in the actual game is ranked in the top 50 in the program's list of suggestions (sometimes also top 3 or top 10). In a full-fledged Go-playing model this would be followed by an elaborate evaluation stage which would consider the effects of the candidate moves, typically using some form of lookahead search. Go players perform this lookahead on 1.5 moves on average (Saito & Yoshikawa, 1996), whereas most computer Go programs will evaluate between 20 and 70 moves.

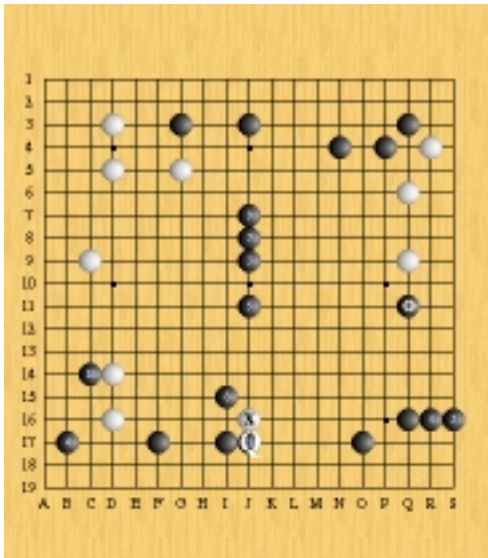


Figure 2: Sample results to a target problem. Numbers indicate ranked suggestions; X and O are the last two moves; and Q is the expert's move.

### Materials

We simulated the human's experience as consisting of 1500 exemplar cases drawn from ten tournament games (freely available on the Internet Go Server, IGS, 2000). These source cases, *ten-games*, are each a LISP-style description in the vocabulary outlined above, recording the neighbourhood of the last two moves with an indication of the move the expert made at that point in the game. One-hundred target problems, *query-random-100*, were compiled from random turns (before the end-game) in other IGS tournament games in the same manner but without any indication of the expert's response.

### Results

Figure 2 illustrates one of the better responses to a target problem. The program's top suggestion (1) is the expert's "right" answer (Q), and is quite close to the opponent's previous move (X). Many of the other candidates were in other sections of the board, reflecting problems locally similar to this one where the expert did play in other areas. (In this situation, playing elsewhere might be better, but as this model attends to stones near X and O, it cannot draw very good conclusions about distal play.)

The program solved 51 of the 100 problems after running for 13 hours (taking the top 50 suggestions). If all suggestions are considered, solutions to 93 problems were found; however, the program made an average of 6791 suggestions per problem. (There are only 361 positions on the board). There were a large number of repetitions — on average, the right answer is suggested 104 times. Looking at different suggestion depths gives a better picture of the program's performance: among the top 3 suggestions, the right answer was found for 7 problems; among the top 10, for 20; and among the top 50, for 51. Figure 3 shows the performance as up to 200 suggestions are considered.

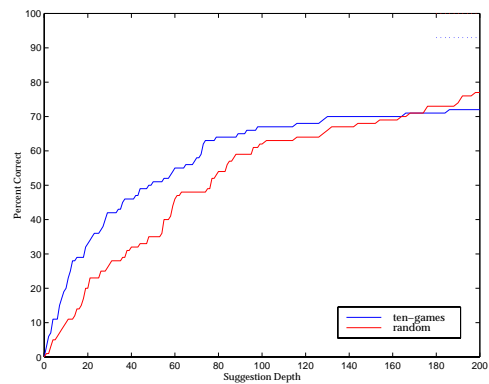


Figure 3: Experiment 1 results. This graph shows the percentage of problems solved by considering the  $x$  highest-scoring suggestions. The heavy line is our program, and the light line a chance player based on the 1970 Zobrist program. The dotted line is the asymptotic percent solved when all inferences were considered.

To put these numbers in perspective, we developed an informed random heuristic on the basis of the 1970 Zobrist program (Burmeister, 2000), also shown in Figure 3. This program selected positions at random, weighted heavily nearby the two prior moves. It took 6 minutes to query the same set, finding solutions to 3 of the 100 problems within its top 3 suggestions; 10 in its top 10; and 35 in its top 50. These numbers are not very sensitive to the precise weights used, so we take this as a baseline for how a rudimentary statistical learning algorithm would perform given only  $X$ ,  $O$ , and  $Q$ .

There were a large number of repeated suggestions in the lists of candidate moves, as well as quite a few invalid moves, either off the board or in an occupied square. A human player would immediately filter these out, and a machine evaluation routine would also quickly eliminate them from the lookahead search; it is interesting to review the effect this has. For our system, after removing these candidates, 13 of the correct solutions were in the top 3 suggestions, 32 in the top 10, and 64 in the top 50. For the informed random player, 5 were in the top 3 suggestions, 14 in the top 10, and 51 in the top 50.

### Analysis

A comparison of our analogical problem solver with the weighted random player shows that, for very small numbers of suggestions, our program performs much better, solving twice as many problems, though taking orders of magnitudes longer. At greater suggestion depths, the chance player improves relative to our program, for the trivial reason that will eventually guess every space on the board; considering more than 50 or 100 suggestions is not very practical either for input to a machine evaluation routine or as even a rough model of human candidate move generation. If we further focus on the source and target problems where the expert played within three stones of one of the last two moves (between 30 and 50 possibilities), the correct answer is in the top 10 different valid suggestions for 67 of the 100 problems. (This compares with 20 for the chance player). This shows that our analogical Go solver performs best on localized problems, which are in fact those instances where our solver has the largest amount of relevant information. This points to the fact that the representation was the biggest weakness when reasoning from exemplars.

### Schema-Based Processing

The second set of experiments was designed to explore the use of schemas in memory-based problem solving. Hundreds of thousands of pairs of cases from the *ten-games* set were passed to our Structure-Mapping Engine schema induction module, and the highest-scoring schemas (after normalisation) were kept as the set *schemas-1500*. This set was then used to solve the same selection of 100 random problems as in the previous experiments, using the same procedure as described above. Next, we examined the schemas which were most effective in solving problems and repeated

the induction process to investigate how well the generalisation technique captures the essential, common aspects in families of cases.

### Comparison with Exemplars

The most significant result with the schema-based run was that the computations took about half as long, achieving approximately the same success rate. On the same problem set (*query-random-100*), this run took 7 hours; in the top three suggestions, the *schemas-1500* source set found answers to 7 of the 100 problems; in the top 10, 23; and in the top 50, 51. In the limit, schemas suggested the solution to 88 problems.

Two main factors explain the speed difference. Firstly, the schema cases are much smaller, about 1/3 the size. Secondly, a much smaller number of suggestions were made per problem, 3204 on average. Nonetheless, the performance at low suggestion depths was about the same. In fact, as a percentage of the total number of suggestions, the correct move was suggested 60 more often by the schemas (77 of 3204) than by the exemplars (104 of 6791). This implies two things:

- When the schema set contained a case which solved the problem, it contained a lot of cases which solved the problem.
- Schemas were more likely to make suggestions to appropriate problems than were exemplars; *i.e.* they were less likely to give wrong answers.

The first point tells us that there was even more repetition in the schema set than in the exemplar set, which might have been expected, considering that the schemas encode common patterns among the exemplar set. The second point was quite surprising, though: one might expect the schemas to be more general, and hence more applicable than the exemplars. What happened, however, was that the exemplars, because they contained so much background information, could match more situations. With many exemplar cases, analogies were based on irrelevant criteria but were still strong enough to form inferences — inappropriately — about  $Q$ . Asymptotically, however, exemplars solved 93 of the 100 problems; in some cases, exemplars appropriately made inferences about  $Q$  based on information that had been lost in the schema formation process.

After filtering out repetitions and invalid moves, the schema-based run was slightly better than the exemplar-based run. The top three suggestions gave solutions to 14 problems (compared to 13 for exemplars); the top 10 solved 39 (versus 32); and the top 50 solved 70 (versus 64). This strengthens our conclusion that in this experiment, the schema induction process was somewhat successful in discarding irrelevant, distracting information from cases and achieving better performance much more quickly.

## The Effectiveness of Schemas

If this conclusion is correct, then many of the schemas should correspond to common Go situations or aphorisms. Figure 4 below shows one schema that was

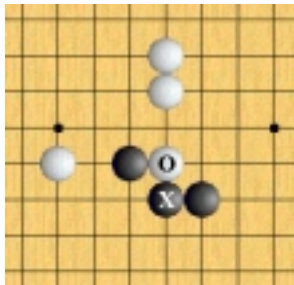


Figure 4: A particularly good schema. Playing Q to the left of X in a situation like this solved many of the target problems.

particularly useful: it gave correct suggestions to 20% of the posed problems, usually in the top 50 and several times in the top 10. Other effective schemas also helped to solve a large number of problems, to a much greater extent than individual exemplars. On the other hand, there were some problems which no schemas matched but which were matched closely and solved by some exemplars. In summary, it appears that some good schemas can effectively replace a large number of common exemplars, but that in outlying cases, exemplars are important to keep around.

An approach we are currently investigating is to lump exemplars and schemas together, developing new schemas at random (weighted by the SME match score) and adding them to the pool. Some of the high-generation schemas, *i.e.* those formed after multiple generalisations, match patterns in standard Go reference books. One of these is the principle to “hane at the head of two stones”, to jump out in front of an opponent’s line of stones (shown in the left of figure 5 below). Another interesting configuration left out the colour of X and Q, suggesting that whether black or white played X relative to the other two stones, then the other player should set Q down to the lower right of X.

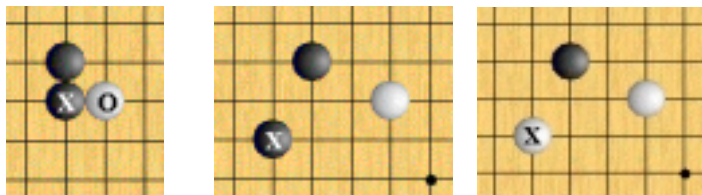


Figure 5: Some very generalised schemas. Interesting schemas found included the “hane”, on the left (play just below X), and the loose rhombi, both on the right (play to the lower right of X).

However, not all the high-generational schemas correspond to nicely stated principles or even to reasonable Go play. Anytime there is a large intersection between cases, even if it is meaningless and accidental, a schema

can form: for reliable, iterable induction, a better technique is needed.

## Conclusion

This memory-based model of the candidate move generation phase of Go has promising and interesting results. One of the three highest-scoring valid suggestions matched the expert’s move in one out of eight problems, and the best 50 suggestions solved more than half the problems despite a simple and locally-confined representation. Still, a large number of problems could not be solved by this model, and in considering future research directions it is useful to analyse these failings.

## Reminding and Richer Representations

Most of the problems that were not solved were ones where the subsequent move was in a different region of the board than either of the two previous moves. In these instances, the source exemplars simply did not store the information to enable our approach to find the answer. It seems likely that human players attend to a much wider range of features; if more of these could be recognised and encoded by the routines that generate the descriptions, we might expect better performance.

On the other hand, it is expensive to keep all this information: the cases would become too large to perform analogies on the entire set. One possible resolution would be to encode initially only the most salient features of the target; a reminding stage, such as in MAC/FAC (Forbus *et al.*, 1995), could select a small number of cases on which to perform the analogical matching. Some inferences would posit the presence of certain features in the target problem, which could be added to the initial representation if they hold, and the analogy could continue iteratively, re-representing, re-reminding, and re-matching, until it flounders or suggests something about where to play.

There is also the question of where these features will come from. Most Go programs have an extensive feature recognition routine, and it would be possible at first just to duplicate some of these. Ideally, these features could then evolve, with better ones developing as the program collects more experience. Schemas might be useful here, to replace common, structured phrases in the representation by a single reference to a schema, similar in a way to chunking. This model would be interesting to test, in Go, or in any domain where expertise might take the form of good feature recognition for case retrieval.

## Evaluation and Improving Abstraction

A major criticism of this particular approach to schema formation is that it only identifies patterns in static input descriptions. It does this without any regard for the significance of stones, and so encodes a lot of useless, coincidental substructures. An interesting AI perspective would be to grade cases and individual description lines according to their performance. A similar, more cognitive approach comes from Riesbeck & Schank (1989) who stress the importance of building logical explanations for generalisations to eliminate this sort of spurious abstraction. This is precisely what is done in the evaluation phase. While we have so

far ignored this phase as distinct and unrelated to representation and matching, it could conceivably be used to build explanations for good schemas; by storing this information as part of exemplar cases, it could also serve to enrich the representation.

### General Discussion

This model gives an example of how experiences and generalisations might be used, by people and by machines, to solve difficult problems. It essentially performs pattern matching using analogy, with good initial results in a very complex domain. Additionally, it embeds some powerful logic and machine learning techniques in a cognitive framework of schema induction. The major weaknesses of our model seem to be in the simplicity of the representation and the absence of the reminding and evaluation stages. These issues, sometimes considered separate from the matching stage, must on the contrary be addressed simultaneously and in depth when developing a memory-based problem solver.

Our approach can also be viewed as finding solution categories for a target problem, as analogical problem solving and categorisation are closely related. In this light, our results suggest that it is more efficient (in terms of time, memory, and to a lesser extent success rate) to define categories on the basis of schemas when there are many similar cases. This offers circumstantial evidence in favour of multiple-prototype and theory-based views of categorisation with the added benefit of describing how schema definitions might be formed from structural and surface features of exemplars. Instead of relying on a complete set of episodic exemplars, a memory-based approach can benefit from the clustering and compression given by analogical induction and the formation of schematic (semantic?) memories.

On the other hand, our schemas did not even suggest solutions to some of the problems that were easily solved by exemplars; forming good schemas, if it is possible in most cases, is more difficult than our technique recognises. No matter what, exemplars will always be needed for those areas where experience is minimal and where categories are not neatly defined. For Go, the data suggest that the best categorisation and problem solving would be achieved by a mixed source set containing a few very general schemas, more specific schemas, and exemplars in areas not well represented by the schemas.

In conclusion, we have implemented and analysed a model of memory-based cognition — in a symbolic architecture — and applied it to complex problem solving in Go, achieving better-than-chance performance with a very limited representation. At this stage, it seems that schemas can assist but not supplant pure exemplars in this type of problem solving. It seems also that the central matching stage may be more intricately dependent on the reminding and the evaluation stage than is typically acknowledged, particularly regarding the representation. This indicates compelling research directions both for Computer Go and for the psychology of problem solving.

### References<sup>1</sup>

- Bozulich, R., ed. (1992) *The Go Player's Almanac*. San Jose: Ishi.
- Burmeister, J. (2000) Research Page, <http://www.psy.uq.edu.au/~jay/>
- Cheng, P. W., & K. J. Holyoak. (1985) Pragmatic reasoning schemas, *Cognitive Psychology* 17:391-416.
- Daelemans, W., A. van den Bosch, & T. Weijters. (1997) Empirical Learning of Natural Language Processing Task, workshop position paper, The 9th European Conference on Machine Learning.
- Falkenhainer, B., K. D. Forbus, & D. Gentner. (1989) The structure-mapping engine: algorithm and examples, *Artificial Intelligence* 41:1-63.
- Forbus, K. D., D. Gentner, & K. Law. (1995) MAC/FAC: a model of similarity based retrieval, *Cognitive Science* 19:141-205.
- Gentner, D. (1983) Structure-mapping: a theoretical framework for analogy, *Cognitive Science* 7:155-170.
- Holyoak, K. J., & P. Thagard. (1989) *Mental Leaps*. Cambridge, MA: MIT Press.
- IGS (2000) The Internet Go Server. <http://igs.joyjoy.net/>
- Komatsu, L. K. (1992) Recent views of conceptual structure, *Psychological Bulletin* 112:500-526.
- Kuehne, S. E., K. D. Forbus, & D. Gentner. (1999) Category Learning as Incremental Abstraction using Structure-Mapping, poster presentation, 21st Annual Meeting of the Cognitive Science Society. Vancouver.
- Nosofsky, R. M. (1984) Choice, similarity, and the context theory of classification, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10:104-114.
- Plotkin, G. D. (1971) *Automatic Methods of Inductive Inference*, PhD Thesis, University of Edinburgh.
- Reiss, M. (2000) Mick's Computer Go Page, <http://www.reiss.demon.co.uk/webgo/compgo.htm>
- Riesbeck, C. K., & R. C. Schank. (1989) *Inside Case-Based Reasoning*. Hillsdale, NJ: Erlbaum.
- Saito, Y., & A. Yoshikawa (1996) A Protocol Study of Problem Solving in Go, Poster presentation, abstract in *Proceedings of the 18th Annual Conference of the Cognitive Science Society* 833.
- Stoutamire, D. (1991) Machine Learning, Game Playing, and Go, Technical Report TR 91-128. Cleveland, OH: Case Western Reserve University.

---

<sup>1</sup> In addition to the referees, we are indebted to Jon Oberlander, Ian Frank (now at ETL Japan), and others in the Edinburgh Informatics community for helpful discussions; and to Ken Forbus's group at Northwestern University for their insights and the SME code. The work was funded principally by a British Marshall Scholarship.

# Toward an Integrated Account of Reflexive and Reflective Reasoning

John E. Hummel (jhummel@lifesci.ucla.edu)  
Department of Psychology  
University of California Los Angeles  
405 Hilgard Ave.  
Los Angeles, CA 90095-1563

Jesse M. Choplin (choplin@lifesci.ucla.edu)  
Department of Psychology  
University of California Los Angeles  
405 Hilgard Ave.  
Los Angeles, CA 90095-1563

## Abstract

Some inferences are seemingly automatic (*reflexive*; Shastri & Ajjanagadde, 1993), whereas others require more effort (i.e., are *reflective*). We present the beginnings of an integrated account of reflexive and reflective reasoning, based on the LISA model of analogical reasoning (Hummel & Holyoak, 1997). The account holds that reflexive inferences are those that can be generated automatically based on existing knowledge in long-term memory, whereas reflective inferences require explicit structure-mapping and therefore demand greater attention and working memory. According to this account, reflexive inferences manifest themselves in the semantic encoding of objects and predicates, whereas reflective inferences manifest themselves as explicit propositions. In contrast to reflexive inferences, which are equally reflexive, reflective inferences may require more or less effort. We present preliminary simulation results demonstrating that both kinds of inference can be modeled in a single architecture for representing propositional knowledge.

## Reflexive vs. Reflective Reasoning

Some inferences are so effortless that we are barely aware of making them. Told that *Bill sold Mary his car*, you will infer that Mary now owns the car so automatically that Shastri and Ajjanagadde (1993) describe the inference as *reflexive*. Even more reflexive is the inference that Bill is probably an adult human male, and Mary an adult human female. Other inferences require more effort. Told that *Bill loves Mary* and *Mary loves Tom*, it is natural to infer that Bill is likely to be jealous of Tom, but this inference arguably requires a bit more reflection (and is less certain) than the inference that Mary is a woman. More effortful still are many kinds of inferences made in the context of scientific and mathematical reasoning, planning, and so forth. What is the relationship between reflexive inferences, such as *Bill is male* or *Mary owns the car*, and more *reflective* inferences, such as *Bill may be jealous of Tom*, or *matter and energy must be special cases of a common physical principle*? And what is the process by which

reflective inferences become more reflexive with experience? To a young child, it may not be immediately obvious that Bill's selling Mary his car implies that she now owns the car; but after a sufficient number of examples, the child will eventually induce a schema that makes the relationship between buying and owning reflexive (if evidenced only by the fact that the inference is reflexive for an adult).

In the literature on human cognition, the study of reflexive and reflective reasoning have been largely separate, with the former more common in the study of (for instance) story comprehension (e.g., Kintsch & van Dijk, 1978; Shastri & Ajjanagadde, 1993; St. John, 1992; St. John & McClelland, 1990), and the latter predominating in the study of problem solving (e.g., De Soto, London & Handel, 1965; Byrne & Johnson-Laird, 1989; Newell & Simon, 1976) and reasoning by analogy (e.g., Forbus et al., 1995; Gentner, 1983; Holyoak & Thagard, 1989; Hummel & Holyoak, 1997). Similarly, computational accounts of reflexive inference (e.g., Shastri & Ajjanagadde, 1993; St. John, 1992) have typically had little to say about more reflective forms of reasoning, and models of reflective (e.g., analogical) reasoning have had little to say about the nature of reflexive reasoning.

The most reflexive form of inference is encoding—inferring, for example, that "Mary" in "Bill loves Mary" is an adult human female. It is this most reflexive form of inference that has been most neglected in models of reflective reasoning. One consequence is that these models must be given, in full detail, the representations they are to use for reasoning. For example, the models of Forbus et al. (1995), Holyoak and Thagard (1989) and Hummel and Holyoak (1997) draw analogies between situations whose representations are fully specified for them. In contrast to human reasoners, who can read a sentence such as "Bill loves Mary, but Mary loves Tom" and infer the details for themselves (e.g., that Bill and Tom are adult human males, etc.), these models must be handed all this information for each analogy they are asked to solve.<sup>1</sup> One reason for this division between models of reflective and reflexive reasoning may be that the two kinds of reasoning obey different computational constraints, and therefore demand different kinds of algorithms. At the same time, however, both kinds of inference take place within the same cognitive architecture, and must operate on the same mental representations.

This paper presents the beginnings of an algorithmic account of the relationship between reflexive and reflective reasoning. In broad strokes the account holds that both kinds of reasoning require the capacity to dynamically bind

---

<sup>1</sup>One notable exception to this generalization is Hofstadter & Mitchel's (1994) CopyCat model, which solves analogy problems of the form X:Y::Z:?, and uses routines to change its representation of X, Y and Z in order to find the best possible analogy. In contrast to other models of analogy, CopyCat is not "stuck" with fixed representations of the elements of its analogies. At the same time, however, this model cannot simulate the kind of encoding discussed here, or the type of reflexive inferences discussed by Shastri & Ajjanagadde (1993).

variables to values (or equivalently, roles to fillers) in order to permit flexible (rule-like) generalization (cf. Shastri & Ajjanagadde, 1993, on the role of variable binding in reflexive reasoning; Holyoak & Hummel, 2000, and Hummel & Holyoak, 1997, on the role of variable binding in reflective reasoning). That is, both reflexive and reflective inferences are operations on symbolic representations. In addition, we propose that what makes reflective reasoning more effortful than reflexive reasoning is, at least in part, that the most reflexive inferences result from a kind of structured memory retrieval (i.e., retrieval that exploits and maintains variable-value bindings), whereas more reflective inferences require explicit structure mapping. That is, as illustrated in the simulations below, we propose that an inference will be fully reflexive when the to-be-inferred information is already available in long-term memory (LTM), and that it becomes progressively more reflective as the to-be-inferred information must be constructed on the basis of mapping large, multi-proposition structures.

The starting point for this effort is Hummel & Holyoak's (1997) LISA model of analogical reasoning, so we will briefly sketch that model's approach to knowledge representation and reflective inference (including memory retrieval, structure mapping, and schema induction). Mapping and retrieval are described in detail in Hummel and Holyoak (1997), and inference and schema induction are described in detail in Holyoak and Hummel (2000).

### The LISA Model

The core of LISA's architecture is a system for representing propositions in working memory (WM) by dynamically binding roles to their fillers, and encoding those bindings in LTM. LISA uses synchrony of firing for dynamic binding in WM (Hummel & Holyoak, 1992; Shastri & Ajjanagadde, 1993). Case roles and objects are represented in WM as distributed patterns of activation on a collection of *semantic units* (small circles in Figure 1); case roles and objects fire in synchrony when they are bound together and out of synchrony when they are not. For example, to represent the proposition *sell-to (Bill, Mary, car)* in WM, semantic units representing the *seller* role of the *sell-to* relation (e.g., *transaction*, *exchange*, etc.) fire in synchrony with units representing *Bill* while units representing the *buyer* role fire in synchrony with units representing *Mary*, and units for the *object* role fire in synchrony with units representing *car*. The three sets of units (*Bill+seller*, *Mary+buyer* and *car+object*) must be mutually de-synchronized with one another.

A proposition is encoded in LTM by a hierarchy of *structure units* (Figures 1 and 2). At the bottom of the hierarchy are *predicate* and *object* units (triangles and large circles, respectively, in Figure 1). Each predicate unit locally codes one case role of one predicate. For example, *seller* represents the first (seller) role of the predicate *sell-to*, and has bi-directional excitatory connections to all the semantic units representing that role; *buyer* and *sell-object*

represents the buyer and object roles, respectively, and are connected to the corresponding semantics. Semantically-related predicates share units in corresponding roles (e.g., *seller* and *giver* share many units), making the semantic similarity of different predicates explicit. Object units are like predicate units except that they are connected to semantic units describing things rather than roles. For example, *Mary* might be connected to units for *human*, *adult*, *female*, etc., whereas *car* might be connected to *object*, *vehicle*, etc.

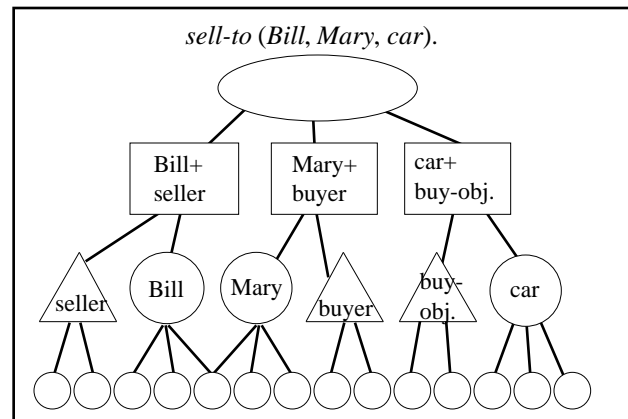


Figure 1. The LISA LTM representation of the proposition *sell-to (Bill, Mary, car)*.

*Sub-proposition* units (*SPs*; rectangles in Figure 1) bind roles to objects in LTM. *Sell-to (Bill, Mary, car)* would be represented by three *SPs*, one binding *Bill* to *seller*, one binding *Mary* to *buyer*, and one binding *car* to *sell-object*. *SPs* have bi-directional excitatory connections with the object and predicate units they bind together. *Proposition (P)* units (oval in Figure 1) reside at the top of the hierarchy and have bi-directional excitatory connections with the corresponding *SPs*. Complete, multi-proposition analogs (i.e., situations, events or schemas) are represented by collections of structure units (see Figure 2).

The final component of LISA's architecture is a set of *mapping connections* between structure units of the same type in different analogs. Every *P* unit in one analog may share a mapping connection with every *P* unit in every other analog; likewise, *SPs* share connections across analogs, as do objects and predicates. For the purposes of mapping and retrieval, analogs are divided into two mutually exclusive sets: a *driver* and one or more *recipients*. Retrieval and mapping are controlled by the driver. LISA performs retrieval and mapping as a form of guided pattern matching. As *P* units in the driver become active, they generate (via their *SP*, predicate and object units) synchronized patterns of activation on the semantic units (one pattern for each role-filler binding). The semantic units are shared by all propositions, so the patterns generated by one proposition tend to activate one or more similar propositions in LTM (retrieval) or in working memory (analogical mapping). Mapping differs from retrieval solely by the addition of the



modifiable mapping connections. During mapping, the weights on the mapping connections grow larger when the units they link are active simultaneously, permitting LISA to learn the correspondences generated during retrieval. These connection weights also serve to constrain subsequent memory access. By the end of a simulation run, corresponding structure units will have large positive weights on their mapping connections, and non-corresponding units will have strongly negative weights. Hummel & Holyoak (1997) showed that these operations account for a large body of findings in the literature on human analogical reasoning.

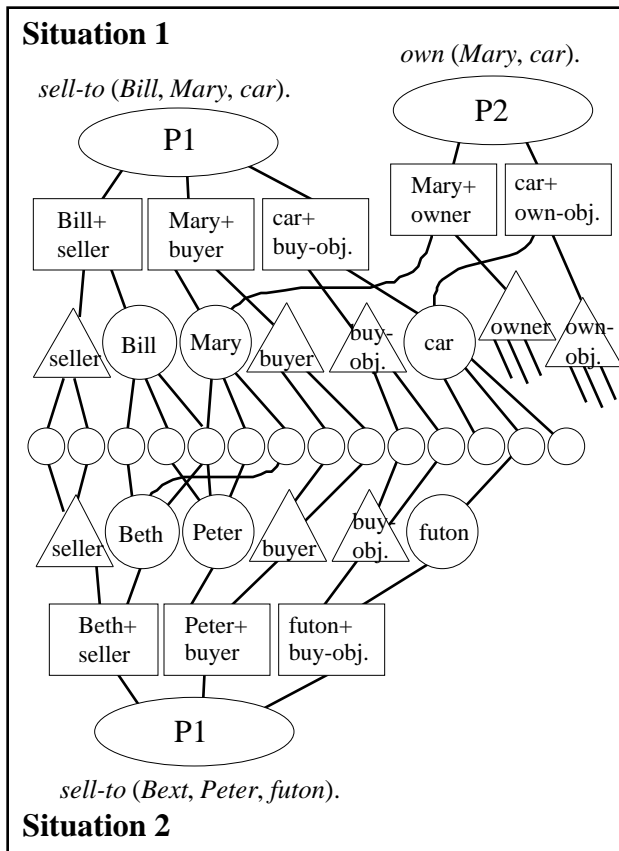


Figure 2. LISA LTM representation of *sell-to (Bill, Mary, car)* and *own (Mary, car)* (Situation 1; top) and *sell-to (Beth, Peter, futon)* (Situation 2; bottom).

Augmented with unsupervised learning and intersection discovery, LISA's approach to mapping supports inference and schema induction as a natural consequence (Holyoak & Hummel, 2000). Consider an analogy between two situations (Figure 2): In situation 1, Bill sells his car to Mary (proposition P1), so Mary now owns the car (P2); in situation 2, Beth sells her futon to Peter (P1), but there is no explicit statement that Peter now owns the futon. During mapping, corresponding elements in the two analogs will become active simultaneously. For instance, *sell-to (Bill, Mary, car)* in the driver, will activate *sell-to (Beth, Peter, futon)* in the recipient, so corresponding elements

(such as *Bill* and *Beth*) will fire in synchrony with one another, and non-corresponding elements (e.g., *Bill* and *futon*) will fire out of synchrony. As a result, LISA learns mapping connections from *Bill* to *Beth*, *Mary* to *Peter*, and *car* to *futon*. Likewise, the roles of *sell-to* in situation 1 map to the corresponding roles of *sell-to* in situation 2. However, nothing in situation 2 maps to the roles of *owns* in situation 1. Therefore, when *owns (Mary car)* fires in situation 1, LISA will build units in situation 2 to correspond to the structures in situation 1 representing that proposition: It will build units corresponding to *owner* and *owned*, and connect them to the semantic units representing those roles; it will build SPs corresponding to *owner+Mary* and *owned+car*, and connect them to *owner* and *Peter* and *owned* and *futon*, respectively; finally, it will also build a P unit corresponding to the whole proposition, connecting it to the newly created SPs. (LISA "knows" what to connect to what simply by virtue of which units are firing in synchrony with one another; see Holyoak & Hummel, 2000.) That is, it will infer that Peter now owns the futon.

The same operations permit LISA to perform schema induction in a third "schema" analog. Although we have described the activation of semantic units only from the perspective of the driver, recipient analogs also feed activation to the semantic units. The activation of a semantic unit is a linear function of its input, so any semantic unit that is common to both the driver and recipient will receive input from both and become roughly twice as active as any semantic unit receiving input from only one analog. Shared semantic elements are thus tagged as such by their activations. These shared elements are encoded into the schema by the same unsupervised learning algorithm that performs analogical inference: Units in the schema connect themselves to semantic units and to one another based on their co-activity. Because the learning algorithm is sensitive to the activations of the semantic units, object and predicate units in the schema preferentially learn connections to the semantic that are common to—the intersection of—the corresponding units in the known situations. In the case of the current example, the induced schema would be roughly *sell-to (person1, person2, object)*, and *own (person2, object)*.

## Extension to Reflexive Reasoning

As described above, LISA is a model of reflective reasoning that makes inferences about novel situations based on explicit analogies (i.e., structure-mappings) to familiar situations. However, the operations it uses for analogy, inference and schema induction—most notably, the feedback from the recipient analog to the semantic units (henceforth *recipient feedback*)—suggest themselves as the beginnings of an account of reflexive inference. The basic idea is to use the recipient feedback from structures in LTM (including both general schemas and specific situations) to encode the semantic representation of predicates and objects in the driver. That is, encoding is seen as a collection of reflexive inferences about the properties of the predicates and objects.

Using the recipient feedback in this way solves only one of several problems that must be solved in order to provide a general integrated account of reflexive and reflective reasoning; however, the simulations reported here suggest that it is a useful first step.

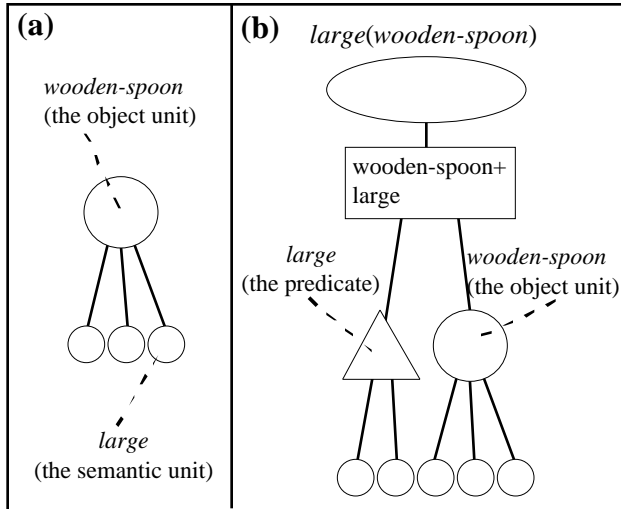


Figure 3. Ways to represent the fact that wooden spoons are large in LISA: (a) *large* as a semantic feature connected to the object unit *wooden-spoon*; (b) *large* as a predicate in the proposition *large (wooden-spoon)*.

Consider the concept of a wooden spoon. The statement "wooden spoon" makes exactly two properties of the object *wooden spoon* explicit: it is wooden and it is a spoon. But upon encountering these properties, it is irresistible to infer additional properties, such as that it is large (as spoons go), it is more likely to be used for cooking than for eating, etc. (cf. Medin & Shoben, 1988). These properties may be represented in two qualitatively different ways: as semantic "features" of wooden spoons, or as explicit propositions. In terms of the LISA architecture, these ways of representing the properties of wooden spoons are, respectively, as connections from the *wooden-spoon* object unit to semantic units for *large*, *cooking*, etc., (Figure 3a) and as full propositions, complete with predicates, SPs and P units (Figure 3b). We hypothesize that the former type of (semantic feature) representation is established reflexively, as an automatic part of encoding the representation of wooden spoons, whereas the latter (propositional) form is established more reflectively, by thinking explicitly about the properties of wooden spoons. It is important to note that inferring the properties of wooden spoons (either reflexively or reflectively) it is not a simple matter of replacing the default value of the *material* slot in the spoon schema (i.e., *metal*) with the value *wooden* (e.g., as suggested by Smith & Osherson, 1984; Smith et al., 1988), because the attributes of spoons (i.e., the values bound to the slots of the schema) are correlated: Other attributes such as *size=large* will have to be inferred,

and these attributes will have to replace the corresponding default values in the schema.

We simulated the reflexive form of this inference as follows. We generated five situations (analog), one corresponding to a "new" situation (thinking about a wooden spoon; analog 1), and the other four corresponding to various schemas in LTM. Analog 1 consists of the single proposition *exist (wooden-spoon)*, where the object *wooden-spoon* is connected to a single semantic unit, *wooden-spoon*, which serves to represent the type *wooden spoons*; the predicate *exist* is not connected to any semantic units. (*Exist* is a vehicle for instantiating *wooden-spoon* in a proposition so that it can be activated—it allows LISA to "think about" wooden spoons devoid of any particular context.) Analog 2 is a schema for wooden spoons consisting of two propositions: *wooden (wooden-spoon)* and *big (wooden-spoon)*. The object unit *wooden-spoon* has positive connections to the type semantic *wooden-spoon* and to semantics for *utensil*, *spoon*, *wooden*, and *big*. It has negative (inhibitory) connections to semantics for *metal*. The predicate *wooden* has positive connections to semantics for *material* and *wood*, and an inhibitory connection to *metal*; *big* excites semantics for *size* and *big*, and inhibits *small*. Analog 3 is a schema for metal spoons, and consists of *metal (metal-spoon)* and *small (metal-spoon)*. The semantic representations of *metal-spoon*, *metal* and *small* are analogous to those of *wooden-spoon*, *wooden* and *big*, respectively, except that the appropriate semantics are reversed (e.g., the predicate *small* is connected to the semantic *small* rather than *big*, etc.). Analog 4 is a schema for spoons in general, and consists of the propositions *utensil (spoon)* and *concave (spoon)*. Analog 5 is a schema for horseback riding, consisting of the single proposition *ride (horse)*. Analog 5 serves as a foil to ensure that the model will not simply activate all knowledge in LTM in the course of drawing reflexive inferences about the (semantically empty) wooden spoon in analog 1.

The goal of the simulation is to activate *exist (wooden-spoon)* in analog 1 and observe which schemas it activates in LTM, and whether the recipient feedback from the objects and predicates in those schemas allow *wooden-spoon* in analog 1 to learn an appropriate semantic encoding. We therefore set analog 1 to be the driver, and left analogs 2 - 4 "dormant" in LTM (see Hummel & Holyoak, 1997). The proposition *exist (wooden-spoon)* was then fired, and propositions in LTM were allowed to respond, feeding activation back to the semantic units. When *exist (wooden-spoon)* first fired, both *wooden (wooden-spoon)* and *big (wooden-spoon)* became active in analog 2 (the wooden spoon schema). *Exist* is semantically empty, so the only semantic feature of analog 1 that activated anything in analog 2 is the type semantic *wooden-spoon* (which is shared by the object *wooden-spoon* in analog 2). As a result, *wooden (wooden-spoon)* and *big (wooden-spoon)* became equally active in analog 2. The feedback from these propositions to the semantic units began to activate other schemas in LTM: the semantics *utensil* and *spoon* activated units in analogs 3 (the metal spoon schema) and 4 (the

generic spoon schema). At the same time, the object *wooden-spoon* (in analog 2) *inhibited* the semantics for *metal*. This inhibition propagated into analog 3 (the metal spoon schema), preventing that schema from becoming active and in turn, preventing it from activating its own semantics. When the pattern of activation settled, analogs 2 and 4 were fully active (i.e., both propositions were active in both analogs), along with all the semantic units to which they are connected. As a result, the object *wooden-spoon* in analog 1 learned connections to the semantics for *utensil*, *spoon*, *wooden* and *big* (due to the feedback from the wooden spoon schema), and to *concave* and *utensil* (based on the generic spoon schema): The model reflexively inferred the semantic properties of the wooden spoon.

In a second simulation, analog 1 consisted of the proposition *exist (metal-spoon)*—this time having LISA "think about" metal spoons—and we ran the same operations described above. This time, analog 3 (the metal spoon schema) and analog 4 (the generic spoon schema) became active, and the model inferred the properties of the *metal-spoon* in analog 1: *metal-spoon* learned connections to the semantic units for *utensil*, *spoon*, *metal* and *small* (due to the feedback from the metal spoon schema), and to *concave* and *utensil* (based on the generic spoon schema). In both these simulations, it is interesting to note that LISA assigned each object (the metal spoon or the wooden spoon) to the most general category appropriate (by activating the generic *spoon* schema), but it did not categorize metal spoons as wooden spoons, or vice versa. As a result, it made appropriate inferences about the objects at multiple levels of abstraction (e.g., that the wooden spoon would be big [which is specific to wooden spoons] and that it would be concave [which is general to all spoons]).

In the previous simulations, the inferences were purely reflective, in the sense that we did not allow LISA to retrieve the schemas from memory and map them back onto analog 1. When we allowed the model to reflect on the properties of wooden spoons—by making the wooden spoon schema the driver, the wooden spoon version of analog 1 the recipient, and allowing it to explicitly map the schema onto analog 1—it inferred the explicit propositions *wooden (wooden-spoon)* and *big (wooden-spoon)* in analog 1. (It did do by exactly the same operations described previously in the discussion of LISA's operation.) Similarly, when we allowed it to reflect on the fact that wooden spoons are spoons—by mapping the spoon schema into analog 1—it inferred the propositions *utensil (wooden-spoon)* and *big (wooden-spoon)*. But importantly, it did not infer any of these propositions until it explicitly brought the corresponding schema into WM and mapped it onto analog 1. This property is interesting in combination with the model's ability to reason reflexively to the most generic category applicable (e.g., to assign wooden spoons semantic features that are true of all spoons based on the generic spoon schema): Together, they predict that reflexive inferences—which manifest themselves in the (implicit) semantic encoding of an object or predicate—will automatically take place across multiple levels of category

abstraction, whereas reflective inferences—which cause the construction of explicit propositional structures—will only take place when the reasoner explicitly reflects on the fact that the object belongs to the category (i.e., explicitly maps the category schema onto the object). To our knowledge, no one has yet tested this prediction of the model.

## Discussion

Using simple operations already in place to simulate reflective analogy-based inference—namely, recipient feedback and unsupervised learning—LISA was able to reflexively infer the meaning of "wooden spoon" and "metal spoon" based on examples in LTM. These inferences were reflexive in the sense that they did not require the model to explicitly map the structures in the new example (analog 1) to the structures in LTM. Instead, they were drawn in the course of what is analog retrieval in LISA (i.e., the process of retrieving a source analog or schema from LTM given a novel target as a cue; see Hummel & Holyoak, 1997). By the end of the first two simulations, the objects *wooden-spoon* (in the first simulation) and *metal spoon* (second simulation) had semantic encodings that were richer than what was provided at the beginning of the simulation. In each case, the object unit started with a single semantic feature (*wooden-spoon* or *metal-spoon*) and ended with a semantic encoding specifying its size (*big* or *small*), material (*wooden* or *metal*), shape (*concave*) and use (*utensil*). However, in neither of the first two simulations did analog 1 end up with any new propositions. By contrast, in the third and fourth simulations, when the schemas were called into WM and allowed to map to analog 1, the model inferred propositions that explicitly stated the properties of the wooden spoon. According to the present account, inferring a new proposition (e.g., one stating explicitly that the wooden spoon is big) is a reflective process that requires retrieval of a schema (or specific situation), and an explicit mapping of the structures in that schema to the structures in the new example.

In this respect, our use of the term "reflexive" is somewhat more restrictive than Shastri & Ajjanagadde's (1993). On our account, an inference such as "Mary now owns the car" is not strictly reflexive unless it is represented strictly as features in the *semantic* representation of Mary (i.e., as connections from the unit *Mary* to units representing *ownership*). If instead (or in addition) the inference is represented as an explicit proposition (*own (Mary car)*), we would classify it as "reflective but easy" (as noted previously, some reflective inferences are easier than others). This distinction between our account and that of Shastri & Ajjanagadde stems primarily from the fact that LISA represents objects and predicates as distributed patterns of activation in WM, which precludes binding an object to more than one predicate role at a time (see Hummel & Holyoak, 1997). As a result, LISA makes a strong distinction between properties *qua semantic features* and properties *qua explicit propositions*. By contrast, Shastri & Ajjanagadde's model represents each object or

predicate as a localist unit, making it possible to "stack" predicates on objects, effectively representing multiple predicate-object bindings (i.e., multiple propositions) in parallel (cf. Hummel & Holyoak, 1997). Whether the human mind makes a strong distinction between features and propositions (like LISA), or permits "stacking" of predicates (like Shastri & Ajjanagadde's model) is an empirical question.

## The Origins of Object Features

To this point our discussion of reflexive inference—based on learning connections to features activated by feedback from structures in LTM—has begged a major question: If objects learn the features that describe them by "comparing themselves to" other objects in LTM, then how did the objects in LTM learn the features that describe themselves in the first place? Answering this question "by comparing themselves to objects in LTM when *they* were encoded" is unsatisfying because it brings to mind the infinite regress, "and where did *those* objects learn *their* features?" etc. Although we are far from being able to provide a complete answer to this very difficult question, one aspect of the model that we have not yet discussed may provide a partial answer. Specifically, we allow some semantic features that belong to predicate units to attach themselves to object units during the course of reflexive inference. (In the original LISA, predicate semantics were attached strictly to predicates and object semantics strictly to objects [see Hummel & Holyoak, 1997]; this approach is a departure from that convention.) As a result, roles to which an object is attached in many situations or schemas in LTM (e.g., the role *big*, in the case of *big* (*wooden-spoon*)) can become attached directly to new instances of those objects as semantic features. In this way, an inference like *own* (*Mary*, *car*) can become truly reflexive in the sense of the definition suggested here: If, in several examples in LTM, the buyer of a product is also represented as the owner of that product (i.e., in a separate *own* (*person*, *object*) proposition), then the semantics of the predicate *own* will tend to become attached as semantic features of subsequent objects bound to the *buyer* role of a *sell-to* relation. We have yet to work out fully the details of this proposal, but preliminary simulations have so far been very promising.

## References

- Byrne, R. M. J., & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory and Language*, 28, 564-575.
- DeSoto, C., London, M., & Handel, S. (1965). Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology*, 2, 513-521.
- Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19, 141-205.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.
- Hofstadter, D. R., & Mitchell, M. (1994). An overview of the Copycat project. In K. J. Holyoak & J. A. Barnden (Eds.), *Advances in connectionist and neural computation theory, Vol. 2: Analogical connections*. Norwood, NJ: Erlbaum.
- Holyoak, K. J., & Hummel, J. E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Dietrich and A. Markman (Eds.), *Cognitive Dynamics: Conceptual Change in Humans and Machines*. Hillsdale, NJ: Erlbaum.
- Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, 13, 295-355.
- Hummel, J.E., & Holyoak, K. J. (1992). Indirect analogical mapping. *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, pp 516 - 521.
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104, 427-466.
- Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363-394.
- Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, 20, 158-190.
- Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19, 113-126.
- Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings. *Behavioral and Brain Sciences*, 16, 417-494.
- Smith, E. E. & Osherson, D. N. (1984). Conceptual Combination with prototype concepts, *Cognitive Science*, 8, 337-361.
- Smith, E. E., Osherson, D. N., Rips, L. J. & Keane, M. (1988). Combining prototypes: A selective modification model, *Cognitive Science*, 12, 485-527.
- St. John, M. F. (1992). The Story Gestalt: A model of knowledge-intensive processes in text comprehension. *Cognitive Science*, 16, 271-302.
- St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46, 217-257.

## Acknowledgments

This research was supported by NSF Grant SBR-9729023, and by grants from the UCLA Academic Senate and HRL Laboratories.

# Constituent Structure in Mathematical Expressions

**Anthony R. Jansen** ([tonyj@csse.monash.edu.au](mailto:tonyj@csse.monash.edu.au))  
School of Computer Science and Software Engineering  
Monash University, Victoria, Australia

**Kim Marriott** ([marriott@csse.monash.edu.au](mailto:marriott@csse.monash.edu.au))  
School of Computer Science and Software Engineering  
Monash University, Victoria, Australia

**Greg W. Yelland** ([Greg.w.Yelland@sci.monash.edu.au](mailto:Greg.w.Yelland@sci.monash.edu.au))  
Department of Psychology  
Monash University, Victoria, Australia

## Abstract

Previous research has suggested that human perception of mathematical expressions is based on syntactic structure. Here, we extend our understanding of how humans perceive algebraic equations in two ways. First, we examined the hypothesis that the internal representation used by experienced mathematicians is based on the phrasal structure of the parse tree. This was tested using a memory recognition task, and the results supported the hypothesis. Second, we explored how much experience with mathematics is necessary before such representations become established. Participants were young students with very little experience with algebra. Surprisingly, the students appeared to encode equations in a manner similar to experienced mathematicians.

## Introduction

Mathematical notation and natural language share many common features. Both have a well-defined syntax and semantics, and both allow for the expression of abstract information. However, an important difference is that the layout of mathematical notation is two dimensional in nature, with equations relying on both vertical and horizontal adjacency relationships between the symbols to provide the meaning. It is natural then to ask how humans comprehend mathematical expressions.

The present paper extends our understanding of how humans perceive mathematical expressions in two ways. First we explore the nature of the internal representation used to encode equations. Specifically, we examine whether the information recovered from equations has a parse tree structure similar to that used to represent sentences of natural language. Second, we explore how much experience with mathematics is necessary before such representations become established.

For many years now, phrase structure grammars have been used to understand the way that humans parse natural language sentences (for example, see Akmajian, Demers and Harnish, 1984). This allows the constituent structure of a natural language sentence to be represented diagrammatically by a parse tree, containing various phrases such as noun and verb phrases. Although phrase structure grammars can only

be applied to sequential languages, variations of such grammars have been proposed in an attempt to enable computers to understand mathematical notation (for example, see Anderson, 1977). Analogously to natural language, parse trees for equations can be created based on mathematical syntax. An example parse tree is given in Figure 1.

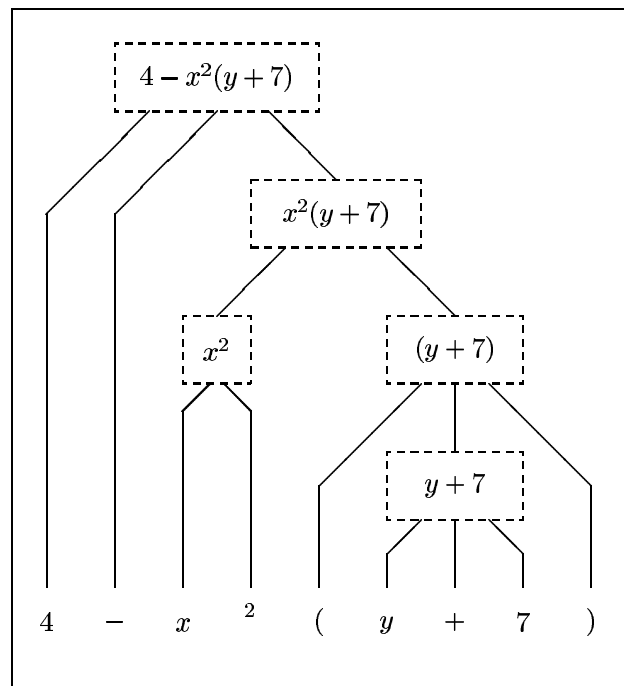


Figure 1: Parse Tree for  $4 - x^2(y + 7)$

Our previous work on the comprehension of mathematical expressions (Jansen, Marriott and Yelland, 1999) has shown that in a memory recognition task, experienced users of mathematics can more readily identify those parts of a previously seen equation that are syntactically well-formed (that is, which have a coherent mathematical meaning, such as  $y + 7$  in the above example), than those that are not well-

formed (for example,  $-x^2(y$  which is also part of the equation, but does not convey any coherent mathematical meaning on its own). This result provides support for the notion that the internal representation used by mathematicians is based on mathematical syntax. This accords with results by Johnson (1968, 1970) which show that in the context of natural language, chunking of sentences is also guided by syntax. The work of Ranney (1987) also shows that even after only brief exposure, the structure of algebra expressions provide information about the category of the symbols in that expression (whether they are variables, numbers, operators, etc.). This indicates that the parsing of such expressions is based on structural content.

## Experiment 1

It is clear that mathematical syntax plays an important role in encoding equations, however it does not necessarily follow that a parse tree structure underlies the internal representation. The first experiment explores this idea with respect to moderately complex algebra expressions. Our hypothesis is that the internal representation used by experienced mathematicians is based on the phrasal structure of the parse tree. To test this hypothesis, we have set up a recognition task to see if participants can more readily recognize sub-expressions of an equation that form a phrasal node on the parse tree (for example,  $y + 7$  in the previous example) as opposed to sub-expressions that are also syntactically valid, but do not form a phrasal node on the parse tree (such as  $4 - x^2$ ). If our hypothesis is correct, we would expect to see a recognition advantage for the phrasal sub-expressions.

### Method

**Participants** Twenty-four participants successfully completed the experiment. All were staff members, graduate or undergraduate students from the Computer Science department, all competent mathematicians who were experienced with algebra. All participants were volunteers between the ages of 18 and 35 years, with normal or corrected-to-normal vision. Data from an additional 8 participants were not included due to excessive error rates.<sup>1</sup>

**Materials and Design** Seventy-five equations were constructed, all consisting of between twelve and fourteen characters. The equations contained at most one fraction and the variable names were  $x$  and  $y$ , since these are most commonly used. For each equation, sub-expressions of three types were constructed.

- a) A *phrasal sub-expression*, which is a syntactically well-formed component of its equation, which conveys the same meaning on its own that it conveyed in the equation. It is a phrasal node in the equation's parse tree.

- b) A *non-phrasal sub-expression*, which is also a well-formed component of its equation, but does not convey the same meaning on its own that it conveyed in the equation. It is not a phrasal node in the equation's parse tree.

- c) An *incorrect sub-expression*, which was not part of the original equation. It is also a well-formed expression. These act as fillers.

Each of the sub-expressions contained between four and six characters (the average for phrasal sub-expressions was 4.78; for non-phrasal, 4.54; for incorrect, 4.60). See Table 1 for examples of equations and sub-expressions used. As the examples show, a variety of sub-expressions were used, some of which were bracketed, but most of which were not.

In order to present all three sub-expression types for each equation, but ensuring that participants were presented with each equation only once to avoid practice effects, three counterbalanced versions of the experiment were constructed. For each version, there were twenty-five instances of each type of sub-expression. Two additional equations were constructed as practice items. The same practice items were used in each version. Eight participants completed each version, each receiving the items in a different pseudo-random order.

**Procedure** Participants were seated comfortably in an isolated booth. Items were displayed as black text on a white background on a 17" monitor at a resolution of 1024x768, controlled by an IBM compatible computer running a purpose designed computer program. The average width of the equations in pixels was 187 (range 91–244) with an average height of 45 (range 26–59). The average width of the sub-expressions in pixels was 74 (range 25–111) with an average height of 23 (range 16–52).

Participants were given a statement of instructions before the experiment began. Practice items preceded the experimental items, and the participants took approximately fifteen minutes to complete the task. Progress was self-paced, with participants pressing the space bar to initiate the presentation of each trial.

Each item was presented in the centre of the monitor in the following sequence. First, a simple algebra equation was shown to the participant for 2500ms. The equation then disappeared and the screen remained blank for 1000ms. Then the sub-expression was shown, remaining on the screen until a response was made. The participant was required to decide whether the sub-expression was in that equation, responding via a timed selective button press. They pressed the green button, (the '/' key on the right side of the keyboard), to indicate that the sub-expression was part of the original equation, and the red button, (the 'Z' key on the left of the keyboard), to indicate that the sub-expression was not part of the original equation. Participants were instructed to respond as quickly as possible, while taking care not to make too many errors.

The response time recorded was the time between the onset of the sub-expression and the participant's response. After the response, the participant received feedback. If the re-

<sup>1</sup>Data from participants with an overall error rate of over 30%, or making in excess of 50% errors for any given sub-expression type, were excluded from the final analysis.

Table 1: Example equations and sub-expressions used in examining phrasal properties.

Equation	Sub-Expression		
	Phrasal	Non-Phrasal	Incorrect
$y = 8 + \frac{8y - 9x^2}{7y^6}$	$8y - 9x^2$	$8y - 9$	$8x + 9$
$3 - \frac{5}{7 - x^2(8 - y)}$	$(8 - y)$	$7 - x^2$	$2 - x^4$

sponse was correct then the word ‘‘CORRECT’’ appeared on the screen. Otherwise, the word ‘‘INCORRECT’’ appeared on the screen. In both cases, the participant’s response time in milliseconds also appeared on the screen.

**Data Treatment** Two measures were employed to reduce the unwanted effects of outlying data points. Absolute upper and lower cut-offs were applied to response latencies, such that any response longer than 2500ms or shorter than 500ms was excluded from the response time data analysis and designated as an error. Secondly, standard deviation cut-offs were applied, so that any response time lying more than two standard deviations above or below a participant’s overall mean response time was truncated to the value of the cut-off point.

It was necessary to exclude two items from the final analysis due to error rates in excess of 75%. One further item also had to be removed in order to balance the number of items in each version of the experiment. As a result, the final analyses were over twenty-four items per condition, not the original twenty-five. Response time and error data were analysed by a series of analyses of variance (ANOVAs), over both participant and item data. Where both the subject-based and item-based analyses were significant they were combined in the *minF*’ statistic to ensure the generalisability of results over both these domains (Clark, 1973).

## Results and Discussion

The mean correct response time (in milliseconds) and error rate for the three sub-expression types are summarised in Table 2, along with the corresponding standard deviations (in parentheses). Planned comparisons of the data were conducted using two-way ANOVAs (versions  $\times$  sub-expression), carried out separately over subject and item data.

As expected, the participants performed significantly better for phrasal sub-expressions than for non-phrasal sub-expressions. This superior performance was seen in the response times with a 196ms recognition advantage ( $minF'(1, 66) = 33.06, p < .01$ ). This advantage held for error rates also ( $minF'(1, 69) = 8.83, p < .01$ ). This indicates that the equations are perceived in a way that allows for faster and more accurate recognition of phrasal sub-expressions than non-phrasal sub-expressions.

Table 2: Mean correct response times (ms) and error rates (%) as a function of sub-expression type for Experiment 1.

Sub-Expression	RT(ms)	%Error
Phrasal	1153 (178)	14.8 (8.7)
Non-Phrasal	1349 (205)	25.2 (11.2)
Incorrect	1382 (246)	20.3 (9.5)

There was also a significant response time advantage for phrasal sub-expressions over incorrect sub-expressions ( $minF'(1, 57) = 35.97, p < .01$ ). However, there was no corresponding overall advantage for error rates, despite the fact that the item-based analysis was significant ( $F_1(1, 21) = 4.21, p = .053, F_2(1, 69) = 5.46, p < .05$ ). There was no significant difference between non-phrasal and incorrect sub-expressions for either response times or error rates.

The results of Experiment 1 provide support for our hypothesis that the internal representation used by experienced mathematicians is based on the phrasal structure of a parse tree. This comes from the logic of the experiment. Encoding of the equations significantly favours recognition of phrasal sub-expressions, indicating that knowledge of the constituent structure that underlies a parse tree is relied upon in the encoding process.

This outcome and those of previous work (Jansen et al., 1999) indicate that experienced mathematicians use an internal representation based on mathematical syntax and a parse tree structure. One interesting issue is just how much experience with mathematics is necessary before such representations become established. This is the focus of our second experiment.

## Experiment 2

Our hypothesis here is that considerable experience is necessary before humans can parse an equation based on its mathematical syntax. To test this hypothesis, recognition tasks were designed to examine the influence of both syntactic well-formedness and phrasal properties in identifying sub-

Table 3: Example equations and sub-expressions used in examining well-formedness.

Equation	Sub-Expression		
	Well-Formed	Non-Well-Formed	Incorrect
$\frac{9}{x(2y-5)} - 4x^3$	$(2y-5)$	$\frac{\quad}{x(2$	$x(4y+$
$x = 6yx - \frac{2x+2}{x}$	$2x+2$	$= 6yx-$	$\frac{6x+2}{y}$

expressions of equations. The participants in these experiments were students in their first year of high school (Year 7). This year level was chosen because it is one year before algebra becomes a major component of their mathematics syllabus (in Australia). The students had been introduced to the notion of a variable, but had not been introduced to the exponent notation and had dealt only with very simple expressions.

Due to the complex nature of the equations (at least by Year 7 standards), we expect to see no significant performance advantages for one type of sub-expression over another, indicating that the internal representations of the students are not based on mathematical syntax or parse tree structures. However if any advantages are present, this would indicate a predisposition towards encoding equations into syntactically based constituent chunks, even with very little experience.

## Method

**Participants** Eighteen participants successfully completed these experiments. All were Year 7 students, aged 12 to 13 years, with only limited knowledge of algebra. All participants were volunteers with normal or corrected-to-normal vision.

**Materials and Design** Experiment 2 consisted of two parts. Part A looked at syntactic well-formedness, the design of the experiment being similar to the experiment described in Jansen et al. (1999) which was conducted with competent adult mathematicians. Sixty equations were used, and sub-expressions of three types were generated for each.

- a) A *well-formed sub-expression*, which is a component of its equation, and conveys the same meaning on its own that it conveys in the equation.
- b) A *non-well-formed sub-expression*, which is also a component of its equation, but does not convey any coherent mathematical meaning on its own.
- c) An *incorrect sub-expression*, which was not part of the original equation. It can be either well-formed or non-well-formed. These act as fillers.

See Table 3 for examples. The equations consisted of between twelve and fourteen characters, and each of the sub-expressions contained between four and six characters (the average for well-formed sub-expressions was 4.77; for non-well-formed, 4.50; for incorrect, 4.66). Only the variable names  $x$  and  $y$  were used, with at most one fraction being present in any equation.

Part B of Experiment 2 again examined phrasal properties, and was based on Experiment 1. Sixty equations were constructed, along with three sub-expressions per equation (phrasal, non-phrasal and incorrect). The properties of the equations and sub-expressions are the same as described in Experiment 1, with the sub-expressions again containing between four and six characters (the average for phrasal sub-expressions was 4.79; for non-phrasal, 4.50; for incorrect, 4.60). Table 1 contains examples of these. The equations used in part A and part B of this experiment were all different.

For each part of the experiment, three counterbalanced versions were created allowing the presentation of all three sub-expression types for each equation, but ensuring that participants were presented with each equation only once to avoid practice effects. For each version, there were twenty instances of each type of sub-expression. Two additional equations were constructed as practice items. The same practice items were used in each version. The items of each version were presented in a different pseudo-random order for each participant.

Participants did both parts of the experiment in the one sitting, one after the other. Due to the tasks in part A and part B being so similar, the order in which they were done was balanced over all of the participants. Thus half of the participants did part A before part B, with the other half doing the experiment in the reverse order.

**Procedure** The experiments were carried out in a quiet room, with groups of four or five students at a time. Each participant was seated in front of an IBM compatible computer with a 14" monitor, running at a resolution of 800x600. All items were black text on a white background, presented by a purpose designed computer program.

For part A, the average width of the equations in pixels was



178 (range 135–219) with an average height of 47 (range 26–61). The average width of the sub-expressions in pixels was 72 (range 39–179) with an average height of 26 (range 18–51). For part B, the average width of the equations in pixels was 187 (range 97–244) with an average height of 46 (range 26–59). The average width of the sub-expressions in pixels was 72 (range 25–111) with an average height of 24 (range 16–52).

The procedure for each part was very similar to that used for Experiment 1. There was no difference in the display timing of the stimuli or the response mechanism. However, since the students were not expected to perform very well in the task, they may lose confidence in performance if continually reminded of errors. Consequently, no feedback was given. Otherwise, the experimental procedure was the same. Participants were also given a brief rest period between the two parts and took approximately 25 minutes to complete the entire experiment.

**Data Treatment** To reduce the unwanted effects of outlying data points, absolute upper and lower cut-offs were applied to response latencies, such that any response longer than 4000ms or shorter than 500ms was designated as an error. The maximum cutoff time here is longer for the Year 7 students than for the experienced mathematicians in previous experiments.

As expected the participants did not perform well in this task, with the accuracy achieved for many sub-expression types being no better than chance. Given that many students were clearly guessing when presented with these sub-expressions, an analysis of response time data would be meaningless. Analysis was therefore only conducted on error rate data, by a series of analyses of variance (ANOVAs) over both participant and item data. Where these were significant, they were combined in the  $\min F'$  statistic. No participants data was excluded from the analysis.

## Results and Discussion

The error rate for the three sub-expression types in part A (which examined well-formedness) is summarised in Table 4, along with the corresponding standard deviations (in parentheses). Planned comparisons of the data were conducted using two-way ANOVAs (versions  $\times$  sub-expression), carried out separately over subject and item data.

Table 4: Error rates (%) as a function of sub-expression type for Experiment 2A.

Sub-Expression	%Error
Well-Formed	36.1 (21.5)
Non-Well-Formed	51.7 (14.9)
Incorrect	50.6 (19.9)

The results of interest are the performance differences between well-formed and non-well-formed sub-expressions. Participants performed significantly better in recognizing well-formed sub-expressions than their non-well-formed counterparts with an advantage of 15.6% ( $\min F'(1, 37) = 7.50, p < .01$ ). In fact, since in each trial participants had a 50–50 chance of success, it is clear that for both non-well-formed and incorrect sub-expressions, participants were doing no better than random guessing. It is only for well-formed sub-expressions that they were performing better than chance.

Table 5 summarises the error rate data for the three sub-expression types in part B of the experiment (which examined phrasal properties), along with the corresponding standard deviations (in parentheses).

Table 5: Error rates (%) as a function of sub-expression type for Experiment 2B.

Sub-Expression	%Error
Phrasal	34.1 (15.4)
Non-Phrasal	53.6 (16.5)
Incorrect	49.4 (22.9)

The results show a significant 19.4% error rate advantage for phrasal sub-expressions over non-phrasal sub-expressions ( $\min F'(1, 59) = 12.07, p < .01$ ). As in part A, the incorrect and also the non-phrasal results indicate that participants are doing no better than chance in responding to these sub-expression types. However, performance was clearly above chance for phrasal sub-expressions.

Given the limited mathematical experience of the Year 7 students, these results are unexpected. The fact that the overall accuracy of the Year 7 students is far lower than for competent adult mathematicians, indicates that the development of their internal representation still has a long way to go. However, superior performance in recognizing syntactically well-formed and phrasal sub-expressions provides support for the notion that mathematical syntax plays an important role in the way that these students encode equations. This result therefore does not support our hypothesis that considerable experience is necessary before students can parse an equation based on its mathematical syntax.

Despite the significance of these results, it is not clear whether the students represent a heterogeneous or a homogeneous population with respect to their performances in this task. Therefore, a further analysis of the error rate data from this experiment was conducted. For part A of the experiment, a three-way split was carried out based on the difference in accuracy in recognizing well-formed and non-well-formed sub-expressions. The results of participants in each version were divided into three groups. The top group contained participants with the greatest performance advantage in recognizing well-formed sub-expressions over non-well-formed

sub-expressions. The bottom group contained those with the least advantage, or possibly even a disadvantage in recognizing well-formed sub-expressions over their non-well-formed counterparts. The remaining participants formed a middle group, but the results of this group were not of interest. Since there were six participants per version, each group contained the results of two participants from each version. ANOVAs were then conducted to compare the performance of the top and bottom group.

As expected, an even greater performance advantage of 32.5% in identifying well-formed over non-well-formed sub-expressions was found in the top group ( $\min.F'(1, 30) = 19.55, p < .01$ ). However, the performance of the bottom group revealed a slight disadvantage of 1.7% in recognizing well-formed sub-expressions over their non-well-formed counterparts. This result was not statistically significant ( $F < 1$  for analysis by both subject and item).

A similar analysis was conducted for part B of the experiment, with the three way split based on the accuracy difference between recognizing phrasal and non-phrasal sub-expressions. The top and bottom groups reflect the participants with the greatest and least performance advantage respectively, in recognizing phrasal sub-expressions over non-phrasal sub-expressions. The top group again had a significant performance advantage of 32.7% in identifying phrasal over non-phrasal sub-expressions ( $\min.F'(1, 60) = 20.17, p < .01$ ). For the bottom group however, the advantage was only 5.8% which was not statistically significant ( $F < 1$  for analysis by both subject and item).

This result indicates that within the population sample for Experiment 2, there are two distinct groups, one of students who encode equations based on mathematical syntax, and one of those who appear not to. One possible explanation for this result is that some students have more previous experience with algebra and mathematics than others. However, another possibility is that some students might have a stronger predisposition for using knowledge of mathematical syntax to guide construction of internal representations. Certainly more research will be needed before the cause of this result can be resolved.

## Conclusions

Previous research has suggested that adults competent in mathematics encode equations into constituents that have syntactically well-formed structure (Jansen et al., 1999). We have extended upon these results by providing support for the hypothesis that the internal representation used by mathematicians is based on the constituent structure of a parse tree. Evidence has also been presented which indicates that this encoding mechanism is present in young students. This result is surprising given that the students have very little experience in dealing with complex algebraic expressions.

The future direction of this research is to further investigate the encoding mechanisms and internal representations used to process equations, and in particular to examine how the rep-

resentations of equations are used in mathematical problem solving. Also, the positive result with the Year 7 students leads to the question of just how little mathematical experience is necessary before mathematical syntax begins to play a role in encoding equations. Whether or not the students are establishing representations based on mathematical syntax, or their performance reflects a more general encoding mechanism for such complex stimuli, can only be resolved by conducting similar experiments with children who have no experience with algebraic equations.

## Acknowledgements

The authors wish to thank the staff and students at the Prahran Campus of Wesley College in Victoria, Australia, for their helpful co-operation and assistance in conducting the experiments involving the Year 7 students.

## References

- Akmajian, A., Demers, R.A., & Harnish, R.M. (1984). *Linguistics: An Introduction to Language and Communication* (2nd ed.). Massachusetts: MIT Press.
- Anderson, R.H. (1977). Two-dimensional mathematical notation. In K.S. Fu (Ed.), *Syntactic Pattern Recognition Applications*. New York: Springer-Verlag.
- Clark, H.H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12, 335–359.
- Jansen, A.R., Marriott, K., & Yelland, G.W. (1999). Perceiving structure in mathematical expressions. In M. Hahn & S.C. Stoness (Eds.), *Proceedings of the twenty first annual conference of the cognitive science society*. Lawrence Erlbaum Associates.
- Johnson, N.F. (1968). The influence of grammatical units on learning. *Journal of Verbal Learning and Verbal Behavior*, 7, 236–240.
- Johnson, N.F. (1970). Chunking and organization in the process of recall. In G.H. Bower (Ed.), *The Psychology of Learning and Motivation*, (Vol. 4). New York: Academic Press.
- Ranney, M. (1987). The role of structural context in perception: Syntax in the recognition of algebraic expressions. *Memory and Cognition*, 15(1), 29–41.

# Algorithm, Heuristic or Exemplar: Process and Representation in Multiple-Cue Judgment

**Sari Jones (Sari.Jones@psyk.uu.se)**

Department of Psychology, Uppsala University  
Box 1225, SE-751 42, Uppsala, Sweden

**Peter Juslin (Peter.Juslin@psy.umu.se)**

Department of Psychology, Umeå University  
SE-901 87, Umeå, Sweden

**Henrik Olsson (Henrik.Olsson@psy.umu.se)**

Department of Psychology, Umeå University  
SE-901 87, Umeå, Sweden

**Anders Winman (Anders.Winman@psyk.uu.se)**

Department of Psychology, Uppsala University  
Box 1225, SE-751 42, Uppsala, Sweden

## Abstract

We present an experimental design that allows us to investigate the representations and processes used in human multiple-cue judgment. We compare three ideal models of how knowledge is stored and applied in a judgment: A linear additive model (LAM), a heuristic model, Take-the-best (TTB) and a generic exemplar-based model (EBM). The results show that people adaptively change processing depending on what information is present in the learning phase and whether or not the learning situation is compatible with the test. Feedback on a continuous variable provides information sufficient to estimate a LAM that can be used both when learning is and is not compatible with the test. When only dichotomous feedback is provided, the processes differ depending on the learning-test compatibility. At high compatibility, the processing is best described by EBM, but at low compatibility heuristic processes such as TTB become more frequent alternatives to LAM.

## Introduction

In the 1950's and 60's two new research paradigms emerged in cognitive science, categorization research (e.g., Shepard, Hovland, & Jenkins, 1961) and research on multiple-cue judgment (e.g., Hammond, 1955). While the former has continued to flourish, the Brunswikian inspired judgment research quietly left the arena in the 80's, although with a re-emergence in studies on realism of confidence (Gigerenzer, Kleinbölting, & Hoffrage, 1991; Juslin, 1994). The two paradigms have a lot in common, but there is seldom cross-reference between them (but see Kruschke & Johansen, 1999). A major conclusion from research on multiple-cue judgment is that linear models fit judgment data well (Brehmer, 1994) but in regard to knowledge representation and processes, there has been little research. In the categorization literature in contrast, a variety of models with explicit representational and process assumptions have been proposed (see Medin, 1989).

In this article, we bring the two paradigms together by combining multiple-cue learning with theories and methods

from research on categorization. We present an experimental design that allows us to investigate knowledge representations and processes in human judgment. As a point of departure we take research that postulates multiple levels of representation (e.g., exemplars, rules) that compete to control the judgments in a specific task (Ashby et al., 1998; Erickson & Kruschke, 1998; Logan, 1988). The idea is that experience with some domain may lead to co-existing representations at several levels. A general hypothesis is that the process and representation that dominates at the time of judgment is contingent on an interaction between the learning environment and the judgment task to which the knowledge is later applied. We offer some preliminary ideas in regard to the principles that determine this interaction. The crucial question is: In what circumstances will a particular level of representation dominate the judgments?

We compare three ideal models of how knowledge is stored and applied in a multiple-cue judgment task. *Linear Additive Models* (LAM) suggest that we store information in memory about: (a) the weight or *cue validity* attached to each cue in the form of a linear coefficient, and (b) an *algebraic rule* for the combination of the cues, in this case a linear additive rule (Brehmer, 1994). The process at the time of judgment is *cue-integration*. Recently, a simpler and more heuristic alternative has been proposed in terms of the *Take The Best* algorithm (TTB; Gigerenzer & Goldstein, 1996; Gigerenzer, Todd, & ABC Group, 1999). TTB suggests that the single most valid cue that is applicable is used and that no information is integrated. The knowledge in memory are *cue validities* and the process amounts to *cue-substitution*. Finally, *Exemplar-Based Models* (EBM) from the categorization literature (e.g., Medin & Schaffer, 1978; Nosofsky & Palmeri, 1997) assert that the memory traces of each encountered object are stored in memory and that judgments are based on the similarity between the new object and the already stored exemplars. In this case, the representations are *exemplars* and the process is similarity-based *retrieval* from memory.

## Overview of the Experimental Design

The design presented in this article is based on the idea that an object is judged according to cues. The participants learn that there is a species of frogs that vary in degree of toxicity (0 to 100 %). This attribute depends on four characteristics of the frog; color of the back (green or brown), shape of a spot on the back (wedge shaped or round), size of glands above the eyes (large or small) and color of the abdomen (white or light yellow). These characteristics are binary cues that have the weights .4, .3, .2 and .1 respectively, in a linear equation: Toxicity =  $.4 \times \text{Cue 1} + .3 \times \text{Cue 2} + .2 \times \text{Cue 3} + .1 \times \text{Cue 4}$ . The weights can be understood as the proportions of poison that each cue adds to the total amount of poison.

Table 1: The exemplars and total proportion of poison when the weights on Cues 1 to 4 are .4, .3, .2, and .1 respectively.

Exemplar	Cue 1	Cue 2	Cue 3	Cue 4	Total
1	1	1	1	1	1
2	1	1	1	0	0.9
3	1	1	0	1	0.8
4	1	1	0	0	0.7
5	1	0	1	1	0.7
6	1	0	1	0	0.6
7	1	0	0	1	0.5
8	1	0	0	0	0.4
9	0	1	1	1	0.6
10	0	1	1	0	0.5
11	0	1	0	1	0.4
12	0	1	0	0	0.3
13	0	0	1	1	0.3
14	0	0	1	0	0.2
15	0	0	0	1	0.1
16	0	0	0	0	0.0

Adding the weights of all cues with positive cue values (1) gives the total proportion of poison for each frog. For example, if a frog has a green back (1), round spot (0), large glands (1) and light yellow abdomen (0) the total proportion of poison is  $1 \times .4 + 0 \times .3 + 1 \times .2 + 0 \times .1 = .6$ . It is convenient to describe the frogs according to their binary code. The above frog is Exemplar 6 (1010) in Table 1. With four cues, there are 16 possible exemplars. The proportion of poison varies between 0.0 and 1.0 (see Table 1).

In the *learning phase* participants learn to judge whether an exemplar is dangerous or not. Exemplars with a proportion of poison above .5 belong to the category Dangerous, whereas exemplars with a proportion of poison below .5 belong to the category Not dangerous. The participants receive *dichotomous feedback* about the accuracy of their prediction (e.g., “Correct” or “Wrong”). In addition they may or may not receive *continuous feedback* about the exact proportion of poison (e.g., “The amount of poison is 70%”). Exemplars that have a proportion of poison of exactly .5 are randomly assigned as dangerous or not. Three exemplars are omitted in training to test EBM, as described below.

The *single-object test* is the same task as in the learning phase, except that also exemplars that were omitted in the training phase are introduced. In the *pair-comparisons test*, two exemplars are compared in regard to degree of toxicity (i.e., which is the most dangerous one). Feedback is with-

held in the tests. The two tests allow us to investigate how knowledge acquired in one task is applied to a new judgment task, and how this affects the choice of representation.

## Model Specifications

**LAM** The characteristic of LAM is that cue validities and a linear, additive integration rule are retrieved from memory and all four cues are weighed together to calculate a total proportion of poison for each exemplar. Clearly, LAM provides the optimal algorithm for the task: If participants have accurate estimates of the cue validities (linear coefficients) they will categorize all exemplars in the single-object test correctly, except when the proportion of poison is .5. With a correctly estimated LAM the accuracy is .94 (15 out of 16 judgments correct), which provides a ceiling on the accuracy that can be attained.

In the pair-comparisons test, LAM computes an estimated toxicity of each of the two exemplars and decides on the exemplar with the higher estimated toxicity. With a correctly estimated LAM all judgments are correct, except when comparing exemplars with the same poison proportion, where the judgment is a guess (e.g., when comparing Exemplars 4 and 5). When all exemplars are compared to each other once (120 comparisons) the success rate is .98 (5 out of 120 comparisons will be guesses). The representations—the cue validities and the integration rule—and the process of cue integration are the same for all exemplars. Therefore, in the test phase LAM predicts that there should be no difference in the accuracy for (old) exemplars previously presented in the learning phase and (new) exemplars that have not been encountered previously.

**TTB** The characteristic of TTB is that only one of the four cues is used to make a judgment: The cue considered to be the most valid one. In the context of TTB, cue validity is defined as the conditional probability of a correct choice across all cases where the cue is applicable. In the single-object test, TTB implies that an exemplar with cue value 1 on the most valid cue is categorized as dangerous, and an exemplar with cue value 0 as not dangerous. In this application, the most valid cue (Cue 1) has validity .81 when applied to all 16 exemplars. The highest accuracy attainable with TTB in the single-object test is therefore .81.

In the pair-comparisons test, a cue is applicable if one exemplar has cue value 1 and the other 0 on this cue. If the best cue is applicable a judgment can be made, but if the cue values are identical for both exemplars the second to best cue has to be considered, and so on. The most valid cue (Cue 1) has a cue validity of .95 in the application to the pair-comparisons test. The second to best cue (Cue 2) has a validity of .93. In the pair-comparisons test the highest possible accuracy with TTB is .95. This presumes consistent application of the most valid cue that is applicable, correct conception of cue-directions, and correct rank ordering of cue validities. Because the representation and process is the same, TTB predicts that there should be no difference between new and old exemplars.

In addition to these global indices, there are critical exemplars for each cue that discriminate TTB from the other models. In the single-object test, there are two critical ex-

emplars for each cue. If TTB is used, both of these exemplars will be judged incorrectly, whereas if LAM is used they will be judged correctly. Exemplars 8 (1000) and 9 (0111) will be incorrectly categorized if Cue 1 is attended to, because the judgment is based on the binary value of Cue 1 instead of the total poison proportion of the exemplar. Similarly, in the pair-comparisons test there are comparisons that signal the use of TTB. There are three critical comparisons that will be judged incorrectly when Cue 1 is used, because the judgment is based on the binary value of Cue 1 rather than the overall poison proportion of the exemplars. When Cue 2 is used as the best cue, 10 comparisons are critical.<sup>1</sup>

**EBM** With EBM the new exemplar is judged by similarity to retrieved memory traces of previous exemplars. In the single-object test we assume that the process is well-captured by a standard exemplar-based model from the literature, *the context model* (Medin & Schaffer, 1978). On this account, the new exemplar (probe) is assigned to the category Dangerous with a probability equal to the proportion of the summed similarity to the stored exemplars in the category Dangerous, relative to the summed similarity to all stored exemplars. When the context model is applied to the single-object test with parameters that imply extreme specificity (i.e., all similarity parameters equal to 0), the model only retrieves identical exemplars. In this case, the context model produces the same accuracy as LAM, .94.

Application to the pair-comparisons test, where the two objects (probes) are compared in regard to a continuous variable, is more complicated. In a learning environment with only dichotomous feedback, an EBM like the context model can merely differentiate between pair-comparisons that contrast exemplars from different categories. This holds for 50 percent of the comparisons, and for these an EBM can attain perfect accuracy. For the remaining comparisons that involve two dangerous or two non-dangerous exemplars the judgment has to be based on a guess associated with .5 accuracy. Thus: in a condition with only dichotomous feedback an EBM can at most attain an accuracy of .75 ( $.5 \times 1 + .5 \times .5$ ). When feedback is continuous, on the other hand, an EBM can potentially store also the continuous value with the exemplar. An extension of the original context model that applies to estimation of continuous variables is *PROBEX* (for *PROBABILITIES* from *EXEMPLARS*, Juslin & Persson, 1999). When *PROBEX* is applied to pair comparison it makes one estimate of the continuous variable for each of the two probes. Each estimate is a weighted average of the values on the variable that have been stored with previous exemplars, where the weights are the similarities to the probe. The rule for computation of similarity is the multiplicative similarity rule of the context model. In a pair-comparisons task, *PROBEX* decides on the probe with the higher estimated value on the continuous variable. Again, if parameters are set so that only identical exemplars are retrieved (weighted), *PROBEX* allows the same accuracy as

LAM, that is, .98.<sup>2</sup> The reader is referred to Medin and Schaffer (1978) and Juslin and Persson (1999) for a complete specification of the models.

For current purposes, it is sufficient to highlight a general property of many exemplar-based models: Because the judgments are based on similarity to stored exemplars and the multiplicative similarity rule implies a particular sensitivity to identical exemplars, accuracy should be higher for old exemplars that correspond exactly to stored exemplars than for new exemplars. It should thus be easier to categorize exemplars, or compare two exemplars, that have been encountered previously, than new exemplars. EBM can be tested by omitting exemplars in the learning phase and later introducing them in the test phase. If EBM is used, the proportion correct should be higher for old exemplars than for new exemplars.

### Cost-Benefit Considerations and Learning-Test Compatibility

We will concentrate on two principles that determines the representation and process that dominates in a task: *cost-benefit considerations* and *learning-test compatibility*. As a preliminary step to this analysis, we have to make a few additional assumptions about the three models. We interpret both LAM and TTB to involve conscious, controlled and analytical processes constrained by short-term memory. We expect that short-term memory can hold at most a few elements (e.g., cue validities) active at any moment and that the process requires active mental effort. EBM is memory-based and the retrieval processes are assumed to be preconscious, automatic and to require little mental effort. It is possible, however, that EBM requires a longer period of learning to accumulate a sufficient set of exemplars. We assume that these processes are present simultaneously and compete to control a specific judgment.

The principle of cost-benefit consideration implies that the relative gain of applying a process is weighted against the cost of applying it. In the context of a design like the present one, the gain is accuracy and the cost is mental effort. The cost involved in applying a process concerns both investment in the learning phase (e.g., the effort to estimate linear coefficients with LAM), and in the test phase (e.g., cue integration). Payne, Bettman, and Johnson (1993) have studied cost-benefit considerations in choice of decision rules in multi-attribute decision making. This research suggests that as the cost of applying a mental algorithm (e.g., LAM) increases people adapt and turn to heuristic processes (e.g., TTB). The principle of learning-test compatibility implies that for memory-based processes, the conditions for successful retrieval are optimal when the circumstances at test match those at learning. This principle is supported by an extensive literature on memory, and illustrated by concepts such as the principle of encoding specificity (Thomson

<sup>1</sup> The results for Cues 3 and 4 are not reported because analyses of the data show that, if TTB is used Cue 1 and Cue 2 are most frequently used as the best cue.

<sup>2</sup> The limits on accuracy are conditional on complete knowledge of all 16 exemplars. Omitting 3 exemplars in the learning phase constrains the possibility to estimate linear coefficients and cue validities, and to store exemplars. These deviations are minor and have no effect on the conclusions.

& Tulving, 1970) and transfer-appropriate processing (Morris, Bransford, & Franks, 1977).

We apply these principles to three experimental manipulations: (a) *Feedback quality*: Presentation of continuous or dichotomous feedback in the learning phase. (b) *Test format*: Single-object or pair-comparisons test. (c) *Cue order*: Fixed or varied presentation order of the cues.

### Feedback Quality

Presentation of continuous feedback in the learning phase should enhance the use of LAM. Continuous feedback facilitates estimation of linear coefficients for each cue. This decreases the cost required to use LAM and thereby increases its prevalence. Because learning the cue weights with dichotomous feedback is arduous, cost-benefit considerations suggest that participants are likely to resort to a computationally simpler process like EBM or TTB.

### Test Format

In single-object tests, learning and test consist of the same task, whereas in pair-comparisons tests the learning phase and the test differ. The principle of learning-test compatibility implies that EBM should be more common in the single-object test, where the conditions at learning and test are identical. This increases the probability of successful retrieval of stored exemplars. Note that LAM and EBM, in principle, allow the same accuracy, but at different costs: LAM allows rapid learning but requires larger mental effort. EBM requires little mental effort but, presumably, more extended training to attain the same level of performance.

The change of context in the pair-comparisons test should increase the rate of responses guided by LAM or TTB, which are not dependent on episodic retrieval. Participants should have less opportunity to rely on memory (EBM), and turn to the analytic processes implied by LAM and TTB. Moreover, when only dichotomous feedback is provided in the learning phase, EBM provides poor guidance in a pair-comparison that concerns continuous values of the exemplars.

The principles of cost-benefit consideration and learning-test compatibility should interact in a specific way. In pair-comparisons, more mental effort is needed to make the judgment compared to in a single-object test. If LAM is used in a single-object test, four cues are weighted and added. In a pair-comparisons tests, the cognitive effort is doubled. This should make participants who use LAM for single-object judgments swap into a less demanding process in the pair-comparisons test. Because of lower training-test compatibility in the pair-comparisons task, however, they are likely to divert to a heuristic algorithm such as TTB rather than to EBM.

Predictions for the first two manipulations are summarized in Table 2. The provision of continuous feedback should allow the participants to estimate a LAM, and this should be particularly evident in the single-object test where application of LAM demands less effort. With dichotomous feedback, EBM should dominate in the single-object test where training and test conditions match, whereas TTB

should dominate in the pair-comparisons test where this match is lower.

Table 2: Processes predicted to dominate as a function of the manipulation of feedback quality and type of test.

Test format	Feedback quality	
	Continuous	Dichotomous
Single-object	LAM	EBM
Pair-comparisons	LAM/TTB	TTB

### Cue Order

In the experiment, we presented the cues in fixed or randomly varied presentation-order across trials. Our hypothesis was that a fixed cue order should enhance the use of EBM because this should maximize training-test compatibility. This manipulation produced no effects, an observation to which we return in the discussion.

## Method

### Participants

Sixty-four persons (41 women and 23 men, mean age = 24.4) participated. All but 4 were undergraduate students at Uppsala University. Participants received a course credit or a cinema voucher worth approximately 75 SEK for participating.

### Design and Procedure

Each learning trial consisted of presentation of an exemplar of a fictitious frog species with 4 different attributes (described above) with the information presented in written text. Three exemplars were omitted and the remaining exemplars were judged 10 times each in the learning phase, making a total of 130 trials. For half of the participants, exemplars 4, 9 and 10 were omitted, for the other half exemplars 5, 6 and 7. These exemplars are equal in poison percentage (i.e., 4 is equal to 5, 9 is equal to 6 and 10 to 7). The omission was thus counterbalanced.

The participants answered the question “Is the frog dangerous or not?” and received dichotomous feedback “Correct answer“ or “Wrong answer“. Half of the participants also received continuous feedback about the percentage of poison of the frog, for example, “70% poison” The weights, .4, .3, .2 and .1, were randomized to different cues for each participant. For half of the participants cues were presented in fixed order (in the same order and spatial location on the list) and for the other half cues were presented in a varied order.

After the learning phase, participants received a single-object test structurally identical to the learning phase, but without feedback. This phase consisted of 16 trials as all exemplars were judged once. Finally, a pair-comparisons test in which 2 exemplars were contrasted, (also without feedback) was administered. The question was: “Which frog is most dangerous?”. This test consisted of 120 trials (all exemplars were compared to each other once). Each session lasted 45 min to 1h and 30 min.

## Results

### Proportion Correct

In the single-object test, the use of LAM and EBM was predicted and the results support these predictions. Table 3 displays the mean proportions correct (M) and 95 % Confidence Intervals (CI) for each condition.

Proportions correct clearly refute the use of TTB in the continuous feedback condition since the confidence intervals do not include .81, the maximum performance possible for TTB. The proportion correct is significantly higher in the continuous feedback condition than in the dichotomous feedback condition,  $t(62) = 2.04$ , (*one-tail*)  $p = .03$ . This is expected if participants, to some extent, rely on LAM but not if they uniformly rely on EBM. In the continuous feedback condition it is easier to estimate cue validities and therefore the proportion correct is expected to be higher in this condition if LAM is used.

Table 3: Proportions correct (M) and 95 % Confidence Intervals (CI) for each condition of the single-object test.

Feedback	Cue order		Total
	Varied	Fixed	
<b>Continuous</b>	M=.86 * (CI: .81-.91)	M=.88 (CI: .80-.88)	M=.87 * (CI: .83-.92)
<b>Dichotomous</b>	M=.82 (CI: .78-.87)	M=.79 (CI: .70-.88)	M=.80 (CI: .76-.85)
<b>Total</b>	M=.84 (CI: .77-.89)	M=.84 (CI: .80-.87)	M=.84 (CI: .80-.87)

\*The CI does not include .81, the maximum performance possible for TTB.

In pair-comparisons, the participants were predicted to use LAM in the continuous feedback condition and TTB in the dichotomous feedback condition. The mean proportion correct is .90 (CI: .87-.94) in the continuous feedback condition and .80 (CI: .75-.85) in the dichotomous feedback condition. The proportions correct are consistent with TTB. In sum: The proportions correct falsify TTB in the single-object condition with dichotomous feedback, but the proportions correct are compatible with LAM and EBM in all conditions. Presentation order of the cues had no effect on accuracy.

### Old and New Exemplars

The differences in proportions correct for old and new exemplars are presented in Figure 1. The only condition in which there is a substantial advantage for old exemplars over new exemplars is the single-object/dichotomous feedback condition, where EBM is predicted to dominate. In the other three conditions in Figure 1, EBM is not supported. When feedback quality is high and the cost of applying a complex algorithm is low, participants can rely on the powerful (but choosy on data) LAM. When feedback is of poorer quality, the cost of applying LAM is too high and the participants resort to the less demanding EBM. This is particularly likely to occur when learning-test compatibility is high (i.e., in the single-object/dichotomous condition).

In Figure 2, proportions correct are displayed separately for new, mixed (comparing 1 new and 1 old exemplar) and

old exemplars in the pair-comparison test. In the continuous feedback condition, there is no difference between new, mixed or old exemplars, supporting LAM (TTB is refuted by the proportion correct, see Table 3). More perplexing, in the dichotomous feedback condition there is no difference between old and new exemplars, but the proportion correct on mixed exemplars is significantly lower. This effect reflects a bias to choose the old exemplar, resembling the *recognition principle* discussed in the context of TTB (Gigerenzer & Goldstein, 1996). This principle states that when presented with a pair comparison between two objects, only one of which is recognized, the participants will guess on the recognized object. In sum: In the dichotomous feedback condition, the comparison of old and new exemplars supports EBM in single-object tests and, potentially, TTB in the pair-comparison tests. In the continuous feedback conditions, the results support LAM.

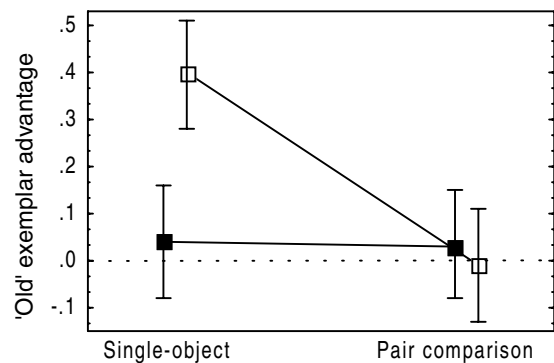


Figure 1: Mean difference between proportion correct for old and new exemplars and 95 % Confidence Intervals for the continuous (filled squares) and dichotomous (open squares) conditions.

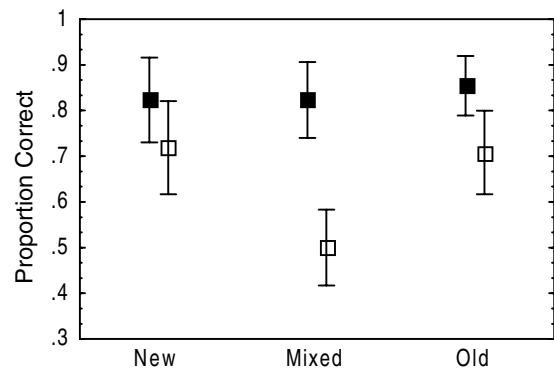


Figure 2: Mean proportion correct and 95 % Confidence Intervals of pair comparisons with new, mixed and old exemplars for the continuous (filled squares) and dichotomous (open squares) feedback conditions.

### Critical Exemplars

The critical exemplars do not provide much support for TTB in the single-object conditions. TTB with Cue 1 as the best

cue, should yield wrong judgments on Exemplars 8 (1000) and 9 (0111) in the single-object test. The proportion correct on these Exemplars is well above 0 in both conditions, specifically .83 (CI: .72-.94) in the continuous feedback condition and .73 (CI: .62-.85) in the dichotomous feedback condition. On individual level, five participants (7.8%) used TTB, four judged both critical exemplars of Cue 1 incorrectly and one the critical exemplars of Cue 2. These participants were equally distributed in the dichotomous and continuous feedback conditions. In the pair-comparisons test, TTB receives some support from the critical comparisons. Although a group level analysis shows no consistent use of TTB, on an individual level 16 participants (25%) used TTB: Nine with Cue 1 as the best cue and, surprisingly, seven with Cue 2 as the best cue.

### Discussion

We have introduced a design in which the knowledge representations and processes in a multiple-cue learning task can be studied. The results suggest that humans change cognitive processing, as a function of the information present during learning and the compatibility of learning and test, in a way that is consistent with the principles of cost-benefit and learning-test compatibility derived from previous research.

Specifically, the presentation of continuous feedback in the learning phase provided participants with information that allowed them more easily to estimate a LAM. A LAM is applicable both when learning is, and is not, compatible with the test, but the application is more demanding in the pair-comparisons test. Thus, there was more support for the domination of LAM when continuous feedback was provided, but less so in the pair-comparisons than the single-object test.

When only dichotomous feedback is available, the estimation of LAM becomes demanding and other processes come to dominate the judgments. The other processes correspond to the two classical ways of circumventing the limited capacity of controlled thought processes: memory-based performance, or automatization, and heuristic processing. When the test is similar enough to the learning task, processing is memory-based and relies on exemplar representations. When the test is different from learning, heuristic processes, such as TTB, increase in frequency. Overall, however, we found little evidence in support of TTB (i.e., a minority of participants in the pair-comparisons task seemed to rely on it). This may perhaps be explained by the relatively simple task used in the experiment and, indeed, as the task became more complex, evidence in favor of TTB seemed to increase. Nonetheless, at present there is little empirical data that provide support for the empirical validity of TTB.

To our surprise, the manipulation of fixed or varied order of the cues had no effect. Together with the clear effect of old and new exemplars in the dichotomous/single-object condition, this suggests that the judgments are sometimes guided by exemplar-memory, but the representation of the exemplars may be more conceptual than visual in its character.

### Acknowledgments

This research was supported by the Swedish Council for Research in Humanities and Social Sciences.

### References

- Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442-481.
- Brehmer, B. (1994). The psychology of linear judgment models. *Acta Psychologica*, *87*, 137-154.
- Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107-140.
- Gigerenzer, G., & Goldstein, D., G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650-669.
- Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A brunswikian theory of confidence. *Psychological Review*, *98*, 506-528.
- Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.
- Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review*, *62*, 255-262.
- Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, *57*, 226-246.
- Juslin, P., & Persson, M. (1999). *Probabilities from exemplars: On the role of similarity and frequency in probability judgment*. Manuscript submitted for publication.
- Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1083-1119.
- Logan, D. G. (1988). Towards an instance theory of automatization. *Psychological Review*, *95*, 492-527.
- Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, *44*, 1469-1481.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.
- Morris, C. D., Bransford, J.D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519-533.
- Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266-300.
- Payne, J. W., Bettman, J. R., & Johnson E. J. (1993). *The adaptive decision maker*. New York: Cambridge University Press.
- Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75* (13, Whole No. 517).
- Thomson, D. M., & Tulving, E. (1970). Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology*, *86*, 255-262.



# Influences on Attribute Selection in Redescriptions: A Corpus Study

Pamela W. Jordan (jordan@isp.pitt.edu)

Intelligent Systems Program

University of Pittsburgh

Pittsburgh PA 15260

## Abstract

We report on the results of the first stage of a corpus analysis that tests five hypotheses about how domain and discourse goals and the functions of repetition can influence the content of redescrptions in dialogue. We found a positive correlation between the attributes expressed in redescrptions and contexts in which three of these types of goals are predicted. These results provide us with guidance on the selection strategies we will test in the next stage of analysis.

## Introduction

In an extended discourse, speakers often redescribe objects that were introduced earlier in order to say something more about the object or the event in which it participates. The main goal when redescribing an entity is to re-voke the appropriate discourse entity. However, a goal-directed view of sentence generation suggests that speakers can attempt to satisfy multiple goals with each utterance [Appelt, 1985] and that a single linguistic form can opportunistically contribute to the satisfaction of multiple goals [Stone and Webber, 1998]. The possibility that goals besides identification could influence the content of a nominal expression<sup>1</sup> have not been fully addressed in computational work on generating nominal expressions.

The many-one mapping of goals to linguistic forms is more generally referred to as *overloading intentions* [Pollack, 1991]. Overloading can involve tradeoff across linguistic levels [Di Eugenio and Webber, 1996, Stone and Webber, 1998]. For example, an intention which is achieved by complicating a form at the semantic level may allow the speaker to simplify at the syntactic level by omitting important information [Stone and Webber, 1998].

Although we have learned that overloading is natural and perhaps even necessary, we have no well supported account of what degree of overloading is reasonable and what forms can more readily address multiple goals in dialogue. Without such an account, we have no principled way to deploy overloading in the automatic generation of natural language. Without well supported constraints on overloading, we are liable to create overloads in unnatural ways which will actually impede effective communication.

To investigate whether overloading applies to redescrptions, we examined 166 non-pronominal redescrptions found in 13 dialogues of the COCONUT

corpus [Di Eugenio et al., 2000]. This corpus contains computer-mediated dialogues in which two people collaborate on a simple design task, buying furniture for two rooms of a house. The participants' main goal is to negotiate the purchases; the items of highest priority are a sofa for the living room and a table and four chairs for the dining room. The participants also have specific secondary goals which further complicate the problem solving task. Participants are instructed to try to meet as many of these goals as possible, and are motivated to do so by associating points with satisfied goals. The secondary goals are: (1) Match colors within a room, (2) Buy as much furniture as you can, (3) Spend all your money. Each participant is given a separate budget and inventory of furniture and must decide what to make mutually known. Every furniture item in the inventory is described by five attributes; type, color, price, owner, and quantity.

In this article, we report on the results of the first stage of a corpus analysis that tests five hypotheses about how domain and discourse goals and the functions of repetition can influence the content of redescrptions in dialogue. We found a positive correlation between the attributes expressed in redescrptions and contexts in which three of these types of goals are predicted. The positive correlations will guide us in implementing and testing selection strategies for generating nominal expressions.

## Hypotheses about influences on redescrptions

Our hypotheses reflect non-identification goals that could influence the choice of attributes for a redescription. These goals are derived from work on the functions of repetition at the utterance or propositional level [Walker, 1993, Johnstone, 1994] and from observations about task intentions and constraint changes (e.g. matching colors) that were not directly communicated by the dialogue participants.

Our first hypothesis is based on the observation that in the COCONUT corpus, people often adjusted task constraints with no explicit discussion (38%).<sup>2</sup> Similarly to [Walker, 1993] for the propositional level, we suggest that the hearer is inferring changes from the redundancies in the redescription. So the repeated property could

<sup>1</sup>Identification being satisfied by more than a nominal expression also deserves consideration.

<sup>2</sup>Constraints are found in planning and scheduling tasks as well as design tasks [Di Eugenio et al., 2000, Jordan, 2000].

both help uniquely identify the intended discourse entity and enable the hearer to infer constraint changes.

To illustrate this possibility for COCONUT, assume that there is an initial constraint setting to match colors, that the speaker just discussed using a red table and prior to that introduced tables of various colors, four \$100 red chairs, and four \$75 green chairs. Finally, assume that she has decided to drop the color match constraint and suggest the cheaper \$75 chairs. When she communicates her suggestion, we hypothesize that she will prefer saying *green chairs* or *\$75 green chairs* over the more economical *\$75 chairs*. By choosing “green,” she adequately identifies the target chairs while also enabling the hearer to infer that she intends to drop the color match constraint. She has eliminated having to explicitly communicate the information [Walker, 1993] and reduced the risk of the hearer missing the inference [Carletta, 1992].

DOMAIN CONSTRAINT CHANGES HYPOTHESIS: Properties related to constraint changes are expressed in a context where the change must be inferred by the hearer.

Intentional relations indicate that an utterance, for example, elaborates or motivates other utterances, helping to bind together utterances to form a coherent discourse [Mann and Thompson, 1987]. Since the relations between utterances can influence the content and form of utterances [McKeown, 1985, Moser and Moore, 1995], we suggest that they could influence redescriptions as well. For example, in a context where all the object information is mutually known, a table needs to be chosen and red chairs have already been selected, (1b) can be considered motivation for the choice made in (1a).

- (1) a. Let’s use my table.  
b. It is red.

We know that it is beneficial under certain cognitive resource limitations to make motivations explicit even when the motivation is mutually known [Walker, 1993]. When we replace (1) with (2), the redescription may subsume the motivation and do the same thing.

- (2) Let’s use my red table.

PERSUASION HYPOTHESIS: Property values that are pivotal for deliberation are expressed in the context of goals to communicate a proposed action.

If a speaker repeats an utterance and provides no new information, this can show that a stage of the interaction is complete [Whittaker and Stenton, 1988, Jordan and Di Eugenio, 1997]. Repeating properties for a recently evoked item could show that the current stage has just been completed while doing so for an

older item could indicate that a higher level subproblem has been completed. In (3), S’s second utterance appears to end a stage in the interaction, in this case the end of the agreement process for a *select sofa* action [Di Eugenio et al., 2000].

- (3) S: ...I have a \$300 yellow sofa...

G: My sofa’s are more expensive so buy *your \$300 yellow sofa*. Also...

S: ... I will go ahead and buy *the \$300 yellow sofa*.

COMMITMENT HYPOTHESIS: In the context of a commitment to a proposal, all the properties expressed in the proposal will be repeated.

The second case of indicating that a higher level subproblem has been completed is a summary. Given the goals for the COCONUT design problem, the participants need to agree on the furniture items selected, whether colors should match in a room and whether they have selected as many different items as they can. In addition they have the hard constraint of not overspending. We suggest that a speaker would need to review all of these decisions in order to summarize the currently agreed upon solution state. Since the furniture items all have attributes related to each of these decisions, we hypothesize that a speaker will economize his summarization by including all the attributes that relate to decisions when listing the agreed upon items in the current solution. For example, in (4), the participants have previously decided on the dining room items and are completing their selections for the living room. Note that when G requests a summary of the living room, D includes all the attributes that relate to decision making for the task. D also decides to review all the items that they had previously decided upon for the dining room.

- (4) G: I got the rug. What do you have in the living room and what are the prices of the items

D: the green sofa in the living room 350. dining room—> 3 yellow chairs 75 each, 1 high-table yellow, 1 yellow rug

SUMMARIZATION HYPOTHESIS: In the context of a previously completed problem or subproblem, all decision relevant, mutually known properties for an item will be repeated.

Finally, a speaker might also repeat an utterance to show that it was understood [Clark and Schaefer, 1989, Brennan, 1990, Walker, 1992, Walker, 1993]. In the COCONUT corpus, the hearer sometimes repeats the description in the turn immediately following. For example, in (3), G repeats S’s description of the sofa, although the sofa was introduced by S. We claim that this type of redescription could help verify that the property information was correctly understood.

VERIFICATION HYPOTHESIS: In the context of a newly introduced entity, all the properties expressed will be repeated by the hearer in his/her next turn.

### Analyzing the Corpus

To identify the contexts and attribute usages we described in the hypotheses we used annotated features<sup>3</sup> and other easily extractable features of the corpus (e.g. the utterance speaker and the proximity of a redescription to its last mention). We used two types of corpus annotation features to support our study: (1) discourse entity level annotations that capture (a) the definitions and updates for discourse entities as a dialogue progresses and (b) the properties selected to redescribe discourse entities, and (2) utterance level annotations that capture (a) the problem solving state in terms of goals and constraint changes, and (b) discourse features such as commitments and offers. All of the features we used were found to have good intercoder reliability [Di Eugenio et al., 1998, Jordan, 1999]. The annotation features are described in detail in [Jordan, 2000].

Of these feature, the discourse features are the most complicated. They are based on elements of the agreement process described in [Di Eugenio et al., 2000]. The high-level definitions for these features are:

- propose: The speaker offers an item in a context where he already knows the hearer’s alternatives.
- partner decidable option: The speaker describes or offers an item but does not know the hearer’s alternatives.
- unconditional commit: The speaker indicates his commitment to using an item.
- unendorsed option: The speaker offers an item in a context where he knows the hearer’s alternatives and indicates an alternative is better.

Next we will describe how we used the annotation features to identify the contexts and redescrptions indicated in our hypotheses.

### Results of Corpus Analysis

We used chi-square and the Fisher exact tests<sup>4</sup> to check for correlations between factors. Although these tests assume independence, we feel we can violate this assumption given that the dependencies between redescrptions aren’t necessarily direct and obvious.<sup>5</sup> In all of the contingency tables, the counts are restricted to utterances

<sup>3</sup>The annotators for the corpus only knew the high-level goal of the study. Because of this, their annotation decisions were not influenced by the hypotheses we are studying here.

<sup>4</sup>We use the Fisher exact test when  $N < 20$  and an expected frequency is  $\leq 5$ .

<sup>5</sup>On average, the dialogues have 42 utterances, 25 discourse entities, and 6 utterances between redescrptions.

Changes	Related Properties
Room Color Limit	color
Price Upper Limit	price
Price Evaluator	price
Property Limit	color, price

Table 1: Associated Properties and Changes

that contain redescrptions. Finally the counts were all done automatically using software that interpreted the annotation features since the contextual factors generally involved multiple annotation features.

**Domain Constraint Changes Hypothesis** For this hypothesis we test whether there is a difference in property usage when a constraint change is communicated implicitly or explicitly. COCONUT is annotated with features indicating (1) whether a constraint change was communicated and whether this was accomplished implicitly or explicitly (2) which properties were included in the redescrptions. We examined each utterance for every constraint change that is generally possible for the domain when populating the cells of the contingency table.

We only count properties that relate to constraints. For example, we only look at the usage of the color property for the color match constraint or price for placing price limits. In Table 1, we list each of the constraint types that we examined and the property that we expected would be useful for inferring that change.<sup>6</sup> Our expectations derive from the instructions given to the COCONUT dialogue participants.

	Property Used	Property not Used
Implicit change	9	0
Explicit change	2	11

Table 2: Contingencies for Domain Constraint Changes Hypothesis

Table 2 shows that in the context of an implicit constraint change, properties related to the change are more likely to be used in the description than when the change is explicit (Fisher Exact Test,  $p < 0.0002$ ).

**Persuasion Hypothesis** For the Persuasion hypothesis, we wish to test whether expressing a property in a redescription is related to whether the expressed property makes the redescrbed item more desirable as a solution for a goal than the alternatives. For example, the cost of the item being redescrbed might be lower than any of the alternatives that have been discussed so far.

<sup>6</sup>The relevant property for the property limit constraint is indicated in the annotation for the constraint change.

A persuasion context exists when a proposal is to be made and alternate solutions exist and there is a contrast between the colors or prices that make the proposed item clearly a better choice. Given the analysis of the agreement process in [Di Eugenio et al., 2000], we first look for either a propose utterance, or an unconditional commitment utterance where the previous state is an unendorsed option, a partner decidable option or a list of options in which the speaker intentions are unclear.

For each of the unconditional commitment cases, we present examples. First, in (5), A’s partner decidable option is followed by B’s unconditional commitment.

(5) A: I have a blue sofa for \$200.

B: I have a yellow sofa for \$250. Let’s go with your \$200 sofa.

In (6), B does not endorse the option he presents but A overrides his objection with an unconditional commitment to it.

(6) A: We have \$100 left. I still have that \$50 blue chair.

B: I have a rug for \$100, but it is yellow.

A: We don’t need to match. Let’s get your \$100 rug.

Finally, in (7), A lists all of the items he has available. From the perspective of the agreement structure, lists such as this have no high-level task goals associated with them. However, the items do become part of the dialogue participants shared knowledge allowing all the items to be considered during problem solving so that they can become alternative options for the goals they are implicitly associated with. Because B is a position to deliberate<sup>7</sup>, his second utterance is annotated as an unconditional commitment. In this case there are two possibilities for what sofa to select, a persuasion context arises.

(7) A: I only have 2 red tables for \$200, 1 green table for \$350 and 4 \$50 blue chairs. I don’t have any rugs or lamps but I have 1 yellow sofa for \$200.

B: I have yellow rug for \$75 and a blue sofa for \$200. Let’s buy your yellow sofa and my rug.

Once we have identified possible persuasion contexts, we need to check for contrasts with alternatives. The alternatives are approximated by accumulating a list of the items evoked for each action. After a propose or unconditional commitment, all the items in the list for an action get flushed before starting over with the proposed item.

<sup>7</sup>This deliberation requirement for unconditional commitment is related to the problem solving architecture and is justified in [Di Eugenio et al., 2000].

Contrast	Related Property
Matches room but not alternatives	color
Cheaper than alternatives	price
More expensive than alternatives (near end of problem)	price

Table 3: Associated Properties and Contrasts

Next we check for contrasts. The contrast possibilities are shown in table 3 and arise from the COCONUT problem description. We were unable to accurately model the goal of buying as much as possible with the annotations available. For color we compare the color of the proposed item to those items already selected for the room and the alternate items. If the proposed item matches items already selected for the room while none of the alternates do, then a persuasion context exists. For prices there are two possibilities that depend on whether or not the end of the problem solving effort is nearing. An item may be a better choice when either (1) the price of the proposed item is greater than that of each alternate (i.e. it may be helping to spend out the budget) or (2) the price of the proposed item is less than that of each alternate (i.e. the cheaper item may be preferred since it leaves some money for other purchases).

Table 4 shows support for the persuasion hypothesis ( $\chi^2 = 5, p < .05, df = 1$ ).

	Property Not Used	Property Used
no contrast	18	9
contrast	13	24

Table 4: Contingencies for Persuasion Hypothesis

**Commitment Hypothesis** Here we test whether in the context of a commitment to a proposed action all the properties expressed in the proposal are more likely to be repeated. A commitment context exists when either (1) there is a previous proposal or unconditional commitment for the action involving the entity in the immediately previous turn and no other items must have been discussed for the action in the interim or (2) a speaker unconditionally commits again after doing so in his previous turn.

When determining repeated properties, we discount the type and owner properties. The type property is excluded because it involves pronominalization and zero anaphora; issues we are not addressing in this research. We exclude the owner property because its only function is identification in this domain.

Table 5 indicates that in contexts where a commitment is predicted, all mutually known properties are more likely to be included in redescrptions (Fisher Ex-

act Test,  $p < .0171$ ).

	Not Repeat Properties	Repeat Properties
No Commitment	7	8
Commitment	2	20

Table 5: Contingencies for Commitment Hypothesis

**Summarization Hypothesis** Here we test if the previous completion of a problem or subproblem correlates with expressing all the decision related, mutually known properties in a redescription. First, we must isolate redescriptions that occur after an agreement has been reached for the action.

A summarization context exists when an agreement has been reached for the action without the action being readdressed between the agreement and the current turn. The achievement of an agreement state is approximated when either (1) a propose or partner decidable option was the last state for the action and it happened more than two turns ago or (2) an unconditional commit was the last state and it happened two or more turns ago. In the first case, the agreement must be inferred and in the other the agreement is more explicit.

For the agreement state under condition (1), we require more than two turns to intervene because we want to allow for the cases where the partner left the decision pending by moving on to a dependent action (e.g. a final table decision may be left pending until the chair options are explored). We are estimating that if the action is not revisited after three turns, then it was not put on hold pending work on another action and that the partner agreed by moving on to another independent action.<sup>8</sup> This test for agreement takes into consideration that the initiation of the relevant next contribution shows evidence of understanding [Clark and Schaefer, 1987] and possibly joint commitment. For condition (2), we require that there be an intervening turn so that the partner is able to show that he has moved on to some other problem.

As with the commitment hypothesis, the type and owner properties are excluded when determining whether mutually known properties are repeated.

Table 6 indicates there is no correlation between a summarization context as we have characterized it and whether all the mutually known properties that relate to decisions get repeated ( $\chi^2 = 1.49, df = 1, NS$ ).

**Verification Hypothesis** With this hypothesis we test whether the repetition of all the properties presented in a previous description correlate with a con-

<sup>8</sup>In the initial version of the annotation scheme, there was a feature for indicating dependent actions but it was dropped because of poor intercoder reliability.

	All Mutual Properties Used	Not All Mutual Properties Used
Not End of Agreement Process	54	117
End of Agreement Process	8	8

Table 6: Contingencies for Summarization Hypothesis

text in which the entity was just introduced. In this case we collect all the properties that were presented in the turn where the item was first described and check whether this mention of the item was in the immediately previous turn or further back in the dialogue. As with the commitment and summarization hypotheses, the type and owner properties are excluded when determining whether properties are repeated. Table 7 shows no correlation between the verification context and the choice of attributes ( $\chi^2 = .06, df = 1, NS$ ).

	Properties Not All Repeated	Properties All Repeated
initial not in previous turn	1	0
initial in previous turn	44	2

Table 7: Contingencies for Verification Hypothesis

## Conclusion

Our analysis of the COCONUT corpus, shows positive correlations between the content of redescriptions and three of the contexts in which the repetition and domain and discourse goals we considered are expected. In particular, the contexts in which constraint changes, reasons for proposing, and commitment to proposals are predicted, positively correlated with the attributes expressed in the redescriptions of discourse entities. Finally, we found no support for the hypotheses that the properties expressed in a redescription correlate with verification or summarization contexts. In the case of the verification context, it is possible that the non-interruptibility of the COCONUT communications setting makes this sort of repetition function unnecessary. In the case of the summarization context, our ability to accurately detect this context may have been hampered by the estimates we had to make about the current state of the problem solving. Furthermore, there may be additional influences that depend on the reason for the summarization or the point at which it occurs during problem solving.

In future work we will test an attribute selection algorithm that embodies these hypotheses and compare

it against human performance and baseline algorithms that only consider the identification goal for redescrptions (e.g. IDAS [Dale and Reiter, 1995]).

## References

- [Appelt, 1985] Appelt, D. E. (1985). Planning English referring expressions. *Artificial Intelligence*, 26(1):1–33.
- [Brennan, 1990] Brennan, S. E. (1990). *Seeking and Providing Evidence for Mutual Understanding*. PhD thesis, Stanford University Psychology Dept. Unpublished Manuscript.
- [Carletta, 1992] Carletta, J. C. (1992). *Risk Taking and Recovery in Task-Oriented Dialogue*. PhD thesis, Edinburgh University.
- [Clark and Schaefer, 1987] Clark, H. H. and Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2:19–41.
- [Clark and Schaefer, 1989] Clark, H. H. and Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, 13:259–294.
- [Dale and Reiter, 1995] Dale, R. and Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.
- [Di Eugenio et al., 1998] Di Eugenio, B., Jordan, P. W., Moore, J. D., and Thomason, R. H. (1998). An empirical investigation of collaborative dialogues. In *ACL-COLING98, Proceedings of the Thirty-sixth Conference of the Association for Computational Linguistics*, Montreal, Canada.
- [Di Eugenio et al., 2000] Di Eugenio, B., Jordan, P. W., Thomason, R. H., and Moore, J. D. (2000). The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *To Appear in International Journal of Human-Computer Studies*.
- [Di Eugenio and Webber, 1996] Di Eugenio, B. and Webber, B. (1996). Pragmatic overloading in natural language instructions. *International Journal of Expert Systems, Special Issue on Knowledge Representation and Reasoning for Natural Language Processing*, 9(1):53–84.
- [Johnstone, 1994] Johnstone, B. (1994). Repetition in discourse: A dialogue. In Johnstone, B., editor, *Repetition in Discourse: Interdisciplinary Perspectives, Volume 1*, volume XLVII of *Advances in Discourse Processes*, chapter 1. Ablex.
- [Jordan, 1999] Jordan, P. W. (1999). An empirical study of the communicative goals impacting nominal expressions. In *Proceedings of the ESSLLI workshop on The Generation of Nominal Expressions*.
- [Jordan, 2000] Jordan, P. W. (2000). *Intentional Influences on Object Redescrptions in Dialogue: Evidence from an Empirical Study*. PhD thesis, Intelligent Systems Program, University of Pittsburgh.
- [Jordan and Di Eugenio, 1997] Jordan, P. W. and Di Eugenio, B. (1997). Control and initiative in collaborative problem solving dialogues. In *Computational Models for Mixed Initiative Interaction. Papers from the 1997 AAAI Spring Symposium. Technical Report SS-97-04*, pages 81–84. The AAAI Press.
- [Mann and Thompson, 1987] Mann, W. and Thompson, S. (1987). Rhetorical Structure Theory: A Framework for the Analysis of Texts. Technical Report RS-87-190, USC/Information Sciences Institute.
- [McKeown, 1985] McKeown, K. R. (1985). *Text Generation. Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press.
- [Moser and Moore, 1995] Moser, M. and Moore, J. D. (1995). Investigating cue placement and selection in tutorial discourse. In *Proceedings of 33rd Annual Meeting of the Association for Computational Linguistics*, pages 130–135.
- [Pollack, 1991] Pollack, M. E. (1991). Overloading intentions for efficient practical reasoning. *Noûs*, 25:513–536.
- [Stone and Webber, 1998] Stone, M. and Webber, B. (1998). Textual economy through close coupling of syntax and semantics. In *Proceedings of 1998 International Workshop on Natural Language Generation*, Niagra-on-the-Lake, Canada.
- [Walker, 1992] Walker, M. A. (1992). Redundancy in collaborative dialogue. In *Fourteenth International Conference on Computational Linguistics*, pages 345–351.
- [Walker, 1993] Walker, M. A. (1993). *Informational Redundancy and Resource Bounds in Dialogue*. PhD thesis, University of Pennsylvania.
- [Whittaker and Stenton, 1988] Whittaker, S. and Stenton, P. (1988). Cues and control in expert client dialogues. In *Proc. 26th Annual Meeting of the ACL, Association of Computational Linguistics*, pages 123–130.

## Acknowledgments

This research was supported by the National Science Foundation under Grant No. IRI-9314961.

# Verb Meanings, Object Affordances, and the Incremental Restriction of Reference

**Edward Kako** (ekako1@swarthmore.edu)  
Department of Psychology, Swarthmore College  
500 College Ave.  
Swarthmore, PA 19081 USA

**John C. Trueswell** (trueswel@psych.upenn.edu)  
Department of Psychology, University of Pennsylvania  
3401 Walnut St.  
Philadelphia, PA 19104 USA

## Abstract

There has traditionally been significant interest in the role of verb semantic restrictions in both psycholinguistic and computational theorizing about language interpretation (e.g., McRae, Spivey-Knowlton, & Tanenhaus, 1998; Resnik, 1996; Trueswell, Tanenhaus, & Garnsey, 1994). The bulk of this research has focused on how such information influences syntactic choices during parsing. The current paper explores in detail the time-course of, and mechanisms for, on-going referential processing. While their eye movements were recorded, subjects acted upon spoken instructions such as "Now I want you to fold the napkin." The verb was either highly constraining (e.g., "fold") or weakly constraining ("pick up"); the array contained either just one object with the appropriate affordances (the target) or two such objects (the target and a competitor). We provide evidence that listeners are capable of rapidly constraining the domain of reference of upcoming constituents to multiple objects with appropriate semantic affordances, which compete for referential consideration. Moreover, in relation to computational theorizing on this topic, the eyemovement patterns suggest that a verb's informativeness (i.e., the "tightness" of the semantic space of possible constituents, Resnik, 1996) affects the speed with which listeners can compute the domain of reference of upcoming constituents.

## Introduction

Psychologists have been interested in the process of language comprehension since the earliest days of generative grammar (Fodor & Bever, 1965; Miller & Isard, 1963; Slobin, 1966). Most comprehension studies have focused on the problem of syntactic ambiguity resolution – how listeners or readers decide among competing structural analyses (Caplan, Baker, & Dehaut, 1985; Crain & Steedman, 1985; Ferreira & Clifton, 1986; Frazier & Fodor, 1978; MacDonald, Pearlmuter, & Seidenberg, 1994; Trueswell et al., 1994, among many others). In the last several years, however, there has been a growing interest in on-line *semantic* interpretation – in particular, the extent to which listeners can use combinatory semantic information to determine the reference of words and phrases in a rapid, incremental fashion. Much of this work has been conducted in the so-called visual world paradigm, in which listeners manipulate the

contents of a miniature world as their eyes are tracked by a head-mounted visor (Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Tanenhaus, Spivey-Knowlton, & Sedivy, 1995). In this paper we provide evidence, using this paradigm, that the meanings of verbs become available to listeners rapidly enough to constrain the domain of reference for the upcoming direct object.

By virtue of what they mean, words often impose restrictions upon the semantics of other words that appear with them. Many prepositions impose restrictions on the geometric properties of their objects (see especially Landau & Jackendoff, 1993); *through*, for instance, requires that its object have some kind of hole. Verbs are especially picky in this regard: The subject of a verb must be the sort of thing that can perform the denoted action, and the direct object must be the sort of thing that can be sensed, manipulated, or changed in the relevant way. The verb *drink*, for instance, requires a subject capable of drinking, and a direct object capable of being drunk. Hence while *John drank the juice* sounds perfectly natural, both *The table drank the juice* and *John drank the table* register as distinctly odd. Although semantic restrictions have long played a role in linguistic theory (e.g., Chomsky, 1965; Jackendoff, 1972) and in the study of syntactic processing (Boland & Boehm-Jernigan, 1998; McRae et al., 1998; Tabossi, Spivey Knowlton, McRae, & Tanenhaus, 1994; Trueswell et al., 1994) little research has been done until recently to examine their potentially important role in on-line referential processing.

Using the visual world paradigm, Chambers, Eberhard, Carlson, and Filip (1998) have demonstrated rapid access to the meaning of the preposition *inside* and its use to restrict the referential domain of definite noun phrases. Participants in their experiment sat before an array of objects including a duck, a rope, a napkin, a can, and a whistle. When they were instructed to "Put the whistle inside the can," participants launched eye movements to the can even before the onset of the noun. The meaning of *inside* provided enough information for listeners to limit the referential domain of the upcoming noun phrase to the one object with the appropriate physical properties (or *affordances*, in the terms of Gibson, 1977). Crucially, such movements were not found when the preposition was *below* – which does not constrain the affor-

dances of its object – or when the array contained two additional objects with an interior volume (a bowl and a glass).

While highly suggestive, these results are limited in two ways. First, Chambers et al. used only one lexical item (*inside*), and so it is unclear whether semantic restrictions are rapidly available across a range of lexical items. Second, prepositions are generally considered to be closed-class items (Talmy, 1988), which differ from open-class items in a number of important ways, including frequency and semantic richness (Friederici, 1985; Gordon & Caramazza, 1985; Neville, Mills, & Lawson, 1992; Van Petten & Kutas, 1991). Perhaps it is the status of *inside* as closed-class that makes its semantic restrictions so readily available.

Evidence that bears on both of these concerns comes from a recent study by Altmann & Kamide (1999), who used a modified version of the visual world paradigm to explore the online processing of verbs. In their experiment, participants sat before a computer screen displaying several pieces of clip art: for example, a boy, a toy train, a toy car, a birthday cake, and a balloon. Listeners heard the scene described with one of two sentences: "The boy will move the cake," or "The boy will eat the cake." In the first case, multiple objects in the scene satisfied the semantic restrictions of the verb; in the second case, only the cake did so. Altmann and Kamide found that eye movements to the target were launched more rapidly after *eat*-type verbs (where the verb picked out only one object in the array) than after *move*-type verbs (where the verb picked out multiple objects). Looks to the target object were always delayed for *move*-type verbs until after hearing the definite NP *the cake*. These results suggest that semantic restrictions are rapidly available for open-class verbs as well as for closed-class prepositions, and across a range of lexical items.

Like the Chambers et al. (1998) experiment, the Altmann and Kamide study also has some features that limit what we can conclude about the on-line use of semantic restrictions. First, participants in one of their experiments had to indicate (with a button press) whether the sentence matched the visual scene (in half of all trials, the sentence did *not* match the scene). This metalinguistic judgment might have caused participants to process the incoming sentences in a strategic, non-natural fashion, perhaps encouraging them to focus more closely on verb information than they otherwise might have.<sup>1</sup>

More importantly, the two experiments reported in Altmann and Kamide provide conflicting evidence for the use of semantic restrictions to constrain referential domains. If listeners rapidly exploit the semantic restrictions of verbs to constrain the domain of reference, they should spend less time looking at inedible objects following *eat* than following *move*. Their graph of data from Experiment 1 confirms this prediction. But their graph of data from Experiment 2 re-

veals the opposite pattern: Participants spent more time fixating non-target objects after *eat* than after *move*. Further complicating interpretation of their results, Altmann and Kamide include in the category "Other" both non-target objects that meet the restrictions of the verb and non-target objects that do not meet those restrictions. It is therefore impossible to judge whether participants excluded incompatible objects from consideration altogether, as would be predicted by a model in which listeners restrict the referential domain rapidly and incrementally.

In reporting their data, Chambers et al. (1998) separate looks to other containers from looks to non-containers. Their data show some signs of early temporary consideration of the cohort of objects with the appropriate affordances (the target plus the other two containers). However, the proportion of early looks to each of these objects was only slightly greater than the proportion of early looks to an unrelated object. The fragmentation of attention among several objects in the multiple containers condition may have made it difficult to distinguish looks to the competitors from (presumably random) looks to unrelated objects. The precise time-course of referential restriction therefore remains uncertain.

In what follows, we report an experiment on the semantic restrictions of verbs using the visual world paradigm, with multiple lexical items and a condition with a single competitor. In this study, participants acted out spoken instructions like "Now I want you to fold the towel." On half of trials, the array contained just one object with the appropriate affordances (the target). On the other half, it contained both a target and one competitor (in this case, a napkin). Participants also acted out instructions like "Now I'd like you to pick up the towel" with precisely the same manipulation of competitor presence. While some verbs (e.g. *fold*) imposed strong semantic restrictions relative to the scene (picking out just one or two objects), other verbs (e.g. *pick up*) imposed only weak restrictions (potentially picking out all four objects).

Two aspects of our experiment should help to illuminate further both the time course and the causes of rapid referential restriction. First, we separate looks to compatible non-target objects from looks to incompatible non-target objects. Second, we include only one competitor in our trials, making it easier to distinguish looks to the competitor from random looks to unrelated (incompatible) objects in the display.

## Methods

### Participants

Sixteen undergraduates from the University of Pennsylvania participated in this study. They received either course credit or \$6.00. All were native speakers of English and had uncorrected vision or wore soft contact lenses.

### Stimuli

All critical instructions had the form "Now I want you to *verb* the *noun*" (followed in some cases by an additional phrase, such as "into the box"). We chose eight verbs with strong semantic restrictions, and four verbs with weak semantic restrictions (meaning that each weak verb was pre-

<sup>1</sup> Another version of the experiment eliminated the explicit metalinguistic component. But in that experiment, the participants – who did not participate in the prior version – were told that "in this version of the experiment, we aren't asking you to pay any particular attention to the sentences." This allusion to the prior study might have encouraged participants to strategize metalinguistically.



sented twice). As outlined in the Introduction, the experiment had a 2 (Restriction Strength: Strong versus Weak) x 2 (Competitor: Present versus Absent) design. Note that when the verb was Weak, the Competitor acted as such in name only, as the verb lacked the restrictions necessary to pick out a subset of the objects in the array.

Each list contained sixteen target trials; eight with Strong Verbs and eight with Weak Verbs. The design was such that subjects heard each Strong Verb only once, and manipulated each target object only once. The target trials in a list were evenly divided between the four conditions (with four trials per condition); conditions were rotated across lists, resulting in four lists. All trials consisted of two instructions: the critical sentence followed by a second instruction, which asked participants to further manipulate the Target (e.g., "Now I want you to fold the towel. Now cover the box with it").

Target trials were accompanied by sixteen filler trials that used other verbs and involved the manipulation of other objects. Order of target and filler trials within a list was determined by random assignment, with two constraints: first, that there be no more than two consecutive target trials using the same verb type; and second, that critical trials and filler trials alternated. To control for order of presentation, each list was presented in one of two orders, one the reverse of the other.

Prior to each instruction, participants were told to "Look at the cross" (the central fixation point on the table). Instructions were digitally recorded and played from a laptop computer connected to a pair of external speakers. Post experiment interviews revealed that subjects were unaware of the manipulation or intent of the experiment.

## Procedure

Eye movements were monitored with an ISCAN head-mounted eye-tracker. The device had two cameras: One recorded the visual environment from the perspective of the participant's left eye, and the other recorded a close-up image of the left eye. A computer analyzed the eye image in real time, superimposing the horizontal and vertical eye position on the scene image; this composite image was recorded to tape using a frame-accurate digital video recorder. The tracker determined eye position by following the relative positions of the pupil and the corneal surface reflection, thereby canceling out errors in eye position that might result from slippage of the visor. Moreover, because the scene and eye cameras were attached to the visor, tracking accuracy was not affected by movements of the participant's head.

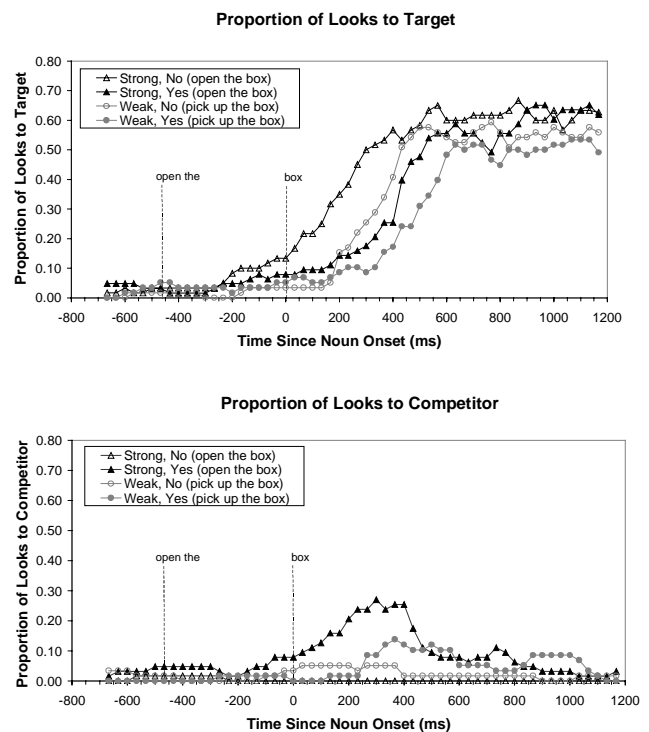
Participants were asked to carry out the instruction as quickly as they could. The entire experiment lasted approximately half an hour.

## Results

The digital videotape of each participant's scene and eye-position was analyzed by using the slow motion and freeze frame viewing on a digital VCR. For each trial, the frame number corresponding to the onset of the spoken instruction was noted. Then, the location and onset time of each successive fixation on an object was recorded by inspecting the video frame images until 1 sec after the offset of the instruc-

tion. Trials were not included in the analysis if the tracking signal became degraded during the critical portion of the sentence, which was defined as lasting from the onset of the verb until 1 sec after the offset of the instruction. Of the 256 trials, 16 (6.25%) were not included in the analyses.

Figure 1 presents the fixation probabilities over time in 33-ms intervals (the sampling rate of the VCR), for the Target (the upper graph) and the Competitor (the lower graph). The data are plotted relative to the onset of the noun, corresponding to zero milliseconds on the X-axis. The onset of the verb occurred an average of 485 milliseconds prior to the noun, and is marked by a vertical bar above the X-axis. The probabilities do not sum to zero because the plot omits the probabilities of fixating the cross or the other two objects. The probability of fixating either the cross or the other two objects did not differ across conditions.



**Figure 1:** Proportion of trials with fixations to target (top) and competitor (bottom).

In the Strong Verb, Competitor Absent condition, there were early looks to the Target (the open circles in the upper graph) and essentially no looks to the Competitor Replacement (the open circles in the lower graph). Early consideration of the Target begins in this condition prior to the onset of the noun, and rises rapidly during the first 250 msec of the noun. By contrast, in the Strong Verb, Competitor Present condition (e.g., when the array contained both a towel and a napkin), looks to the Target (the filled circles in the upper graph) were reduced, as participants temporarily considered the Competitor (the filled circles in the lower graph). Interestingly, participants evenly distributed their early inspection of the scene between the objects that had the appropriate affordances (e.g., the two foldable objects).

Instructions containing Weak verbs (e.g., "pick up") exhibited a different pattern of fixations. Fixations on the Target (the triangles in the upper graph) were delayed until after the onset of the noun. The greatest delay occurred in the Weak Verb, Competitor Present condition (the filled triangles in the upper graph). This time period was marked by some temporary consideration of the Competitor (the filled triangles in the lower graph). This competition presumably reflects minor confusion arising from perceptual similarity between the Target and Competitor. For instance, a few "accidental" looks to the towel ought to be expected upon hearing "napkin" in the instruction "Now I'd like you to pick up the napkin...". Consistent with this explanation, competition in this condition is small and appears after onset of the noun.

### Early Looks to the Target

In order to assess whether early looks to the Target occurred more often in the Strong Verb, Competitor Absent condition than in the other three conditions, we averaged the proportion of time spent fixating the Target during a time slice corresponding to 233 ms after the onset of the verb until 233 ms after the onset of the noun (see Table 1). Because it takes approximately 200-250 ms for the eyes to respond to phonemic input in word recognition studies using this paradigm (e.g., Allopenna, Magnuson, & Tanenhaus, 1998), any significant differences during this portion of the speech are unlikely to be attributable to the perception of the noun (e.g., Allopenna et al., 1998). To test differences, subject and item means were entered into separate Analyses of Variance (ANOVAs) with three factors: Verb Type (Strong, Weak); Competitor (Absent, Present) and Presentation List/Item Group (4 lists in the subject analysis and 4 item groups in the item analysis).<sup>2</sup> These analyses revealed a reliable effect of Verb Type ( $F(1,12)=24.70$ ,  $p<0.001$ ;  $F(1,12)=15.15$ ,  $p<0.005$ ) with Strong verbs showing more early looks to the Target than Weak verbs. There was also a marginal effect of Competitor Presence ( $F(1,12)=3.27$ ,  $p<0.1$ ;  $F(1,12)=3.91$ ,  $p<0.1$ ). There was an interaction between Verb Type and Competitor Presence that was significant in the subject analysis and marginally significant in the item analysis ( $F(1,12)=5.65$ ,  $p<0.05$ ;  $F(1,12)=4.32$ ,  $p=0.06$ ). Simple effects tests showed that Strong verbs had an advantage over Weak verbs when the Competitor was Absent ( $F(1,12)=16.28$ ,  $p<0.005$ ;  $F(1,12)=14.30$ ,  $p<0.005$ ) but not when it was Present ( $F(1,12)=1.64$ ;  $F(1,12)=0.89$ ).<sup>3</sup>

<sup>2</sup> All ANOVAs were conducted on an arcsine transformation of the data, arcsine  $((2*p)-1)$ . This was done to adjust for the fact that the proportion  $p$  is bounded at 0 and 1. ANOVAs conducted on untransformed data yielded similar statistical patterns.

<sup>3</sup> For looks to Target only, there were some uninterpretable interactions with the List factor.

**Table 1A:** Proportion of Looks to the Target

Time Slice 1: (Verb + 233ms) to (Noun+233ms)		
	Competitor Present	
	YES	NO
Strong Verb	0.07	0.17
Weak Verb	0.05	0.04

Time Slice 2: (Noun + 233ms) to (Noun+767ms)		
	Competitor Present	
	YES	NO
Strong Verb	0.39	0.56
Weak Verb	0.29	0.44

**Table 1B:** Proportion of Looks to the Competitor

Time Slice 1: (Verb + 233ms) to (Noun+233ms)		
	Competitor Present	
	YES	NO
Strong Verb	0.09	0.01
Weak Verb	0.03	0.01

Time Slice 2: (Noun + 233ms) to (Noun+767ms)		
	Competitor Present	
	YES	NO
Strong Verb	0.16	0.00
Weak Verb	0.09	0.03

### Early Looks to the Competitor

Similar ANOVAs were conducted on the mean proportion of early looks to the Competitor during this time slice (see Table 1). As can be seen in the table, most looks to the Competitor occurred in the Strong verb condition when the Competitor was Present. The analysis revealed a reliable interaction between Competitor presence and Verb Type ( $F(1,12)=7.97$ ,  $p<0.05$ ;  $F(1,12)=18.99$ ,  $p<0.005$ ), a marginal effect of Competitor Presence ( $F(1,12)=3.98$ ,  $p<0.07$ ;  $F(1,12)=9.79$ ,  $p<0.05$ ) and no effect of Verb Type ( $F(1,12)=3.25$ ;  $F(1,12)=3.06$ ). Simple Effects showed an effect of Verb Type when the Competitor was Present ( $F(1,12)=6.50$ ,  $p<0.05$ ;  $F(1,12)=10.16$ ,  $p<0.01$ ) but not when it was Absent ( $F(1,12)=2.77$ ;  $F(1,12)=2.87$ ).

To assess any preference for looking at the Target over the Competitor during this time slice, two-tailed t-tests on subject and item means were done comparing looks to the Target with looks to the Competitor. To avoid Type I errors, we corrected for the number of tests by dividing the alpha by four. As expected, the only reliable difference arose in the Strong Verb, Competitor Absent condition, where there were significantly more looks to the Target than to the Competitor Replacement (e.g., a Coke can) ( $t(15)=3.48$ ,  $p=0.003$ ;  $t(15)=3.96$ ,  $p=0.001$ ).

### Later Looks to the Target.

We also quantified looks to the Target and Competitor in a second time slice, corresponding to approximately 500 ms after the first time slice (i.e., from 233 ms after the onset of the noun until 767 ms after the onset of the noun; see Table

1). Differences in this region are more likely to be affected by the perception of the target noun phrase. ANOVAs revealed a main effect of Competitor Presence ( $F(1,12)=10.71$ ,  $p<0.01$ ;  $F(1,12)=11.60$ ,  $p<0.01$ ), with more looks to the Target when the Competitor was Absent. In addition, there was a marginal effect of Verb type ( $F(1,12)=9.15$ ,  $p<0.05$ ;  $F(1,12)=3.91$ ,  $p<0.08$ ) with more looks to the Target when the sentence contained a Strong verb. There was no interaction between these factors ( $F_s<1$ ).

### Later Looks to the Competitor

ANOVAs on the mean proportion of time spent looking at the Competitor in this region revealed a main effect of Competitor ( $F(1,12)=11.48$ ,  $p<0.01$ ;  $F(1,12)=39.10$ ,  $p<0.001$ ), no effect of Verb Type ( $F(1,12)=0.38$ ;  $F(1,12)=1.52$ ) and a weak interaction between these factors that was significant only in the item analysis ( $F(1,12)=2.72$ ;  $F(1,12)=6.64$ ,  $p<0.05$ ).

### Discussion

We have presented evidence that the semantic restrictions of verbs become available rapidly enough during comprehension to permit listeners to make predictions about the likely reference of the upcoming direct object. Participants looked more rapidly at the referent of the direct object when the verb had Strong restrictions than when it had Weak ones. For instance, they looked more rapidly at the towel when told to fold it than when told to pick it up. When the scene included a second foldable object, the use of a Strong restrictions verb resulted in early temporary consideration of this second object, which competes with the target object. This pattern replicates the one reported both by Chambers (1998) and by Altmann and Kamide (1999), with several improvements: We used multiple lexical items, a task less likely to induce listener strategies<sup>4</sup>, and a single competitor. The last improvement allowed us to show that listeners rapidly eliminated incompatible non-target objects from consideration.

While it seems clear that semantic restrictions are rapidly available for referential restriction, the precise mechanism of this restriction remains unclear. There are two explanations for the source of this restriction. Listeners might launch eye movements after hearing a strongly constraining verb because they have already assessed the properties of the objects in the display and recognize that only a subset of those objects is compatible with the verb's restrictions. By con-

trast, listeners might launch eye movements simply because a strongly restricting verb is more likely than a weakly restricting *in any context* to pick out a unique referent (or subset of referents). We will refer to these possibilities as *affordance matching* and *informativeness*, respectively. In fact, the notion of informativeness has been quantified in recent computational theorizing by (Resnik, 1996), who also provides evidence that verb informativeness has very real psycholinguistic consequences.

These two possibilities make different predictions about the likelihood of launching an eye movement just after hearing the verb. If listeners actively match affordances, they should launch eye movements as soon as they determine that one or more objects in the scene satisfy the restrictions of the verb. Thus, they should be equally likely to launch eye movements following a weak verb as following a strong verb, because in both cases, at least one object in the array satisfies the restrictions of the verb; in both cases, interrogation of the array can begin immediately. If, on the other hand, eye movements are triggered by a verb's informativeness, listeners should be more likely to launch eye movements following a strongly constraining verb than following a weakly constraining one, as informative verbs carry enough information to identify their direct objects, whereas weakly informative verbs do not.

To test between these possibilities, we examined the proportion of fixations on any object in two time slices: from 233 ms after the onset of the verb until 233 ms after the onset of the noun, and from 233 ms after the noun to 767 ms after the noun (the same slices used in the analyses presented in the Results section). As Table 2 indicates, listeners were more likely to launch a fixation to any object following a Strong verb than following a Weak verb. In Time Slice 1, the effect of verb type was reliable in the subject analysis, and marginal in the item analysis ( $F(1,12)=9.71$ ,  $p<0.01$ ;  $F(1,12)=4.36$ ,  $p<0.06$ ). In Time Slice 2, the effect of verb type was reliable in both analyses ( $F(1,12)=8.28$ ,  $p<0.02$ ;  $F(1,12)=7.96$ ,  $p<0.02$ ).

**Table 2:** Proportion of looks to any object

Time Slice 1: (Verb + 233ms) to (Noun+233ms)		
	Competitor Present	
	YES	NO
Strong Verb	0.22	0.22
Weak Verb	0.16	0.12
Time Slice 2: (Noun + 233ms) to (Noun+767ms)		
	Competitor Present	
	YES	NO
Strong Verb	0.60	0.63
Weak Verb	0.46	0.52

While these data are somewhat preliminary, they suggest that a verb's informativeness, independent of context, contributes to the speed with which listeners can compute the domain of reference of upcoming constituents. Because a Strong verb is highly informative about its upcoming direct object, listeners can begin to interrogate the visual scene for

<sup>4</sup> It is of course possible that listeners in our task developed strategies that resulted in them unnaturally focusing on particular classes of information (see Tanenhaus & Spivey-Knowlton, 1996 for a discussion of this issue). However, if strategies were developed to use verb restrictions, we might expect their effects to emerge over the course of the experiment. We tested this possibility in two ways. We inspected the first half of the trials in the experiment, and we inspected the first of paired items in the experiment. In both cases, the pattern of eye movements was similar to the overall pattern, i.e., early looks to the Target in the Strong Verb Competitor Absent condition, and some early looks to the Competitor in the Strong Verb Competitor Present condition.

an object with the appropriate affordances before they have heard the noun phrase.

Whether early eye movements are driven by informativeness or affordance matching, it is clear that verb meanings can be accessed rapidly enough to make predictions about the reference of an upcoming direct object, and to constrain the set of entities to which the direct object might refer. The current findings contribute to a growing body of data that support a view of semantic interpretation as both incremental and predictive. Words not traditionally thought to carry reference – prepositions (Chambers et al., 1998), adjectives (Sedivy et al., 1999), and verbs (Altmann & Kamide, 1999, the present study) – can be exploited by listeners to predict the reference of upcoming nouns. Indeed, the linking of speech to a mental model of the world appears to be an active, continuous process.

### Acknowledgements

We thank Sarah Brown-Schmidt and Jared Novick for their considerable help in running the experiments, and in coding and analyzing the data. We also thank the members of the eyetracking lab group at the Institute for Research in Cognitive Science at Penn for their many helpful suggestions. This work was supported by NIH National Research Service Award #HD085070-01 to the first author while a postdoctoral fellow at IRCS; and by NSF grant #SBR-96-16833 and NIH grant 1-R01-HD37507001, both to the second author.

### References

Alloppenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38, 419-439.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247-264.

Boland, J. E., & Boehm-Jernigan, H. (1998). Lexical constraints and prepositional phrase attachment. *Journal of Memory and Language*, 39(4), 684-719.

Caplan, D., Baker, C., & Dehaut, F. (1985). Syntactic determinants of sentence comprehension in aphasia. *Cognition*, 21, 117-175.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Carlson, G. N., & Filip, H. (1998). Words and worlds: The construction of context for definite reference. *Proceedings of the Annual Conference of the Cognitive Science Society*, 20.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Crain, S., & Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological syntax processor. In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 320-358). New York: Cambridge University Press.

Ferreira, F., & Clifton, C. J. (1986). The independence of syntactic processing. *Journal of Memory and Language*, 25, 348-368.

Fodor, J. A., & Bever, T. G. (1965). The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior*, 4, 414-420.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage model. *Cognition*, 6, 291-325.

Friederici, A. D. (1985). Levels of processing and vocabulary types: Evidence from on-line comprehension in normals and agrammatics. *Cognition*, 21, 133-166.

Gibson, J. J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 67-82). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gordon, B., & Caramazza, A. (1985). Lexical access and frequency sensitivity: Frequency saturation and open/closed class equivalence. *Cognition*, 21, 95-115.

Jackendoff, R. (1972). *Semantic interpretation in generative grammar*. Cambridge, MA: MIT Press.

Landau, B., & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *The Behavioral and Brain Sciences*, 16, 217-265.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4), 676-703.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38, 283-312.

Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, 2, 217-228.

Neville, H. J., Mills, D. L., & Lawson, D. S. (1992). Fractionating language: Different neural subsystems with different sensitive periods. *Cerebral Cortex*, 2, 244-258.

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61, 127-159.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71, 109-147.

Slobin, D. I. (1966). Grammatical transformations and sentence comprehension in childhood and adulthood. *Journal of Verbal Learning and Verbal Behavior*, 5, 219-227.

Tabossi, P., Spivey Knowlton, M. J., McRae, K., & Tanenhaus, M. K. (1994). Semantic effects on syntactic ambiguity resolution: Evidence for a constraint-based resolution process. In U. Carlo & M. Moscovitch (Eds.), *Attention and performance 15: Conscious and nonconscious information processing* (pp. 589-615). Cambridge, MA: MIT Press.

Talmy, L. (1988). The relation of grammar to cognition. In B. Rudzka-Ostyn (Ed.), *Topics in cognitive linguistics* (pp. 165-205). Philadelphia: John Benjamins Publishing Co.

Tanenhaus, M. K., & Spivey-Knowlton, M. J. (1996). Eye-tracking. *Language and Cognitive Processes*, 11(6), 583-588.

Tanenhaus, M. K., Spivey-Knowlton, M. J., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632-1634.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33, 285-318.

Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory & Cognition*, 19, 95-112.

# Learning the Use of Discourse Markers in Tutorial Dialogue for an Intelligent Tutoring System

Jung Hee Kim  
(janice@steve.iit.edu) IIT  
Michael Glass  
(michael.glass@iit.edu) IIT  
Reva Freedman  
(freedr+@pitt.edu) LRDC  
Martha W. Evens  
(evens@iit.edu) IIT

Department of Computer Science  
Illinois Institute of Technology  
10 W. 31st St.  
Chicago, IL 60616

Learning Research and Development Center  
3939 O'Hara St.  
Pittsburgh, PA 15260

## Abstract

Usage of discourse markers in tutorial language can make the difference between stilted and natural sounding dialogue. In this paper we describe some simple rules for selection of discourse markers. These rules were derived for use in an intelligent tutoring system by applying decision-tree machine learning to human tutoring language. The fact that these selection rules operate within the environment of an intention-based planner encouraged us to derive our decision tree partly based on intention-based features. The resulting tree, when applied to the generation task, is relatively easy to understand because it can be referred to traditional intention-based linguistic explanations of discourse marker behavior.

## Introduction

CIRCSIM-Tutor (CST) is a natural language-based intelligent tutoring system that engages the student in Socratic-style dialogue. The goal of the CST project is to imitate fluent simplified human tutoring language, both in the choice of tutorial dialogue strategies and in the use of language.

One feature of fluent dialogue is the use of discourse markers such as “so,” “and,” and “now,” which often occur at structural boundaries in the discourse. Discourse markers, also known as cue words, have as many different descriptions as people describing them. In Grosz and Sidner's (1986) procedural description of discourse, discourse markers flag changes in both attentional and intentional state. In Rhetorical Structure Theory, discourse markers mark rhetorical relations between segments (Mann and Thompson, 1988). The grammar of Quirk et al. (1985, pp. 632 *ff*) subsumes most discourse markers within conjunctions. Stenstrom's (1994) manual on analyzing

discourse emphasizes their use as marking boundaries of topics and digressions and describes them in concert with interpersonal “interactional signals.” Schiffrin (1987) provides a detailed accounting of the behavior and purpose of eleven discourse markers without being tied to a particular theory of discourse or syntax. Schiffrin also provides an operational definition of discourse markers, giving evidence that discourse markers have functions such as aiding coherence and cohesion in text. Halliday and Hassan (1976) in their book on cohesion describe the function of quite a number of discourse markers in detail.

Recently there have been attempts to describe the behavior of discourse markers in computationally useful ways by applying methods of machine learning and corpus linguistics. Litman (1996) devised rules for distinguishing between semantic and structural uses of discourse markers in transcribed speech. In sharp distinction to the more traditional linguistic accounts, the rules are based largely on observable features such as the length of phrases, preceding and succeeding cue words, and prosodic features. Moser and Moore (1995) divided instructional dialogue into discourse segments and coded various relationships between them according to Relational Discourse Analysis, which combines Grosz and Sidner's type of analysis with Rhetorical Structure Theory. They derived rules for a number of aspects of discourse marker usage, including placement and occurrence vs. omission. Di Eugenio, Moore, and Paolucci (1997) studied the same dialogues toward similar ends. Nakano and Kato (1999) studied Japanese instructional dialogue, using machine learning to derive rules for occurrence of three categories of discourse markers. They divided their text into segments in the same manner as RST, but also coded the instructional goals for

each segment in addition to coding the kinds of features used in previous studies.

The addition of instructional goals in Nakano and Kato's study is important to the CIRCSIM-Tutor project, and should be encouraging from the standpoint of trying to generate (as opposed to analyze) instructional dialogue. One reason is that instructional goals proved to be explanatory. A common feature of the machine learning studies is that the text is coded for a large number of features, of which only a few are incorporated by the machine learning process into the eventual rules or decision tree. In Nakano and Kato's study instructional goals were so incorporated, meaning that they were more explanatory than many of the other features. This is congruent with non-corpus-based linguistic theories that explain discourse markers in terms of the speaker's intentions.

The speaker's intentions are rarely explicit in text; for purposes of analysis intentions are divined by coders. However when the machine tutor is generating dialogue, the machine speaker's "intentions," i.e. the tutorial goals, can be given in the form of planning goals, see for example (Young, Moore & Pollack, 1994). Nakano and Kato have shown that having the tutorial goal structure in hand can potentially lead to better discourse marker selection.

In this paper we use attribute-based machine learning of decision trees, specifically the C4.5 algorithm (Quinlan, 1993), to investigate discourse marker selection. We make use of both structural features and aspects of the sequence of tutorial goals—the "intention" of the machine tutor. Although we learn rules from transcripts of human dialogues, we concentrate on features that are available within the CIRCSIM-Tutor generation environment.

The machine tutor does not reason about rhetorical relations such as are usually used to explain discourse markers. Instead it has planning goals that produce schemata containing patterns of dialogue. These schemata define the dialogue segments. Rhetorical relations are implicit in the patterns, so it is possible to relate goal-structure explanations of discourse markers to the rhetorical relation-based theories.

## The Experiment

We recorded the features surrounding instances of discourse markers in human tutorial dialogue, then derived a decision tree to predict discourse marker selection.

The users of CIRCSIM-Tutor are medical students in a first-year physiology class studying the reflex control of blood pressure. Students are required to predict the changes in a set of physiological variables, after which the tutor endeavors to elicit corrected predictions via Socratic-style dialogue, asking questions and giving hints. CST's conversation can be largely segmented into the correction of individual variables.

The CIRCSIM-Tutor project has transcripts of one- and

two-hour keyboard-to-keyboard tutoring sessions between physiology professors and medical students. Our construction of the computer tutor's planning operators and tutorial language is informed by these transcripts. The transcripts were previously marked up with tutorial goals and language phenomena for this purpose (Kim, Freedman & Evens, 1998a, b; Freedman et al., 1998; Zhou et al., 1999). Tutorial goals consist of global goals for tutoring and local goals for maintaining coherence of dialogues. The global goals used in this study are hierarchically arranged into *method* and *topic* levels. A *method* goal describes one way to remediate a student's incorrectly predicted physiological variable. Within one method, a sequence of *topic* goals describes individual concepts to be expressed. A topic can be expressed by either telling the information to the student or eliciting it from the student. A typical dialogue pattern for the correction of one individual variable is as illustrated in Figure 1. The sequence of tutorial goals is as follows:

- The variable to be corrected is introduced into the conversation.
- Various topic goals are realized by telling them to the student or eliciting them from the student.
- The corrected prediction is elicited from the student.

The discourse markers we study in this paper occur at the boundaries between topic goals, as shown in italics in Figure 1. We are concerned with the selection of these discourse markers in human tutorial dialogues in order to generate them correctly. Placement of discourse markers is not an issue, we ignore discourse markers which occur elsewhere.

It will be noted that in our dialogues the junctures between topic goals do not always coincide with the turn boundaries; in fact in our illustration one topic is spread among three turns and one turn encompasses parts of three topics. One typical tutor turn contains:

- An optional acknowledgment of the student's answer
- Possibly an elaboration on that answer
- Possibly some new information
- A question or instruction to the student

(Freedman & Evens, 1996)

The context of a discourse marker therefore includes not only the structure of topic goals, but also information from the turn structure. Preceding the first discourse marker in a tutor's turn is a possible tutor's acknowledgment to the student and possibly some elaboration. Furthermore there is the student's immediately preceding turn, which usually consists of the answer to the tutor's previous question. Some examples of these features, including our characterization of the correctness of the student's answer, are also annotated in Figure 1.

The human transcripts also contain dialogue that is too complex for us to mark up according to our goal hierarchy

and is therefore excluded from our sample.

We further restricted ourselves to exchanges where the student gave answers that were correct or “near misses.” A near miss is a student answer that is true but not expected, and can be repaired without contradicting the student (Zhou et al., 1999). In the dialogue in Figure 1, the tutor repaired the student’s overly specific answer by echoing back the more general answer. Sometimes the tutor temporarily suspends the current topic goal and interpolates a tutoring schema to repair the unexpected answer. In that case the goal hierarchy would show an inner sequence of topic goals devoted to remediating one outer topic. These instances are included in our sample. The tutor’s responses to incorrect student answers (as opposed to near misses) are too varied for us to obtain any regularities in discourse marker usage, so we excluded them.

We extracted instances of the discourse markers “and,” “so,” and “now” because these are the most frequently used ones in our transcripts. Each instance consists of the context around one discourse marker coinciding with a topic change, coded for the following five attributes:

- Category of the student’s answer preceding the marked topic boundary: correct, near miss, or N/A. The N/A case occurs when the tutor covers several topics within one turn, so the topic preceding the discourse marker does not contain a student answer.
- Presence or absence of acknowledgment preceding the topic boundary: ack, no-ack, N/A.
- Discourse marker: “and,” “now,” “so.”
- Position within the sequence of topic goals of the topic following the discourse marker: introduce, initial, middle, or final.
- Presentation of the topic following the discourse marker: inform or elicit.

Thus the sentence “and the reflex hasn’t started to operate yet” from turn 3 of Figure 1 is coded as:

- Student’s answer category = “near miss”
- Acknowledgement = “present”
- Discourse marker = “and”
- Position in sequence = “middle”
- Type of presentation = “inform”

We supplied 60 cases of these feature-annotated discourse marker occurrences to the C4.5 machine learning program. It produced the following rules for selection of the discourse marker:

- If the topic position is introduce then use “now”
- If the topic position is middle then use “and”
- If the topic position is final then use “so”
- If the topic position is initial  
and if the presentation is inform then use “so”  
else {presentation is elicit} use “and”

These rules misclassified 8 of the 60 cases, for an error rate

of 13.3%.

These rules describe our expert tutors’ linguistic behavior, predicting which discourse marker will be selected in certain contexts. We start with this description in order to produce rules for text generation.

## Discussion

Most of the predictions of the derived rules can be explained by existing discourse marker theories. The “now” on the introduction topic is consistent with the explanation by Grosz and Sidner (1986) of marking an attentional change, creating a new focus space of salient objects and topics. Schiffrin (1987, p. 230) says “...‘now’ marks a speaker’s progression through discourse time by displaying attention to an upcoming idea unit.” In fact, this reading of “now” explains some of the cases of “now” that are misclassified by the derived rules. These are cases where the tutor does not explicitly utter an introduce topic at the beginning of the segment, with the result that the attention-shifting “now” is attached to the initial topic. Here is one example:

Now, what two parameters in the prediction table together determine the value of SV?

Although the derived rules misclassify our marked-up transcripts in these cases, for the purpose of generating sentences in the machine tutor this is a useful discovery. The intention to shift tutoring to a new variable is available in CIRCSIM-Tutor’s tutorial goal structure, even if not always expressed in text, so the text generator can plausibly know to emit “now.”

Most of the remaining predictions of the derived rules can be explained by existing discourse marker theory. Shiffrin (1987) and Halliday and Hassan (1976) and Quirk et al. (1985, p. 638) all describe “so” as indicating a result. In our derived rules, the “so” attached to the final topic is used in this fashion. The final sentence of turn 3 in Figure 1 illustrates this point.

When the rules predict “so” attached to the initial topic it has a different role. It is found in what we call the *present-anomaly* tutoring method used to point out the inconsistent appearance of reported facts, viz:

So, in DR heart rate is up, cardiac output is up, but stroke volume is down. How is this possible?

This “so” is explained by Halliday and Hassan as “a statement about the speaker’s reasoning process” meaning it is logical to be having this thought right now.

The discourse marker “and” usually occurs on medial topics to “coordinate and continue” the topics (Schiffrin, 1987, p. 152), and needs no explaining. The discourse marker “and” occurring on the initial topic seems anomalous, but it occurs in the context of a tutorial schema we call *move forward*. This schema attempts to persuade the

student of the correct value of a new physiological variable based on the result of the immediately preceding discussion of a different variable. Here is an example:

Tu: ...That being the case, what will happen to right atrial pressure in this situation?

St: Increase.

Tu: *And* if right atrial pressure increases, what would happen to stroke volume?

In this example, the final topic in the first segment occurred when the student produced the correct value for right atrial pressure. The tutor skipped introducing the next variable, stroke volume, and proceeded directly to the initial topic of the tutoring schema for its correction, which moves forward in causal physiological reasoning from the final topic in the preceding segment. In this case “and” is warranted, it would seem that “so” would be equally appropriate. This is another instance where the CIRCSIM-Tutor text generator makes use of the discourse goal being processed. Even though tutoring of a new variable usually starts with the discourse marker “now,” when the new variable is taught by the *move forward* method goal then the generator emits “and” instead.

Except for the initial discourse marker (usually “now”) at the beginning of a tutoring method schema, it is possible to apply to our own data Di Eugenio et al.’s (1997) discoveries relating rhetorical structure to discourse marker occurrence. Although we did not perform any rhetorical structure analysis on our texts, most of our method schemas fit one of their patterns, as described next.

Here is an idealized realization of a typical CIRCSIM-Tutor method schema for teaching a variable, called *tutoring via determinants*:

Tu: What are the determinants of cardiac output?

St: Heart rate and stroke volume.

Tu: And what is the relation of stroke volume to cardiac output?

St: Direct.

Tu: And we have already seen that heart rate is unchanged. So what happens to cardiac output?

St: It goes up.

In order to analyze this in terms of rhetorical relations, we write down all the propositions in the sequence they occur as if it were a monologue, thereby exposing the argument in simplest form. Since the intention of each of the tutor’s questions is to cause the student to believe the corresponding assertion, we think this is a reasonable model.

- a) The determinants of cardiac output are heart rate and stroke volume,
- b) And stroke volume affects cardiac output directly,
- c) And heart rate is unchanged,
- d) So cardiac output goes up.

In the terms of Relational Discourse Analysis, proposition d) is the *core* while a), b), and c) are *contributors*. The intentional relationship between each contributor and the core is *convince*. In fact, most of our methods have the same structure: the core is the last statement, where the value of the variable is finally understood, and the contributors all argue for the truth of the core. In (Di Eugenio et al., 1997) these relations are all analyzed in the “core2” class, meaning that the core follows the contributor in the text. Their decision tree on discourse marker occurrence yields a simple answer for these cases: the discourse marker should ordinarily appear.

## Conclusions

We have applied decision tree learning to transcripts of expert tutors in order to learn rules that predict discourse marker selection. Our purpose in this endeavor is not to find rules for analyzing texts, but to produce rules for text generation in CIRCSIM-Tutor. Discourse marker usage has traditionally been explained partly in terms of the intention of the speaker and partly in terms of the rhetorical structure of the text. Neither is explicit in transcripts of discourse, but must be imputed by researchers before analyses of discourse markers can proceed. Recent work in using machine learning to explain discourse marker usage has thus shied away from using intention-based explanations.

However within the context of the machine tutor the generation algorithm has access to the speaker’s intentions. In CIRCSIM-Tutor these intentions are the pedagogical goals. The structure of these goals implies the rhetorical structure of the text to be generated. So without explicit reasoning in the rhetorical terms that usually explain discourse markers, simply examining the current goals enables the text generator to select the correct discourse marker.

Our machine-derived decision tree analysis of discourse marker selection is quite successful. The features that drove the machine learning process included the same pedagogical goal analysis as is used by the machine tutor. The decision tree that results was examined by hand; where it incorrectly predicts observed data the decisions can be enhanced by applying traditional linguistic explanations. The fact that this decision tree is intention-based enables us to correlate it to existing linguistic descriptions of discourse marker usage.

## Acknowledgments

Joel A. Michael and Allen A. Rovick, professors of physiology at Rush Medical College, are responsible for the pedagogical and domain knowledge in CIRCSIM-Tutor, and served as expert tutors for the transcripts. Yujian Zhou helped bring machine learning to the CIRCSIM-Tutor project, and has been helpful in all endeavors.

This work was supported by the Cognitive Science Program, Office of Naval Research under Grant No.



N00014-94-1-0338 to Illinois Institute of Technology. The content does not reflect the position or policy of the government and no official endorsement should be inferred.

## References

- Di Eugenio, Barbara, Johanna D. Moore, and Massimo Paolucci. (1997). Learning features that predict cue usage. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL '97)*, Madrid, Spain, pp. 80–87.
- Freedman, Reva and Martha W. Evens. (1996). Generating and Revising Hierarchical Multi-Turn Text Plans in an ITS. In C. Frasson, G. Gauthier and A. Lesgold, eds., *Intelligent Tutoring Systems: Third International Conference (ITS '96)*, Montreal. Berlin: Springer-Verlag. Lecture Notes in Computer Science, no. 1086.
- Freedman, Reva, Yujian Zhou, Michael Glass, Jung Hee Kim, and Martha W. Evens. (1998). Using rule induction to assist in rule construction for a natural-language based intelligent tutoring system. *Proceedings of Twentieth Annual Conference of the Cognitive Science Society*, Madison, WI, pp. 362–367.
- Grosz, Barbara J. and Candace L. Sidner. (1986). Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3): 175–204. Also published as a technical report in 1986 with the same title: Technical Report no. 380 from the Center for the Study of Reading, University of Illinois at Urbana-Champaign.
- Halliday, M. A. K. and Ruqaiya Hasan. (1976). *Cohesion in English*. London: Longman.
- Kim, Jung Hee, Reva Freedman, and Martha W. Evens. (1998a). Relationship between tutorial goals and sentence structure in a corpus of tutoring transcripts. *Proceedings of Ninth Midwest AI and Cognitive Science Conference(MAICS '98)*, Dayton, OH, pp. 124–131.
- Kim, Jung Hee, Reva Freedman, and Martha W. Evens. (1998b). Responding to unexpected student utterances in CIRCSIM-Tutor v. 3: Analysis of transcripts. *Proceedings of the Eleventh Florida AI Research Symposium (FLAIRS '98)*, Sanibel Island, FL, pp. 153–157.
- Litman, Diane J. (1996). Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research*, 5: 53–94.
- Mann, William C. and Sandra A. Thompson. (1988). Rhetorical Structure Theory: Towards a functional theory of text organization. *Text*, 8(3): 243–281.
- Moser, Megan and Johanna D. Moore. (1995). Investigating cue selection and placement in tutorial discourse. *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics (ACL '95)*, Boston, MA, pp. 130–135.
- Nakano, Yukiko I. and Tsuneaki Kato. (1999). Cue phrase selection in instruction dialogue using machine learning. *Discourse Relations and Discourse Markers, Proceedings of the Workshop of COLING-ACL '98*, Montreal, pp. 100–106.
- Quinlan, J. Ross. (1993). *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- Schiffrin, Deborah. (1987). *Discourse Markers*. Cambridge: Cambridge University Press.
- Stenström, Anna-Brita. (1994). *An Introduction to Spoken Interaction*. London: Longman.
- Young, R. Michael, Johanna D. Moore and Martha E. Pollack. (1994). Towards a principled representation of discourse plans. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, Atlanta.
- Zhou, Yujian, Reva Freedman, Michael Glass, Joel A. Michael, Allen A. Rovick, Martha W. Evens. (1999). What should the tutor do when the student cannot answer a question? *Proceedings of the Twelfth International Florida AI Research Symposium (FLAIRS '99)*, Orlando, pp. 187–191.

Turn	Text	Global Tutoring Goal	Other Features
1. Tu:	Now let's look at your prediction for TPR.	Inform introduce variable	Discourse Marker = <i>Now</i>
	Can you tell me how it is controlled?	Elicit initial topic	
2. St:	Parasympathetics		Answer Category = Near Miss
3. Tu:	Correct, TPR is neurally controlled.		Acknowledgment = <i>Correct</i>
	And the reflex hasn't started to operate yet.	Inform middle topic	Discourse Marker = <i>And</i>
	So what is the value of TPR?	Elicit final topic	Discourse Marker = <i>So</i>
4. St:	Unchanged		Answer Category = Correct
5. Tu:	Great!		Acknowledgment = <i>Great</i>
	What other variables are neurally controlled?	Introduce next variables.	

Figure 1. Annotated Tutorial Dialogue for Correcting One Variable.

# Are Structural Principles Useful for Automatic Disambiguation ?

**Alexandra Kinyon** (Alexandra.Kinyon@linguist.jussieu.fr)  
TALANA-Université Paris 7 / CIS Dpt University of Pennsylvania  
Case 7003 2 Pl Jussieu 75005 Paris. France

## Abstract

In this paper we discuss how structural Preferences can be expressed within an LTAG framework on dependency like structures. We argue that the use of psycholinguistically motivated criteria is useful for building practical parse-ranking applications.

## Introduction

On the one hand computational linguists aim at parsing real texts : it is a difficult task, essentially because of spurious ambiguity. The goal is then to find a single preferred overall analysis for each sentence, either by resorting to general principles or to statistical methods, most of the time by focussing on the efficiency of the technique used, rather than on its theoretical relevance. So most of these disambiguation techniques do not take into account theoretical (i.e. cognitive) relevance, especially the incremental nature of human sentence processing.

On the other hand, psycholinguists aim at modeling the very early preferences which people employ in ambiguity resolution during an incremental parse of a sentence, without being concerned with the development of wide-coverage parsing systems or the integration of their principles in wide-coverage grammars. Importantly, these early decisions that people make may or may not eventually be compatible with the overall analysis of the sentence.

In this paper we argue that the two approaches are not antagonistic : wide-coverage disambiguation systems can integrate psycholinguistically motivated principles and yet be efficient. We present structural preferences which are expressed on dependency structures instead of constituent trees, within the framework of Lexicalized Tree Adjoining Grammars (LTAGs). In the first part of this paper, we briefly introduce the LTAG formalism. Then we present the preference principles used, and show that they work well in practice on large data. In a third part, we show why "pure" lexicalist approaches seem insufficient. In a fourth part, we discuss the interaction between our preference principles.

## 1. Brief overview of LTAGs

A LTAG consists of a finite set of **elementary trees** of finite depth. Each elementary tree must "anchor" one or more lexical item(s). The principal anchor is called "head", other anchors are called "co-heads". All leaves in elementary trees are either "anchor", "foot node" (noted \*) or

"substitution node" (noted  $\downarrow$ ). These trees are of 2 types : **auxiliary** or **initial**<sup>1</sup>. An auxiliary tree has exactly one distinguished leaf, called "foot node" and marked \*. Trees that are not auxiliary are initial. Elementary trees combine with 2 operations : **substitution** and **adjunction**. Substitution is compulsory and is used essentially for arguments (subject, verb and noun complements). It consists in replacing in a tree (elementary or not) a node marked for substitution with an initial tree that has a root of same category. Adjunction is optional (although it can be forbidden or made compulsory using specific constraints) and deals essentially with determiners, modifiers, auxiliaries, modals, raising verbs (e.g. seem) and sentential complements (e.g. object sentential complements). It consists in inserting in a tree in place of a node X an auxiliary tree with a root of same category . The descendants of X then become the descendants of the foot node of the auxiliary tree. Contrary to context-free rewriting rules, the history of derivation must be made explicit since the same derived tree can be obtained using different derivations. This is why parsing LTAGs yields a **derivation tree**, from which a **derived tree** (i.e. constituent tree) can be obtained. Figure 1 shows the elementary trees anchored when parsing "Yesterday John kicked the bucket"<sup>2</sup>, as well as the derivation trees obtained both for the "literal interpretation" and for the "idiomatic interpretations" of the sentence. It also shows that both derivation trees yield the same derived tree<sup>3</sup>. The derivation tree is close to a dependency structure (cf Candito & Kahane 98).

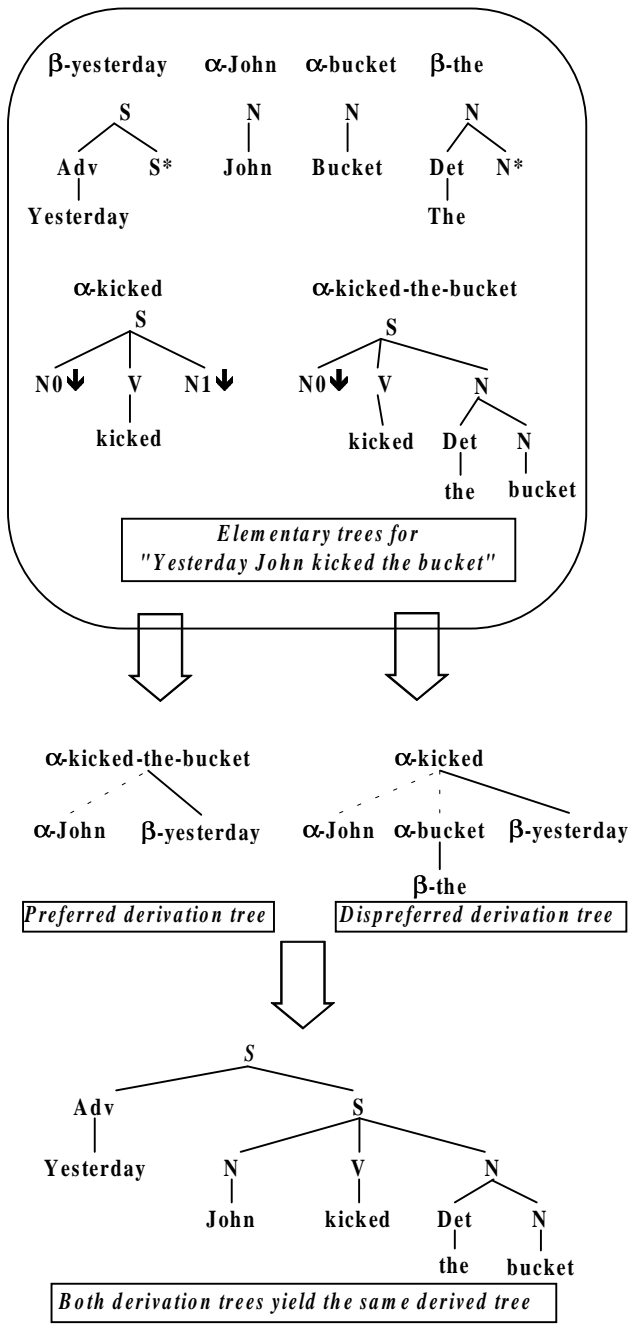
Moreover, linguistic constraints on the well-formedness of elementary trees have been formulated (Abeillé 91) (Frank 92) :

- Predicate Argument Cooccurrence Principle : there must be a leaf node for each realized argument of the head of an elementary tree.

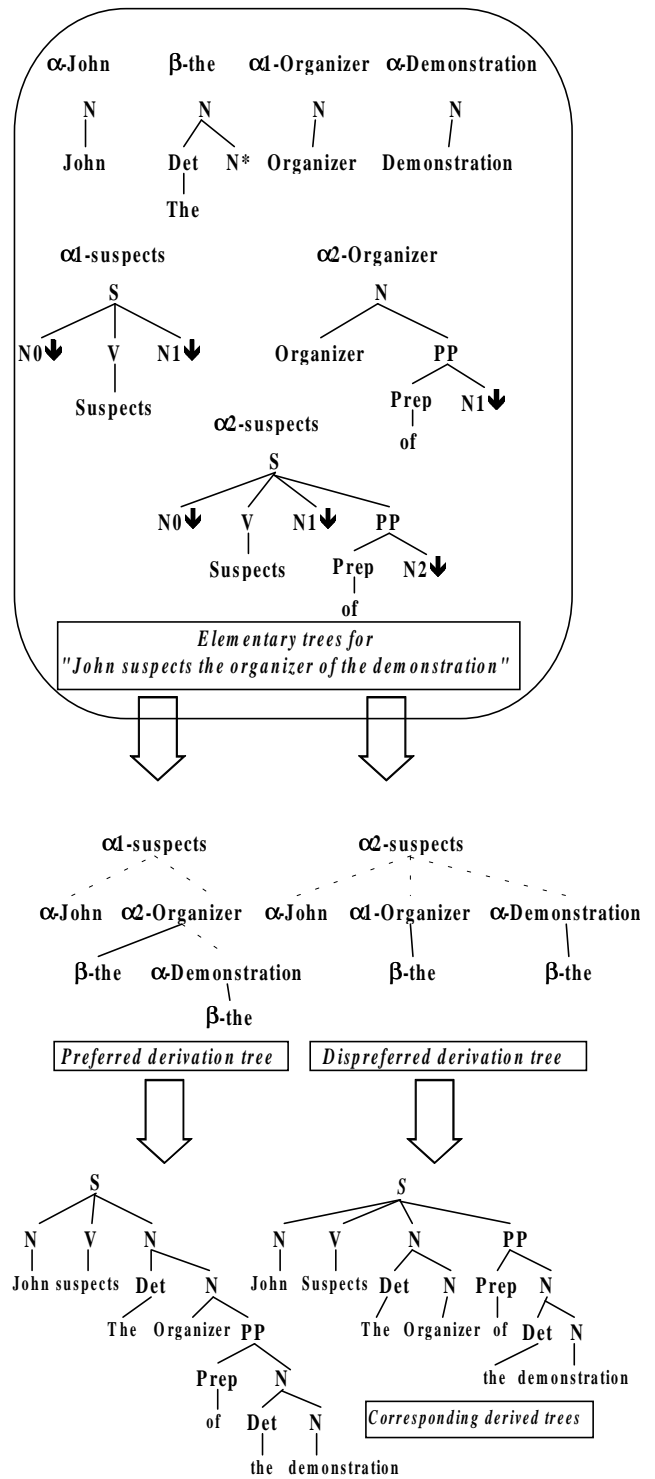
<sup>1</sup> Traditionally initial trees are called  $\alpha$ , and auxiliary trees  $\beta$

<sup>2</sup> All our examples follow linguistic analyses presented in [Abeillé 91]. Thus we use no VP node and no Wh nor NP traces. But this has no incidence on the application of our preference principles.

<sup>3</sup> Dotted lines in derivation trees indicate a substitution, plain lines an adjunction. The number at each node represents the address at which the operation took place, following Gorn convention.



**FIGURE 1 : Illustration of LTAG and of Principle 1**



**FIGURE 2 : Illustration of Principle 2**

- Semantic consistency : No elementary tree is semantically void
- Semantic minimality : an elementary tree corresponds at most to one semantic unit

## 2. Three preference principles expressed on derivation trees

A vast literature, going back as early as (Kimball 73), addresses structural parsing preferences. Older principles, such as right association (RA) and minimal attachment (MA) have been criticized : Among other things, the interaction between these principles is unclear. These principles lack provision for integration with semantics and/or pragmatics (Schubert (84)), do not clearly establish the distinction between arguments and modifiers<sup>4</sup> (Ferreira & Clifton (86)) and are English-biased : evidence against RA has been found for Spanish (Cuetos & Mitchell (88)) and Dutch (Brysbaert & Mitchell (96)). Newer structural principles, on the other hand, such as "Attach anyway" (Fodor & Inoue 98), are not integrated nor implemented into wide-coverage grammars.

So, to account for widely accepted preference principles, which are difficult to formulate in terms of constituents trees (idiomatic interpretation of a sentence favored over its literal interpretation (Abeille 95) (Gibbs 85) (Gibbs & Nayak 89)., arguments favored over adjuncts (Abney 89), (Britt & al. 92) and attachment to closest potential governor), (Kinyon 99a) has formulated the three following principle on dependency-like structures within the LTAG framework :

- 1-Prefer the derivation tree with the fewer number of nodes
- 2-Prefer to attach an  $\alpha$ -tree low in a derivation tree
- 3-Prefer the derivation tree with the fewer number of  $\beta$ -tree nodes<sup>5</sup>

A discussion on the linguistic adequacy of these principles, as well as on why LTAGs are better than other lexicalized formalisms such as LFG to formulate these principles can be found in Kinyon (99b).

Principle 1 accounts for the preference we have for the idiomatic interpretation of a sentence. In LTAGs, all the set elements of the expression are present in a single elementary tree. We have shown in Figure 1 the derivation trees obtained when parsing "Yesterday John kicked the bucket". The derivation tree for the idiomatic interpretation, which is preferred, has fewer nodes than the derivation tree for the literal interpretation..

<sup>4</sup> It is argued that MA and RA do distinguish arguments from modifiers, since arguments will yield a constituent tree with fewer nodes, but this relies very heavily on the underlying syntactic framework : it may be true for an X-bar theory, not necessarily for a more "surfastic " theory of syntax.

<sup>5</sup> This principle was initially presented in (Srinivas & al 95), formulated as "prefer to substitute rather than to adjoin".

Principle 2 captures the preference for an argument to attach to its closest potential governor. So in (a1), "of the demonstration" is preferably attached to "organizer" rather than to "suspect". Similarly, in (a2), "To whom" attaches to "say" rather than to "gives". Figure 2 shows how principle 2 yields the preferred derivation tree for sentence.(a1).

- (a1) John suspects the organizer of the demonstration
- (a2) To whom does Mary say that John gives flowers

Finally, principle 3 accounts for the preference of arguments over adjuncts. So it will allow to retrieve the right attachment in (b1), where "le matin" (the morning) is argument of "regarde" (watches) rather than modifier. It also allows to retrieve the correct attachment in (b2) where "to be honest" is argument of prefer, rather than sentence modifier.

- (b1) Jean regarde le matin (John observes the morning / John watches in the morning)
- (b2) John prefers his daughter to be honest

These principles are easy to implement, so they have yielded practical results<sup>6</sup> : A parse-ranker has been implemented for French within the FTAG project (cf Abeille & al 99,00), using a semi-automatically generated wide coverage grammar of 5000 elementary trees (Candito 96). This parse ranker, tested on 1000 TSNLP sentences, allows to go down from 2.85 derivations trees / sentence to 1.4 derivation trees / sentence without degrading the quality of parsing (i.e. without discarding "correct" parse trees). These results hint that the three principles are well-motivated from a cognitive point of view. This parse ranker is currently being ported to English and tested on the WSJ.

It is important to note that the distinction between arguments and modifiers can be easily expressed within LTAGs, because in derivation trees elementary trees for arguments are essentially initial ( $\alpha$ ), while elementary trees for modifiers are auxiliary ( $\beta$ ).

It is also important to note that (contrary to RA) these structural preferences are language independent, again because they are formulated on dependency-like structures and not on constituent structures : we have just seen that they work well for French, although French is argued to be an "early closure" language (Zagar & al 97).

## 3. Antagonism with lexicalist approaches ?

It has been shown that humans do exhibit frequency effects in language comprehension (Truwell 96), but this does not mean that structural principles are unsound and especially it does not demonstrate that disambiguation systems should resort only to "pure" lexicalist approaches :

One argument against the structural principles presented in 3 would be to say that these structural principles do not

<sup>6</sup> We do not claim, however, that these principles have yielded better results in automatic disambiguation than statistical parsers which integrate lexical information (e.g. Collins 96). Clearly though, our technique is easier to put to use, esp. for languages for which no training data is available.

exist (i.e. are not observable once frequency effects are taken into account). We disagree for the following reasons :

If the use of such principles was just a mere approximation, it would make it hard to explain that the empirical results are so good. Pure lexicalist approaches have not yielded such results to our knowledge on large real-world data (very little data about lexical preferences are available on a large scale esp. for languages other than English).

Also, lexicalist approaches do not allow to explain how two preferred subcategorization frames interact. For example, if "suspect N of N" and "organizer of N" are two preferred realization frames for respectively "suspect" and "organizer", one still needs to account for the fact that "demonstration" will be attached to "organizer" rather than to "suspect" in "John suspects the organizer of the demonstration"<sup>7</sup>. With the same type of reasoning, although "put N in N" is a common realization frame for arguments of "put", the sentence (c) nonetheless seems incomplete. This can also not be accounted for with a pure lexicalist approach

(c) I've put the book that you were reading in the library

Moreover, lexicalist approaches also do not necessarily account for unknown words<sup>8</sup> : in a sentence like (d) "at the man" will most likely be preferred as argument of "plups" rather than modifier, although we know nothing about the preferred subcategorization of "plups". So when considering language acquisition, unknown words are still processed, although no data is available regarding the preference of realization of their arguments. Structural preferences thus appear as a much more economical way to acquire new words : here "at the man" is argument of "plups" so "plups" subcategorizes a PP introduced by "at", whereas if one had to rely on frequency effect, it would take much longer to encounter "plups" many times before formulating a hypothesis about its subcategorization, and verifying it. Also, resorting to very few structural preferences for disambiguation seems much more economical and practical than storing huge quantities of frequency information about the lexicon, especially since contrary to structural preferences, information on the lexicon has to change from language to language.

(d) He plups at the man

Finally, to oppose lexicalist approaches and support the structural principles presented in 3, (Kinyon 00) formulated the following hypothesis :

Regardless of which realization of arguments a verb favors, if it can subcategorize a PP introduced by a given Preposition P, then in practice when the verb and a PP introduced by P appear in the same sentence, the PP is either an argument of the verb, or in a position where it can not be argument (i.e. argument of a closer potential governor, or located

in another clause such as inside a relative, or modifier only if the verb is already saturated).

This hypothesis was validated on LeMonde, a one million words annotated and shallow-parsed corpus for French (Clément & Kinyon 00, Abeillé & al 00b). The 100 most frequent verbs were extracted. 56 of these verbs could subcategorize PPs introduced by one or several prepositions, for a total of 71 subcategorization frames. Then, for each of these subcategorization frames, all the sentences where Verb and Prep co-occur were extracted and examined manually. The main findings are the following :

1- Cases of possible ambiguous attachment remain (13.86 % of the sentences examined)

2- 39% of these ambiguous cases are solved when attaching the PP to the closest potential governor. Moreover, the attachment is deemed correct in all cases.

3- The probability for a verb to realize as an argument a PP introduced by a given Preposition P does not help disambiguation and does not predict the proportion of ambiguous attachments encountered when examining sentences where Verb and P co-occur.

4- Rather, the preposition itself is important : "à" yields much more ambiguity than other prepositions such as "avec" or "pour" because it often introduces a temporal or locational expression (e.g. "à l'assemblée nationale" (*in parliament*) / "à 3 heures" (*at 3 o'clock*)). In fact, 46% of the ambiguous cases remaining after applying structural principles 2 and 3 are solved by resorting to very simple semantic information : à + location nouns , à + time nouns are overwhelmingly adjuncts and not arguments.

Therefore, only 4.6 % of ambiguous attachments remain (mainly set phrases such as "lancer un appel au calme"), which could be disambiguated by refining semantic disambiguation. Thus the hypothesis is validated, which indicates that the use of structural principles + basic semantic information allows very efficient disambiguation and again, in a more economical manner than lexical approaches.

As discussed in (Kinyon 99b) though, some lexical preferences seem useful, but formulated not at the level of lexical items, but rather at the level of parts of speech. So for instance, grammatical categories are preferred over lexical categories. So in (e1) clitic will be preferred over noun for "elle", in (e2) "être" (be) will be an auxiliary rather than a lexical verb, and in (e3) "deux" will be a determiner rather than a noun. General lexical preferences of this type have been incorporated in the parse-ranker discussed in 3. Expressing lexical preferences in such general terms is also economical and allows to eliminate some cases of spurious ambiguity.

(e1) Elle court (She runs / It is her who runs)

(e2) Elle est venue (She has arrived / She is an arrival)

(e3) Je vois deux hommes (I see two men)

<sup>7</sup> Whereas claiming that arguments prefer to attach to their closest potential governors solves this problem.

<sup>8</sup> One may say that more general information is used when encountering unknown lexical items, but this general idea is not implementable as such.

#### 4. Conflicts between structural principles

One of the main argument against "traditional" structural principles is that the interaction between them is unclear. It has been said for example that in case of conflict, minimal attachment prevails over right association in a sentence such as "He repaints the wall with cracks" thus allowing to account for the garden path effect. Of course, this suffers numerous counter-examples.

With the structural principles presented in 3 and expressed on dependency like structures, it is striking that zero conflicts were encountered, both on the 1000 sentences for French, and on 3000 sentences from the wall street journal for English.

This strongly suggests that these principles are relevant from a cognitive point of view.

#### Conclusion

We have presented three parsing preference principles expressed on dependency like structures, and shown that these language-independent principles are both psycholinguistically relevant and useful to disambiguate real-word data on a large scale (which has led to the development of a parser-ranker). We also came to the conclusion that an efficient disambiguation scheme involving these structural preferences as well as limited semantic information and "simplified" lexical principles (i.e. expressed in terms of parts of speech) was more economical than acquiring large amounts of lexical data, thus being more appealing both from a practical and from a cognitive point of view. In fact, these structural preferences are a first step towards a psycholinguistically relevant processing model for LTAGs, which allows among other things to predict garden-path phenomena (cf Kinyon 99c, Kinyon 00b).

#### REFERENCES

- Abeillé A. (1991) Une grammaire lexicalisée d'arbres adjoints pour le français : application à l'analyse automatique. These de doctorat. Université Paris 7.
- Abeillé A. (1995) The flexibility of French idioms. In Idioms LEA. Schenk & al. (eds).
- Abeillé A. Candito M.H. Kinyon A. (1999) FTAG : current status and parsing scheme. Proceedings Vextal'99. Venice.
- Abeillé A. Clément L. Kinyon A. (2000) Building a tree-bank for French. Proc. LREC'2000. Athens
- Abeillé A. Candito M.H. Kinyon A. (2000) The current status of FTAG. Proc. TAG+5. Paris.
- Abney S. (1989) A computational model of human parsing. Journal of psycholinguistic Research, 18, pp. 129-144.
- Britt M, Perfetti C., Garrod S, Rayner K. (1992) Parsing and discourse : Context effects and their limits. Journal of memory and language, 31, 293-314.
- Brysbaert M., Mitchell D.C. (1996) *Modifier Attachment in sentence parsing : Evidence from Dutch*. Quarterly journal of experimental psychology, 49a, 664-695.
- Candito M.H. (1996) A principle based hierarchical representation of LTAG. Proc. 15<sup>th</sup> COLING. Copenhagen.
- Candito M.H., Kahane S. (1998). Can the TAG derivation tree represent a semantic graph ? an answer in the light of MTT. Proc. TAG+5. Philadelphia.
- Clément L. Kinyon A (2000). Chunking, marking and searching a morpho-syntactically annotated corpus for French. Proc. ACIDCA'2000. Monastir.
- Collins M. (1996) Three generative, Lexicalised Models for statistical parsing. Proc. ACL'97. Madrid.
- Cuetos F., Mitchell D.C. (1988) *Cross linguistic differences in parsing : restrictions on the use of the Late Closure strategy in Spanish*. Cognition, 30,73-105.
- Ferreira F. Clifton C. (1986) *The independence of syntactic processing*. Journal of Memory and Language, 25,348-368.
- Fodor J.D. Inoue A. (1998). Attach Anyway. In Reanalysis in Sentence Processing. Fodor & Ferreira (eds). Kluwer academic publishers.
- Frank R. (1992) Syntactic Locality and Tree Adjoining Grammar : Grammatical Acquisition and Processing Perspectives. PhD dissertation. Univ. of Pennsylvania.
- Gibbs R. (1985) On the process of understanding idioms. Journal of psycholinguistic research, 14, pp. 465-477.
- Gibbs R., Nayak (1989) Psycholinguistic studies on the syntactic behaviour of idioms. Cognitive Psychology, 21, pp. 100-138.
- Kimball J. (1973) Seven principles of surface structure parsing in natural language. Cognition 2.
- Kinyon A. (1999a) : Parsing preferences with Lexicalized Tree Adjoining Grammars : exploiting the derivation tree. Proc. ACL'99
- Kinyon A. (1999b) : Hiérarchisation d'analyses basée sur des informations dépendancielles dans le cadre des LTAGs. Proceedings TALN'99.
- Kinyon A. (1999c). Some remarks about the psycholinguistic relevance of LTAGs. CLIN'99. Utrecht
- Kinyon A. (2000a). Structural preferences vs Lexical preferences : some data on French verbs subcategorizing a PP. Poster presented at Cuny'2000. La Jolla. California.
- Kinyon A. (2000b). Hypertags. Proc. COLING'2000. Sarrebrücken.
- Schubert L. (1984). *On parsing preferences*. COLING'84, Stanford. 247-250.
- Srinivas B., Doran C., Kulick S. (1995) : Heuristics and Parse Ranking. Proc IWPT'95. Prag. Czech Republic.

Trueswell, J.C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, 35, 566-585.

Zagar D., Pynte J., Rativeau S. (1997) Evidence for early closure attachment on first pass reading times in French. *Quarterly journal of experimental psychology*, 50(A), 421-438.



# Dynamic Extension of Episode Representation in Analogy-Making in AMBR

**Boicho Kokinov** (bkokinov@nbu.bg)

Central and East European Center for Cognitive Science  
Department of Cognitive Science and Psychology  
New Bulgarian University  
21 Montevideo St, Sofia 1635, Bulgaria

**Alexander Petrov** (apetrov+@andrew.cmu.edu)

Central and East European Center for Cognitive Science  
Department of Cognitive Science and Psychology  
New Bulgarian University  
21 Montevideo St, Sofia 1635, Bulgaria

## Abstract

Models that rely exclusively on static representations cannot account fully for the flexibility of human analogy-making. More sophisticated models should provide mechanisms for dynamic extension, elaboration, and re-representation of episodes. One such mechanism—the *instantiation* mechanism—is described. It uses the target problem as a template for extending the source and vice versa. These extensions are driven and constrained by semantic knowledge about general regularities in the domain. The instantiation mechanism has been developed within a model of analogy-making called AMBR. It relies on AMBR's support for parallel, decentralized, and interactive computation. The instantiation mechanism runs in parallel with the mechanisms for analog access and mapping. Thus, these latter mechanisms guide the instantiation mechanism as to which facts in the large semantic memory are relevant to the specifics of the current situation.

## Re-Representation in Problem-Solving

A substantial body of evidence suggests that people can change their mental representations dynamically during various cognitive tasks. Yet, despite the widespread agreement that human representations are dynamic and flexible, the mechanisms behind these re-representation abilities are not well explored and understood. This paper suggests some mechanisms that can serve that purpose.

In the context of problem solving, there are at least two complementary aspects of re-representation: re-representation of the target problem and re-representation of prior knowledge. These two aspects correspond to Jean Piaget's complementary and related processes of assimilation and accommodation. Re-representation of the target is a process of assimilation because the new information is transformed to comply with the existing knowledge. Conversely, re-representation of the existing knowledge in the face of new experience is a process of accommodation.

*Re-representation of the target problem* has received more attention although it is still not fully understood. Gestalt psychologists (Maier, 1931; Dunker, 1945) demonstrated the importance of dynamic changes of

the target representation for successful problem solving. However, the mechanisms behind this remained unclear and somehow mysterious. Unfortunately, contemporary cognitive science has not been particularly successful to fill this gap, although some progress has been made. Douglas Hofstadter and his group have worked for many years on the integration of analogy-making and what they call high-level perception (Hofstadter, 1995; French, 1995; Mitchell, 1993; Chalmers, French & Hofstadter, 1992). They have proposed a number of mechanisms that work together to build several alternative representations simultaneously, settle gradually on one of them, and radically restructure the representation and settle on an alternative one if necessary. Lange & Wharton (1992) have worked on a similar problem—integrating language comprehension with analogical reminding. They suggested a mechanism for parallel processing of several possible interpretations of an ambiguous phrase. This mechanism would allow for re-interpretation of the phrase if necessary.

The issue of *re-representation of existing knowledge* during the problem solving process has been systematically ignored in the problem solving literature. Unfortunately, most classical models of problem solving take the existing knowledge as constant<sup>1</sup>. Long-term memory structures only have to be retrieved and applied (Ernst & Newell, 1969; Newell, 1990; Anderson, 1983; Anderson & Lebiere, 1998). The same is true for models of analogy-making: they retrieve a ready-made representation of an old episode when looking for a base for analogy (Gentner, 1989; Thagard, Holyoak, Nelson & Gochfeld, 1990; Kokinov, 1994a; Forbus, Gentner & Law, 1995; Hummel & Holyoak, 1997).

The idea that there are ready-made representations of episodes formed during the encoding stage has been challenged by researchers of human memory for quite some time. There is much evidence for the constructive nature of the “retrieved memory traces”. Thus Loftus (1977, 1979) and Neisser and Harsch (1992) demonstrate that people can have vivid “memories” of nonexistent episodes which are clearly constructed during the “retrieval” process. Bartlett

---

<sup>1</sup> We do not discuss learning here—it involves gradual long-term knowledge change rather than short-term accommodation.

(1932) has shown that episode representations are distorted and enriched with information inherited from the schema for the typical event.

The purpose of this paper is to show how episodic representations can be dynamically extended with information derived from general semantic knowledge. AMBR is among the few models of analogy-making which use both semantic and episodic knowledge (Kokinov, 1994a). However, the version reported in (Kokinov, 1994a) used them for different purposes. Episodic memory was a repository of ready-made representations of episodes that might be used as bases for analogy-making. Semantic memory served only to establish semantic similarity between the relations, attributes, and objects participating in the episodes. In the new version of AMBR described in this paper, the boundaries between semantic and episodic memory are blurred and general statements from semantic memory are instantiated to complement and extend the representations of episodes.

## **Main Assumptions of the AMBR Approach to Episode Re-representation**

### **Flexible Dynamic Representations**

The first assumption is that episodes and concepts are not represented by fixed and static complex memory structures such as schemas, lists of propositions or lists of rules. Rather, there are fuzzy and overlapping *coalitions* of simple memory elements. The key distinction is that complex memory structures are retrieved in an all-or-none fashion, while flexible dynamic representations are retrieved and/or constructed element by element. Each element has an activation level associated with it. Depending on the specific pattern of activation over the coalition, various partial representations of the same episode or concept can be retrieved. Some elements of the coalition might be strongly connected with each other and thus tend to be retrieved together, while others might be retrieved only rarely and under specific retrieval conditions. There is also a possibility for *blends* to emerge. This happens when elements belonging to more than one coalition become active together. All this flexibility is important for making re-representation possible.

### **Flexible Dynamic Computation**

The second assumption is that computations are also flexible, parallel, and interactive. There are multiple processes running in parallel and interacting in complex ways that are not specified in advance. Thus, re-representation results from the interplay of many processes including: (i) retrieval of past episodes, (ii) retrieval of generic knowledge, (iii) instantiation of the generic rules or facts, (iv) attempt to build the representation of the past episode in a form that makes it alignable with the representation of the problem at hand, and (v) attempt to build the representation of the target problem in a form that makes it alignable with the representation of the past episode. As stated earlier, it is believed that all these processes should run in parallel and influence each other's

work. In addition, each of the processes described above is quite complex in itself and in turn has to be considered as the result of the interplay of many simpler and more local processes. These requirements lead to a view that complex computation is an emergent interactionist phenomenon rather than pre-specified sequence of algorithmic steps.

### **Integrated Semantic and Episodic Memory**

Semantic memory contains information about concepts and statements about classes of instances. Episodic memory contains information about instances, episodes, and statements about instances. There has been a long-lasting discussion for and against the distinction between these two memories (see, for example, Anderson & Ross, 1980, Herrmann & Harwood, 1980). The third assumption of AMBR is that semantic and episodic memories are integrated so as to allow for coordinated search in both memories. Whenever a cue is provided, both semantic and episodic memory elements can potentially be retrieved. Thus, when a past episode is *recalled* from memory, both specific and general knowledge is used in the recollection process. In this way semantic knowledge can extend the episodic knowledge.

### **Integration of Memory and Reasoning**

The process of re-representation requires that the process of memory access, on one hand, and the processes of mapping and instantiation, on the other, run in parallel and interact with each other. Thus the fourth assumption is that memory and reasoning are highly integrated.

## **AMBR: An Analogy-Making Model Based on the Cognitive Architecture DUAL**

An analogy-making model with re-representation capabilities needs the support of a full cognitive architecture that implements all the assumptions above. The cognitive architecture DUAL is specifically designed to support this decentralized and interactive style of computation (Kokinov, 1994b, 1994c, 1997). The AMBR model of analogy-making (Kokinov, 1994a) is based on this architecture. This paper describes the re-representation extensions that have been added to the model after the original publication. Before explaining how the assumptions are implemented and how they contribute to re-representation, a brief and more general description of DUAL and AMBR is needed.

DUAL is a cognitive architecture based on the *society of mind* idea (Minsky, 1986; see also Hofstadter, 1995). Every DUAL-based system consists of many micro-agents, each of which is quite simple. The micro-agents do not have goals and do not plan their activities; they are simple representation and computation devices. They can establish new links with other agents and some of them can construct new agents. DUAL-agents form coalitions that collectively represent an episode or a generalized concept, or dynamically form coalitions that collectively produce an emergent computation process. Each agent can participate in many coalitions to a various extent depending on the weights of the links connecting that agent to other agents in the coalition.

Knowledge representation in DUAL is highly decentralized. Each episode, concept, general theory, etc. is represented by a coalition of many agents, each of which represents just a small piece of knowledge. Thus a simple episode such as boiling water in the kitchen would be represented by a quite big coalition of agents: an agent for every *concept* related to the situation such as “water”, “kitchen”, “boiling”, “plate”, “pot”, “on”, “in”, “hot”, “cold”, “cause”; an agent for every *instance* of these concepts involved in the particular situation, i.e. “water-1”, “kitchen-3”, “boiling-2”, “plate-3”, “pot-3”; as well as for every single statement such as “on-1(pot-3,plate-3)”, “in-1(water-1,pot-3)”, “hot-1(plate-3)”, “red-1(pot-3)”, etc. However, it should not necessarily be the case that all elements of a coalition become members of the working memory (WM) at certain moment. On the contrary— typically only part of the coalition is activated. Thus each episode is almost always only partially available. Moreover, different subsets of the coalition are active in different contexts. The long-term memory (LTM) of DUAL is the population of all permanent agents, active or inactive. The working memory is simply the active part of LTM plus some newly created temporary agents.

Each agent is a DUAListic computational and representational device: it has a symbolic and a connectionist part. While the symbolic part represents a piece of knowledge (as described above), the connectionist part represents the relevance of this piece of knowledge to the current context. The relevance is represented by the graded activation level computed by the connectionist processor associated with the agent. All the inferences based on the knowledge represented by the agent are computed by the symbolic processor associated with the same agent. These computations are also based only on local interactions with neighboring agents. If necessary, the agents are able to establish new temporary links (and interactions) with other agents. The speed of symbolic processing of a given agent depends on the activation level. In this way the computations are faster if the corresponding knowledge structures are considered relevant to the context and slower or even impossible if they are less relevant or irrelevant.

AMBR is a model of analogy-making based on DUAL, which integrates memory and reasoning. The mechanisms for memory access, mapping, inference, re-representation, etc. are based on emergent computations implemented over a large set of DUAL agents. Memory access is based mainly on the spreading activation mechanism of the connectionist aspect of DUAL. Mapping is based on a number of mechanisms such as marker passing for establishing semantic correspondence, temporary-agent constructors for establishing hypotheses about possible correspondences, link constructors for establishing positive or negative links among hypotheses and existing long-term agents based on structure correspondence, etc. All these mechanisms are running in parallel and influence each other, thus giving rise to various interaction effects.

## AMBR Mechanisms for Dynamic Extension of Episode Representation

Episode representation is dynamically extended in AMBR by the interplay of three processes running in parallel: (i) gradual and partial retrieval of episodic and semantic memory elements, (ii) gradual and partial mapping the retrieved episode and semantic elements onto the target elements, and (iii) gradual and partial instantiation of general statements from semantic memory.

*The gradual retrieval process* is based on the spreading activation over the links between the neighboring agents. When the activation level of certain agent exceeds a given threshold, the agent becomes part of the working memory<sup>2</sup>. It is possible that only part of the coalition passes the threshold, which means that it is possible that only part of the encoded episode elements are retrieved. Thus different representations of the past episode are “constructed” or “retrieved” in different contexts. This differs from other analogy models. Most of them use centralized episode representations (Forbus, Gentner, & Law, 1995, Thagard et al., 1990). Even in LISA (Hummel & Holyoak, 1997) where the episodes are represented in a decentralized way and where the retrieval process is a gradual one, there is a final decision about which episode has won the competition. This decision is done centrally and all elements of the winner are switched from “dormant” into “active” state. Therefore, no partial retrieval of episodes is possible. In AMBR there is even no in-principle possibility to do this form of forced retrieval of whole episodes because the system does not keep any central registry of rosters enumerating the affiliation of elements to episodes.

Since there are tight links between the elements of semantic and episodic memory, activated agents do not necessarily represent elements of an episode. They can also represent pieces of semantic knowledge. Thus, contrary to other models, the retrieval process in AMBR brings both elements of episodic and semantic memory into the WM. Since semantic knowledge is also represented in a decentralized manner, it has the same degree of flexibility. Two scenarios are worth mentioning. The spreading activation mechanism can retrieve (i) a coalition representing schematic knowledge about a typical situation (e.g. “boiling water in the kitchen”) or (ii) single generic statements (such as “a pot is made of metal”). Because the process of instantiation of a schema is much more traditional and well studied, we will focus on the instantiation of single generic statements.

*The gradual mapping process* starts as soon as the first elements from episodic or semantic memory pass the working-memory threshold. An attempt is made to map them onto elements from the target description. An external observer monitoring the behavior of the system as a whole can ascribe different labels to this process depending on the particular kind of prior knowledge that the system happens to use in each particular case. If a past episode is retrieved and mapped to the target, this could be labelled “analogy”. If a general schema is retrieved and used as a source for the

---

<sup>2</sup> After entering the WM, the graded activation continues to play an important role since the speed of symbolic processing performed by the agent depends on its activation level.

mapping, the “analogy-making” mechanism produces an inference that we might prefer to call deductive<sup>3</sup>. The prevailing number of cases will be mixed, however: both episodic and semantic elements will be mapped. These are the cases considered in more detail in this paper. The process of analogy-making is an emergent process. What actually happens in the system at the micro-level is that individual elements of the descriptions try to find their “mates”, i.e. to form correspondence hypotheses between target elements and retrieved elements regardless of whether these elements are originating from episodic or semantic memory. At the same time all the agents participating in this process establish temporary links among themselves in order to cooperate in finding a structurally consistent mapping (Gentner, 1983).

The details of how retrieval and mapping are performed in a decentralized and emergent way will not be presented here because of lack of space. Interested readers can find such descriptions elsewhere (Kokinov, 1994a, Kokinov, Nikolov, & Petrov, 1996). The focus here is on the processing that takes place after a mapping between elements of semantic memory and target elements is established. For example, when an isolated generic statement from semantic memory, such as “made-of-1(teapot,metal)” or “is-hot-1(plate)”, is retrieved it can be mapped onto elements of the target description such as “made-of-2(vessel-1,wood-3)” or “is-burning-2(fire-1)”, respectively.

After the initial correspondence is established, which might be based on the semantic similarity between the predicates (established by the marker-passing mechanism), a generic hypothesis is formed (i.e. a new agent is created) which puts the target proposition—“made-of-2(vessel-1, wood-3)” —in correspondence with the general statement coming from semantic memory that teapots are typically made of metal: “made-of-1(teapot,metal)”. In case that the retrieved episode representation already contains a statement “made-of-3(teapot-1,metal-1)” then most probably it will win the competition and the generic hypothesis will be rejected. However, if such statement is not encoded in the long-term memory since the material of the teapot was not important at the time of experiencing the event, or it was encoded but for some reason it is now not retrieved in WM and therefore does not exist in the current representation of the episode, then it might happen that the generic hypothesis wins the competition or at least is strong enough to start an instantiation process.

**The instantiation process** builds up a new proposition where all the universally quantified variables will be replaced by specific instances-constants, e.g. “made-of-1(teapot,metal)” goes into “made-of-4(teapot-1,metal-prototype)”. How are the constants chosen? If there is a constant (object) of the same type in the retrieved episode, it should be used. In the example above, the episode representation involves such an instance—“teapot-1”. Then the new proposition will use it as an argument. If the episode contains no specific instance of that type then a new instance is constructed which has the properties of the

prototype of the corresponding class. In this example, there is no instance of metal in the episode representation and therefore a new instance is formed—“metal-prototype”. Thus the instantiation mechanism tries to reuse existing instances whenever possible. The DUAL marker-passing mechanism provides information about which instances of the concept “teapot”, if any, are active in the current representation of the episode and hence are available for instantiation.

Instantiation has been used in analogy-making models so far only for adding new objects and propositions to the target problem, i.e. for making analogical inferences (Holyoak & Thagard, 1989, Falkenhainer, Forbus, & Gentner, 1989). In AMBR it is used for extending episode representation and relies heavily on the semantic knowledge of the system.

In summary, the process of extending the representation of the episode emerges from the interaction of several processes that are themselves emergent: the retrieval process which continuously brings up new episodic and semantic memory elements into WM, the mapping process which continuously builds hypotheses about possible correspondences between the retrieved elements and elements from the target description, and the instantiation process which continuously constructs new specific propositions based on generic propositions retrieved from semantic memory.

Why is the continuous interplay between these three processes important? The interactions guide each of the processes and therefore make each of them more effective. They preclude the model from doing exhaustive search. The influence of retrieval on mapping and instantiation is obvious since nothing can be mapped or instantiated if it is not activated (retrieved). The role of mapping is unusual compared to other models of analogy-making. Since the retrieval process in AMBR is a piece-by-piece process that runs continuously and in parallel with the mapping, the latter can influence the former. It is always the case that the retrieved elements of the episode send out activation to the rest of the elements of the episode representation and thus constantly try to activate the whole coalition. However, if the coalition is not tight enough (which is the typical case) they would be able to retrieve only some of their coalition partners. Exactly which elements will be retrieved depends not only on the initial set of elements but also on their mapping status, i.e. which of them are mapped onto target elements and which are not. Elements that are mapped receive abundant activation from the target and therefore will play important role in any further retrieval. In this way the mapping influences the retrieval process.

The importance of this interaction between processes can be demonstrated by contrasting two runs of the system: one with parallelism and interaction and one without. Figure 1 presents such a comparison from a simulation experiment. In the “parallel condition” (thick lines in the figure) all processes are running in the way they have been described so far. In the “retrieval only condition” the mapping process has been intentionally switched off. The important result is that, although the target and the background knowledge were exactly the same in the two runs, two different episodes are retrieved—a more structurally similar one in the first case and a more superficially similar one in the second case.

---

<sup>3</sup> AMBR has been proposed as a unifying mechanism for deductive, inductive, and analogical reasoning (Kokinov, 1988, 1990, 1992, 1994a).

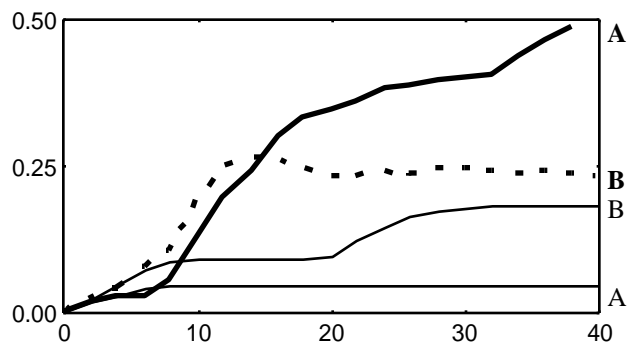


Figure 1. Retrieval indices for two episodes, A and B, in two different conditions as a function of time. The thick lines correspond to the parallel condition in which mapping influences retrieval, the thin lines show 'pure' retrieval.

The mapping influences the instantiation process as well. If there were no such influence, the model would have to build up unrealistically many instantiations—one or more for each generic proposition retrieved from semantic memory. However, the instantiation process is guided by the mapping process—only general propositions that are mapped onto target propositions will be instantiated. On the other hand, once an instantiation is built it supports the mapping and helps in further retrieval of memory elements.

### Conclusions

The mechanisms described above allow for dynamic re-representation of the episodes by: retrieving additional information from episodic memory based on the established mappings; by constructing new memory elements and integrating them into the episode representation based on instantiation of generic statements retrieved from semantic memory and mapped onto the target description; and by retrieving elements from other episodes thus producing a blending between episodes.

In this way AMBR makes the following predictions which can be tested experimentally. The first prediction is that the partial mapping established up to a point influences the further retrieval process. This prediction can be tested by analysing thinking-aloud protocols. Actually, such results have been obtained by Ross and Sofka (1986) as a side effect in a thinking-aloud study. They are summarized in (Ross, 1989) as follows: "... other work (Ross & Sofka, 1986) suggests the possibility that the retrieval may be greatly affected by the use. In particular, we found that subjects, whose task was to recall the details of an earlier example that the current test problem reminded them of, used the test problem not only as an initial reminder but throughout the recall. For instance, the test problem was used to probe for similar objects, and relations and to prompt recall of particular numbers from the earlier example. The retrieval of the earlier example appeared to be interleaved with its use because subjects were setting up correspondences between the earlier example and the test problem during the retrieval." The simulation data described here are obtained

absolutely independently and are based only on the theoretical assumptions of DUAL and AMBR and exhibit exactly the same pattern of interaction.

A second prediction is that people would instantiate generic knowledge in cases where there is missing information from the episode representation and where this information is needed for the mapping, i.e. there is a corresponding piece of information in the target which needs to be mapped onto something from the base. An experiment is currently being prepared to test this hypothesis. McKoon and Ratcliff (1981) demonstrated that people make inferences and extend episode representation during the encoding process, e.g. after listening to a sentence such as "Alice pounded in the nail until the board was safely secured." listeners would infer and encode that "Alice used a hammer." Our prediction is that they would further extend the representation during the recall process when they use that episode in order to map it to the target.

Finally, a third prediction is that people will tend to blend episodic information if the information needed for mapping is missing in the best retrieved episode, but is present in another episode that is also partially retrieved. Another experiment is under development to test this prediction.

### References

- Anderson, J. (1983). *The architecture of cognition*. Cambridge, MA, Harvard University Press.
- Anderson, J. & Lebiere, C. (1998) *The atomic components of thought*. Mahwah, NJ: Erlbaum
- Anderson, J. & Ross, B. (1980). Evidence Against a Semantic-Episodic Distinction. *Journal of Experimental Psychology: Human Learning and Memory*, 6 (5), 441-466.
- Bartlett, F. (1932). *Remembering*. Cambridge: Cambridge University Press.
- Charlmers, D., French, R., & Hofstadter, D. (1992). High-Level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology. *Journal of Experimental and Theoretical AI*, 4 (3), 185-211.
- Dunker, K. (1945) *On problem solving*. In: Psychological Monographs, vol. 58 (whole No 270).
- Ernst, G. & Newell, A. (1969). *GPS: A case study in generality and problem solving*. NY: Academic Press.
- Falkenhainer B., Forbus K., & Gentner D. (1989). The structure mapping engine: Algorithm and examples. *Artificial Intelligence*, 41 (1), 1-63.
- Forbus, K., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19 (2), 141-205.
- French, R. (1995). *The subtlety of sameness: A theory and computer model of analogy-making*. Cambridge, MA: MIT Press
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7 (2), 155-170.
- Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.). *Similarity and analogical reasoning*. New York, NY: Cambridge University Press.

- Herrmann, D. & Harwood, J. (1980). More evidence for the existence of separate semantic and episodic stores in long-term memory. *Journal of Experimental Psychology: Human Learning and Memory*, 6(5), 467-478.
- Hofstadter, D. & the Fluid Analogies Research Group (1995). *Fluid concepts and creative analogies*. New York: Basic Books.
- Holyoak K. & Thagard P. (1989b). Analogical mapping by constraint satisfaction. *Cognitive Science*, 13, 295-355.
- Hummel, J. & Holyoak, K. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3), 427-466.
- Kokinov, B. (1994a). A hybrid model of reasoning by analogy. In K. Holyoak and J. Barnden (Ed). *Advances in connectionist and neural computation theory. Vol. 2: Analogical connections*. Norwood, NJ: Ablex
- Kokinov, B. (1994b). The DUAL cognitive architecture: A hybrid multi-agent approach. *Proceedings of the Eleventh European Conference on Artificial Intelligence*. Wiley.
- Kokinov, B. (1994c). The context-sensitive cognitive architecture DUAL. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kokinov, B. (1997). Micro-level hybridization in the cognitive architecture DUAL. In R. Sun & F. Alexander (Eds). *Connectionist-symbolic integration: From unified to hybrid architectures*. Hillsdale, NJ: Erlbaum
- Kokinov, B., Nikolov, V., & Petrov, A. (1996). Dynamics of emergent computation in DUAL. In A. Ramsey (Ed.). *Artificial intelligence: Methodology, systems, applications*. Amsterdam: IOS Press
- Kolodner, J. (1983). Reconstructive memory: A computer model. *Cognitive Science*, 7, 281-328.
- Lange, T. & Wharton, C. (1992). Dynamic memories: Analysis of an integrated comprehension and episodic memory retrieval model. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum
- Loftus, E. (1977). Shifting human color memory. *Memory and Cognition*, 5, 696-699.
- Loftus, E. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.
- McKoon, G. & Ratcliff, R. (1981). The comprehension process and memory structures involved in instrumental inference. *Journal of Verbal Learning and Verbal Behavior*, 20, 671-682.
- Maier, N. (1931). Reasoning in humans II: The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology*, 12, 181-194
- Minsky, M. (1986). *The society of mind*. New York: Simon & Schuster.
- Mitchell, M. (1993). *Analogy-making as perception: A computer model*. Cambridge, MA: MIT Press
- Neisser, U. & Harsch, N. (1992). Phantom flashbulbs: False recollections of hearing the news about the Challenger. In E. Winograd & U. Neisser (Eds.) *Affect and Accuracy in Recall*. NY: Cambridge University Press.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Ross, B. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 456-468.
- Ross, B. & Sofka, M. (1986). Reminders: Noticing, remembering, and using specific knowledge of earlier problems. Unpublished manuscript.
- Thagard, P., Holyoak, K., Nelson, G., & Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence*, 46, 259-310.

# Controlled Exploration of Alternative Mechanisms in Cognitive Modeling

Rita Kovordányi (ritko@ida.liu.se)

Department of Computer and Information Science  
Linköpings Universitet, SE-581 83 Linköping, Sweden

## Abstract

Overt cognitive behavior arises through a complex interaction between internal, not directly observable, cognitive mechanisms. As there may be several ways of achieving the same overt behavior, it is intrinsically difficult to find the “correct” model. One way to proceed however is to uncover the causal dependencies between a particular configuration of cognitive mechanisms and simulated overt behavior. This can be achieved in controlled simulation experiments where every combination of potentially important cognitive mechanisms is systematically tried out. To illustrate this point, we briefly describe an application of the two-level factorial simulation design on a modeling project in mental imagery. We conclude by discussing the potential of the method as a tool for reliable incremental model development.

## Introduction

The general objective of modeling and simulation is often to correctly predict real-world system performance. In addition to this, cognitive modeling aims at discovering the true nature of cognition (Kieras, 1987; Anderson, 1993; Newell, 1990; Kosslyn, 1980, 1994; Kosslyn et al., 1979). Ideally, this would presuppose either that a cognitive model can be rejected as invalid with respect to empirical data, or that a cognitive model, or a particular cognitive mechanism, can be singled out as being valid within a given theoretical setting.

However, behavioral data often do not cover every necessary aspect of a cognitive phenomenon or are qualitative in nature, and may thus be consistent with a range of possible accounts. This open-endedness poses a severe problem in cognitive theory construction and model building, a problem which is commonly known as the identifiability problem: “The thorny issue of how we can know [that we have arrived at] the correct theory” (Anderson, 1993, p. 10).

Several ways of dealing with this problem have emerged during decades of modeling practice. First, the empirical basis for model construction can be broadened to increase the number of constraints and thereby pin down the gross structure of possible cognitive models. Within the space of possible models which is left, often ad hoc or heuristic search is employed to find a model which satisfies the full range of data (Kieras, 1985). In general, this method increases the probability that the model found is also “correct” in a broader sense.

Second, unified architectures of cognition are incrementally constructed in a team effort and evolve through years of development to accumulate a wide range of empirical data. These architectures outline the main processing subsystems

and the flow of processing in the cognitive system, and in this way support the development of specific, lower-level models (Rosenbloom et al., 1993; Anderson, 1993).

Overt cognitive behavior arises from a complex interaction between internal cognitive mechanisms. Even when model development is guided by assumptions about the overall cognitive architecture, it may be difficult to pin-point which of several possible mechanisms is responsible for a set of empirical observations. For example, should the empirically observed reaction time and error-rate effects of “attending” to visual stimuli be attributed to early or late selection in the visual system, assuming that visual perception is implemented in a hierarchy of mutually interacting stages of processing?

In general, there is a need to untangle the complex interaction between hypothetical cognitive mechanisms. On the one hand, one would like to establish a causal link between central mechanisms and their contribution to overall model behavior. On the other hand, one would like to identify those mechanisms, which either give rise to invalid behavior, or do not significantly contribute to overall model performance. Strictly speaking, this entails an experimentation with cognitive models using an experimental design where every cognitive mechanism in the chain or network of mechanisms involved in a cognitive task is systematically varied so that alternative implementations of individual mechanisms can be fully cross-combined.

For practical reasons, high-dimensional experimental designs are avoided in real-world, psychological experiments. However, in general such practical limitations do not apply to a computer simulation environment. Yet, the full factorial design (cf. section on ‘The two-level factorial design’ below) is not employed in cognitive modeling.

In a modeling project on mental imagery (Kovordányi, 1999b), we have adopted this approach and have systematically simulated alternative embodiments of a generic interactive activation model (McClelland, 1979; McClelland and Rumelhart, 1981, 1994/1988; Rumelhart and McClelland, 1982). Based on our experience with this project, we would like to point to the potentials of this method.

## Simulating cognitive models in a controlled experimental setting

The advocated method for exploring cognitive models may be conceived of as the equivalent of running a high-dimensional real-world experimental design with a multi-way analysis of co-variation (multi-way ANOVA). In this sense, the space of cognitive models is used as a virtual en-

vironment for experimentation: The structure of this environment is partially fixed by what we call the model framework. “The independent variables” correspond to those aspects of the cognitive model which cannot be specified in advance, but which may be potential determinants for the model’s overall behavior. “The dependent variable” constitutes a measure of model performance which, for purposes of model validation, should correspond to experimentally observed behavior in human subjects. Experimentation through systematic model simulation aims to shed light on how some of the “a priori” unknown aspects of the partially specified model interact in affecting the model’s behavior, and most importantly, whether a specific combination of model properties produces valid model behavior.

### The two-level factorial design

Systematic exploration of alternative model instances can be organized according to a full two-level factorial design (Law and Kelton, 1991; Box et al., 1978). This design emphasizes that the question of which model parameters are causally involved in a particular type of simulated behavior can be answered only if all parameters have been fully cross-combined. In order to keep down the computational cost of exploring all parameters, parameter values are varied between a predetermined min- and max-value, in what is called a two-level factorial design.

Note that, for the above reasons, if some model parameters were to be fixed at a given “reasonable value” in order to keep down simulation complexity, the power of the simulation design would diminish. Strictly speaking, such simulations cannot validate conclusions about which model properties are causally involved in the simulated behavior. Simply expressed, parameters may have been fixed at a value where they in fact interact with the central parameters of the model. Hence, for example, if no effect is obtained when the value of one of the central parameters is varied, this could in fact hide a significant negative effect, which is positively modulated by a peripheral parameter, which has been fixed.

Ideally, for a problem with  $k$  degrees of freedom, the minimal number of simulations which needs to be run in order to detect causal dependencies between model parameters is  $2^k$ . However, if the number of simulations turn out to be unmanageably large, a fractal two-level factorial design may be used instead of a full design (cf. Law and Kelton, 1991; Box et al., 1978). In these designs, peripheral parameters are not fixed at an ad hoc value, but are instead defined dynamically to be a function of other, more central parameters.

In addition to providing a minimally sufficient basis for detecting causal relationships in the simulation results, using a two-level factorial design renders the analysis of simulation results computationally simple. A simulation where  $k$  parameters are varied is captured in a design matrix of size  $2^k \times k$  containing +s and –s representing low and high parameter values (cf. Law and Kelton, 1991; Box et al., 1978). The way the matrix is set up, each row will represent a unique combination of parameter values, which in turn corresponds to a particular simulation run. As the design matrix

is regular, it is easy to set up. In addition, once it is computed, the same matrix can be used to control the simulations and to conduct data analysis.

To illustrate the latter case, if the possible interaction between parameters  $p_1$ ,  $p_3$ , and  $p_7$  are inquired, columns 1, 3, and 7 of the design matrix are multiplied value-by-value, and then multiplied with the set of simulation data. The effect of these multiplications is that the correct signs will be added to the data column. A final summation of all the signed entries in the data column, divided by  $2^{k-1}$ , where  $k$  denotes the number of model parameters, yields the desired mean interaction of the parameters involved (cf. figure 1).

run	par 1	par 2	par 3	sim. result
1	–	–	–	$R_1$
2	–	–	+	$R_2$
3	–	+	–	$R_3$
4	–	+	+	$R_4$
5	+	–	–	$R_5$
6	+	–	+	$R_6$
7	+	+	–	$R_7$
8	+	+	+	$R_8$

Figure 1: Example of a two-level full factorial simulation design matrix for three parameters. Each row in the matrix denotes a unique combination of parameter values. The last column in the design matrix designates the outcome of simulating a model (instance) for that particular parameter combination.

### Our modeling project

In our investigation of mental imagery, a full two-level factorial design was used where all parameters not inherently dependent on each other were cross-combined (Kovordányi, 1999b, 2000). While variations in the effect of several possible factors, such as the effect of mental image fading, were taken into account, simulation data analysis was centered around uncovering the effect of focusing early versus late selective attention on part of a mental image in a mental image reinterpretation task. As the empirical results of Finke and colleagues (Finke et al., 1989) and Peterson and colleagues (Peterson et al., 1992) used for model validation were qualitative, no attempt was made to optimize the models towards these data (Kovordányi, 2000). Model validity was instead defined qualitatively, and served as a means for “filtering out” invalid model instances.

### Parameterization of the model design space

The interactive activation model used in our project (cf. Kovordányi, 1998, 1999a) drew its main architectural components from the comprehensive model of mental imagery



forwarded by Kosslyn (1994; Kosslyn et al., 1979; Kosslyn et al., 1990). This model framework enabled us to capture all basic assumptions made at a higher, theoretical level, while enabling a systematic search for algorithmic details, which were left open by the theoretical and empirical basis.

How should an underconstrained model be partially specified so that it allows for a natural variation of model properties? One approach, used in our modeling project, is to set up a generic model framework as a localist network, and let each node in this network encode a holistic property or feature of the modeled phenomenon. In the case of visual perception, one kind of holistic property would be, for example, the individual line segments, which make up more complex line drawings.

One example of localist networks is the interactive activation model developed by McClelland and Rumelhart (McClelland, 1979; McClelland and Rumelhart, 1981, 1994/1988; Rumelhart and McClelland, 1982). In these models, the localist nodes are arranged into reciprocally connected layers of processing, thereby further increasing the structure and penetrability of the model. Units within the same processing layer are assumed to have the same inhibitory/excitatory connection weights. In such a model framework, model parameters can be naturally expressed as connection weights, activation thresholds, resting levels, or simply as “control flags”. These flags could, for example, control whether an individual simulation run should be initiated top-down or bottom-up in the interactive network.

Model parameters can arise naturally also in symbolic

models. Parameters in these models could be represented as alternative (sets of) production rules, or simply alternative definitions (fnc1 – fnc2) of a cognitive mechanism together with some means for activating them at run-time. Hence, in essence, any modularly built computational model can be parameterized with a minimal overhead cost.

## Simulations

Our model framework for mental imagery encompasses three mutually interacting layers of processing (figure 2). At the lowest level, the visual buffer contains detectors for oriented line segments. At the next stage, these feature detectors can evoke (and get feedback from) simple geometric patterns, such as composite lines or triangles, which are stored in visual long-term memory. At the highest level of processing, geometric patterns are combined into abstract concepts stored in amodal, associative long-term memory. In addition to between-layer connections, there is lateral, within-processing-level inhibition between mutually inconsistent (groups of) units. Interpretation in this system entails the dynamic establishment of a correspondence between low-level and higher-level representations.

We simulated mental and perceptual reinterpretation of two composite line drawings from Finke and colleagues (1989, exp. 1). Possible interpretations of these figures were limited to a small set of predefined geometric forms and abstract concepts. For example, possible interpretations of the first figure, formed from an upper case ‘H’ superim-

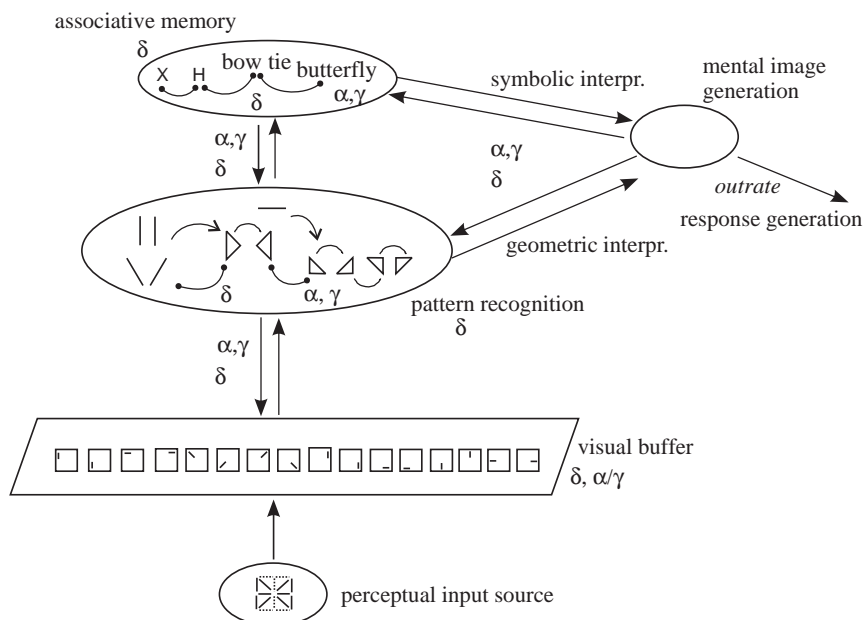


Figure 2: Communication and control structure of our model. Model parameters are shown as tags attached to the corresponding connection or subsystem. Note that model performance is expected to depend not only on how parameters are set, but also on whether the system is initiated top-down or bottom-up. These the two ways of initiating the system correspond to mental imagery and visual perception, respectively.

posed on an upper case 'X', were limited to "four small equilateral triangles", "two large isosceles triangles", "a butterfly", "a tilted hourglass" and "a bow-tie".

As layers in the system were reciprocally interconnected, simulations could be initiated either top-down or bottom-up. This made it possible to compare reinterpretation

performance in visual perception and in mental imagery. When simulations were run in mental mode, a chosen symbolic concept was activated in associative long-term memory, and this activation was projected into the visual buffer, where an activation pattern emerged which represented a visual mental image. When simulation was run in perceptual mode, visual input entered the system at the visual buffer, and was forwarded through consecutive stages of processing, and matched to geometric patterns and abstract concepts. One of these patterns or concepts was selected for verbal report.

Simulations were run through four phases: Mental image generation, followed by mental image reinterpretation, continued with a corresponding perceptual image build-up of the same line-figure, followed by perceptually based reinterpretation. Each simulation was run for 10 simulated seconds, in discrete simulation steps of 50 ms.

Two configurations of the model framework were scrutinized: One where attentional selection occurred late, affect-

ing processing at the level of associative long-term memory, and one where selection occurred early and directly affected the contents of the visual buffer. For these model configurations, the effect of focusing attention (versus not focusing attention) was investigated, taking into account the interaction effects that arose between this central, and other peripheral model parameters.

### Data analysis

In our project, data analysis was based on semi-automatic preparation of the raw simulation data. The prepared data were then visualized. The aim was to facilitate the discovery of significant parameter interactions, and in addition provide a basis for estimating model validity for the different parameter combinations. Below we briefly describe the key stages of this process.

### Identification of interacting parameters

Activation levels of all response units in the interactive activation network were measured for each simulation run, that is for each parameter combination (cf. Kovordányi, 1999b). From these activation values the corresponding probability for mental reinterpretation was calculated. Mental reinterpretation rates were considered valid if they qualitatively matched the reinterpretation rates obtained by Finke and

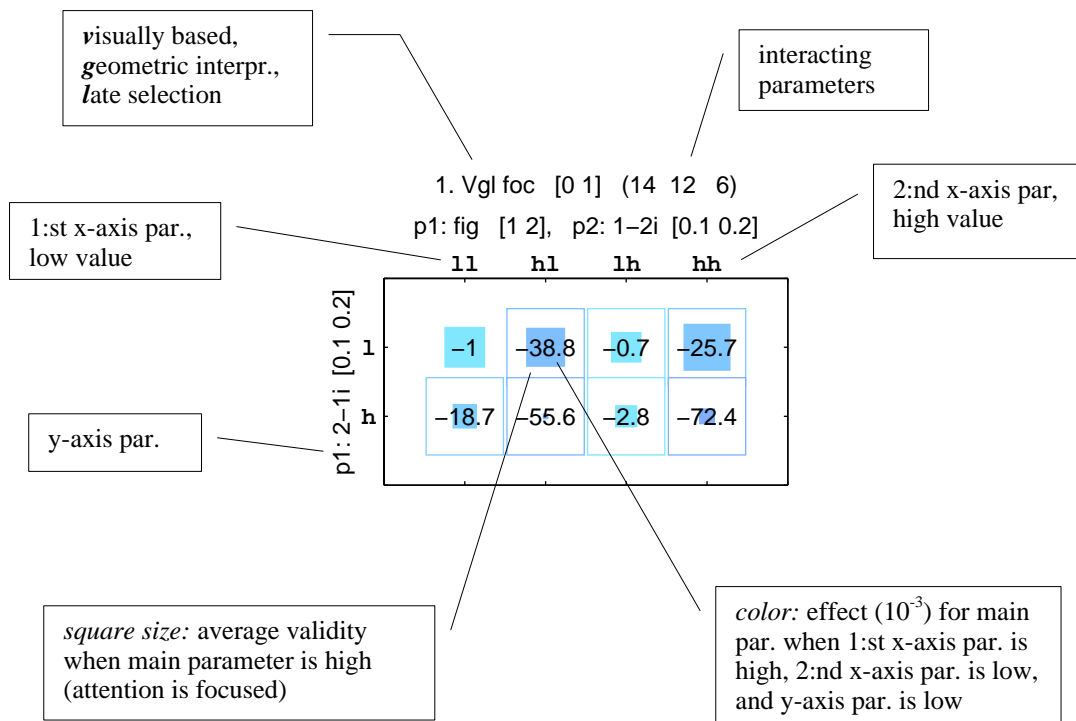


Figure 3: Example visualization of the simulation data. Data values are color-coded to support the understanding of interaction patterns. The area of the square markers reflects the validity of the underlying parameter combinations.

colleagues (1989, exp. 1), and Peterson and colleagues (1992). This amounted to the satisfaction of the following constraints: First, reinterpretation rates were required to be lower for abstract, conceptual interpretations than for geometric interpretations (cf. Finke et al., 1989). In addition, interpretations obtained during mental imagery had to be below those obtained during visual perception.

Second, reinterpretation rates were required to be qualitatively consistent with the findings of Peterson and colleagues (1992), which indicate that reinterpretation rates increase after a de- and refocus of attention.

### Calculation of parameter effects

The calculation of individual parameter effects and parameter interactions was based on a design matrix of  $-s$  and  $+s$ , representing high- and low parameter values (cf. figure 1). In this matrix each column denoted a model parameter and each row represented a specific parameter combination. A measure of model performance, that is simulated mental reinterpretation probability, was associated with each row in the design matrix. In general, in order to obtain a parameter's average effect on overall model performance, those rows in the model performance column of the design matrix which correspond to a low parameter value are summed and subtracted from those rows which correspond to high values. Higher-order interaction effects can be obtained in a similar manner (Law and Kelton, 1991; Box and Hunter, 1978). Given the simulation design matrix, these calculations can be expressed as a sequence of simple matrix operations.

### High-dimensional visualizations

Those groups of interacting parameters whose modulating effect exceeded 20% of the central parameter's effect—in our project this parameter denoted the focusing of attention—were prepared for subsequent visualization. Simulation data was prepared in such a way that parameters which exhibited a stronger mutual interaction with the central parameter would also be visualized closer to each other. This grouping of more related parameters turned out to enhance the understanding of interactions, since stronger interaction patterns emerged as salient color-patches.

The visualizations (illustrated in figure 3) can be conceived of as a high-dimensional cube of changes in model performance, each dimension representing changes caused by one of the interacting parameters. This cube can be sliced and stacked recursively onto a two-dimensional plot (cf. Bosan and Harris, 1996; Harris et al., 1994). Each x-y coordinate in these plots denotes a specific combination of interacting parameters. In our project, the direction of change in model performance was coded along two different color scales, and the magnitude of change was indicated by variations in hue within these scales, with deeper colors depicting a larger change.

The amount of information contained in the visualizations was further increased by the addition of information on model validity. We let the relative area of each colored square reflect the average validity of models corresponding to the central parameter's high value. In our case, this amounted to selective attention being focused. As a result of

including model validity in the visualizations, simulation data contributed to the visual appearance of the plot only to the extent to which they were valid.

### What type of results can be obtained?

Two categories of questions can be addressed using this method. First, simulation results can be approached with a particular hypothesis in mind, as was done in our project. In this case, one would like to make sure that the main effect of a particular embodiment of a cognitive mechanism,  $x_+$  (corresponding to parameter  $x$  at its high value), is as was predicted. For example: Do any of the interactions observed in the simulation results change the fact that parameter  $x$  is generally inhibitory? In addition, one would be interested in mapping out the validity of models where cognitive mechanism  $x_+$  is operating.

Second, simulation results can be openly explored, perhaps focusing on the role of a few central parameters. In this situation, one could, for example, be interested in finding out which cognitive mechanisms work in concert and which work against each other. In the first case the mechanisms would affect model performance in the same direction. In the latter case they would work in opposite direction, canceling out each other's effect. In addition to mapping out such interactions, one would be interested in which combination of mechanisms constitute valid models. This search for valid models can be a powerful way of constraining the space of possible models when several sources for validation are used (for example, a small set of seemingly contradictory experimental results).

### Concluding discussion

The use of distinctive colors, the organization of the visualizations' layout according to the strength of interactions, together with the technique described above for indicating model validity, turned out in practice to facilitate the understanding of the interaction patterns. Strong interactions which also gave rise to valid performance tended to visually coagulate into contiguous color-patches, which "popped-out" from the background of empty squares, marking non-valid cases.

The virtues of this combination of factorial simulation, analysis and visualization method are, in our view, compelling: Although the modeling framework is assumed to be based on a firm empirical basis, model properties which are not well-founded need not be specified in an ad hoc manner.

From a more theoretical perspective, conclusions which can be drawn from a full-factorial investigation will approach the stringency of appropriately conducted "real-world" experiments, with an inevitable difference: The validity of any results obtained will ultimately depend on the validity of the modeling framework itself. Within this framework, causal dependencies between hypothetical cognitive mechanisms and overall model behavior can be correctly mapped out. As a result, the development of subsequent models and/or the construction of cognitive theories can be guided in a stringent way.

As the method itself is qualitative in nature (parameters are varied coarsely between a high and a low value), models

can be validated on the basis of qualitative empirical data. Note that the objective with using this method is not primarily to quantitatively adjust a model's overt performance to empirical data by manually tuning parameters, but instead to single out a combination of internal cognitive mechanisms as the probable cause of empirically observed human behavior.

In a longer perspective, this method can contribute to the incremental development of more and more finely tuned cognitive models. Starting with a firmly based, minimally specified initial model framework, valid cognitive mechanisms can be singled out and subsequently embedded into the framework. Given these additional mechanisms, and/or having refuted some peripheral model properties, the next round of search can be narrowed down, and targeted at a more detailed level. As each increment is reasonably well-founded (validation is based on average simulation results), model development can be more directed.

### Acknowledgments

We would like to thank Sture Hägglund, and two anonymous reviewers for suggestions and valuable comments on an earlier version of this article. We would also like to thank Yvonne Wærn, and Jonas Barklund for fruitful discussions during the development of this project. This work was supported by the Swedish National Board for Industrial and Technical Development and the Swedish Council for Research in the Humanities and Social Sciences.

### References

- Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum.
- Bosan, S. & Harris, T. R. (1996). A visualization-based analysis method for multiparameter models of capillary tissue-exchange. *Annals of Biomedical Engineering*, 24, 124-138.
- Box, G. E. P., Hunter, W. G., & J. S. (1978). *Statistics for experimenters: An introduction design, data analysis, and model building*. New York: Wiley.
- Finke, R. A., Pinker, S. & Farah, M. J. (1989). Reinterpreting visual patterns in mental imagery. *Cognitive Science*, 13, 51-78.
- Harris, P. A., Sorel, B., Harris, T. R., Laughlin, H. & Overholser, K. A. (1994). Parameter identification in coronary pressure flow models: A graphical approach. *Annals of Biomedical Engineering*, 22, 622-637.
- Kieras, D. E. (1985). The why, when, and how of cognitive simulation. *Behavior Research Methods, Instrumentation, and Computers*, 17, 279-285.
- Kieras, D. E. (1987). Cognitive modeling. In Shapiro, S. C & Eckroth, D. (eds): *Encyclopedia of artificial intelligence, vol 1*. New York: Wiley.
- Kosslyn, S. M. (1980). *Image and mind*. Cambridge, MA: Harvard University Press.
- Kosslyn, S. M. (1994). *Image and Brain: The resolution of the imagery debate*. Cambridge, MA: MIT Press.
- Kosslyn, S. M., Pinker, S., Smith, G. E. & Swartz, S. P. (1979). On the demystification of mental imagery. *The Behavioral and Brain Sciences*, 2, 535-581.
- Kosslyn, S. M., Flynn, R. A., Amsterdam, J. B., Wang, G. (1990). Components of high-level vision: A cognitive neuroscience analysis and accounts of neurological syndromes. *Cognition*, 34, 203-277.
- Kovordányi, R. (1998). Is mental imagery symbolic? Exploratory simulations in an interactive activation model. In *Proceedings of Second European Conference on Cognitive Modeling*. Nottingham: Nottingham University Press.
- Kovordányi, R. (1999a). Mental image reinterpretation in the intersection of conceptual and visual constraints. In Paton, R. & Neilson, I. (eds): *Visual representations and interpretation*. London: Springer Verlag.
- Kovordányi, R. (1999b). Modeling and simulating inhibitory mechanisms in mental image reinterpretation—Towards cooperative human-computer creativity. *Linköping Studies in Science and Technology*. Dissertation no. 589. ISBN 91-7219-506-1. Linköping: Linköping University Press.
- Kovordányi, R. (2000). Full factorial simulation modeling of selective attention in mental imagery. Presented at XXVII *International Congress on Psychology*, Stockholm.
- Law, A. M. & Kelton, W. D. (1991). *Simulation modeling and analysis*. New York: McGraw-Hill.
- McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, 86, 4, 287-330.
- McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, 88, 5, 375-407.
- McClelland, J. L. & Rumelhart, D. E. (1994/1988). *Explorations in parallel distributed processing: A handbook of models, programs and exercises*. Cambridge, MA: MIT Press.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Rosenbloom, P. S., Laird, J. E., & Newell, A. (1993). *The Soar papers*. Cambridge, MA: MIT Press.
- Rumelhart, D. E. & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, 89, 1, 60-94.
- Peterson, M. A., Kihlstrom, J. F., Rose, P. M. & Glisky M. L. (1992). Mental images can be ambiguous: Reconstruals and reference-frame reversals. *Memory and Cognition*, 20, 107-123.

# Modeling infant learning via symbolic structural alignment

**Sven E. Kuehne** ([skuehne@ils.nwu.edu](mailto:skuehne@ils.nwu.edu))

Department of Computer Science, Northwestern University  
1890 Maple Avenue, Evanston, IL 60201 USA

**Dedre Gentner** ([gentner@nwu.edu](mailto:gentner@nwu.edu))

Department of Psychology, Northwestern University  
2029 Sheridan Rd., Evanston, IL 60201 USA

**Kenneth D. Forbus** ([forbus@ils.nwu.edu](mailto:forbus@ils.nwu.edu))

Department of Computer Science, Northwestern University  
1890 Maple Avenue, Evanston, IL 60201 USA

## Abstract

Understanding the mechanisms of learning is one of the central questions of Cognitive Science. Recently Marcus et al. showed that seven-month-old infants can learn to recognize regularities in simple language-like stimuli. Marcus proposed that these results could not be modeled via existing connectionist systems, and that such learning requires infants to be constructing rules containing algebraic variables. This paper proposes a third possibility: that such learning can be explained via structural alignment processes operating over structured representations. We demonstrate the plausibility of this approach by describing a simulation, built out of previously tested models of symbolic similarity processing, that models the Marcus data. Unlike existing connectionist simulations, our model learns within the span of stimuli presented to the infants and does not require supervision. It can handle input with and without noise. Contrary to Marcus' proposal, our model does not require the introduction of variables. It incrementally abstracts structural regularities, which do not need to be fully abstract rules for the phenomenon to appear. Our model also proposes a processing explanation for why infants attend longer to the novel stimuli. We describe our model and the simulation results and discuss the role of structural alignment in the development of abstract patterns and rules.

## Introduction

Understanding the mechanisms of learning is one of the central questions of cognitive science. Recent studies (Gomez & Gerken, 1999; Marcus, Vijayan, Rao & Vishton, 1999) have shown that showed that infants as young as seven months can process simple language-like stimuli and build generalizations sufficient to distinguish familiar from unfamiliar patterns in novel test stimuli. In Marcus et al's study, the stimuli were simple 'sentences,' each consisting of three nonsense consonant-vowel 'words' (e.g., 'ba', 'go', 'ka'). All habituation stimuli had a shared grammar, either ABA or ABB. In ABA-type stimuli the first and the third word are the same: e.g., 'pa-ti-pa.' In ABB-type stimuli the second and the third word are identical: e.g., 'le-di-di'. The infants were habituated on 16 such sentences, with three repetitions for each sentence. The infants were then tested on a different

set of sentences that consisted of entirely new words. Half of the test stimuli followed the same grammar as in the habituation phase; the other half followed the non-trained grammar. Marcus et al. found that the infants dishabituated significantly more often to sentences in the non-trained pattern than to sentences in the trained pattern.

Based on these findings Marcus et al. proposed that infants had learned abstract algebraic rules. They noted that these results cannot be accounted for solely by statistical mechanisms that track transitional probabilities. They further argue that their results challenge connectionist models of human learning that use similar information, on two grounds: (1) the infants learn in many fewer trials than are typically needed by connectionist learning systems; (2) more importantly, the infants learn without feedback. In particular, Marcus et al. demonstrated that a simple recurrent network with the same input stimuli could not model this learning task.

In response, several connectionist models have attempted to simulate these findings. Unfortunately, all of them to date include extra assumptions that make them a relatively poor fit for the Marcus et al experiment. For example, Elman (1999; Seidenberg & Elman, 1999) use massive pre-training (50,000 trials) to teach the network the individual stimuli. More importantly, they turn the infants' unsupervised learning task into a supervised learning task by providing the network with external training signals. Other models tailored to capture the data of the study seem unlikely to be applicable to other similar cognitive tasks (Altmann & Dienes, 1999). Using a localist temporal binding scheme, Shastri and Chang (1999) model the infant results without pretraining and without supervision, but still require an order of magnitude more exposure to the stimuli than the infants received.

We propose a third alternative. There is evidence that structural alignment processes operating over symbolic structured representations participate in a number of cognitive processes, including analogy and similarity (Gentner, 1983), categorization (Markman & Gentner, 1993), detection of symmetry and regularity (Ferguson, 1994), and learn-

ing and transfer (Gentner & Medina, 1998). Although these representations and processes are symbolic, they do not need to be rule-like, nor need they involve variables. Instead, we view the notion of correspondence in structural alignment as an interesting cognitive precursor to the notion of variable binding<sup>1</sup>. Correspondences between structured representations can support the projection of inferences, as the analogy literature shows, and therefore a symbolic system can draw inferences about novel situations even without having constructed rules. Moreover, as discussed below, comparison can be used to construct conservative generalizations. Across a series of items with common structure such a process of progressive abstraction can eventually lead to abstract rule-like knowledge. The attainment of rules, in those cases where it occurs, is the result of a gradual process. As we will show, symbolic descriptions can be used with structural alignment to model learning that is initially conservative, but which occurs fast enough to be psychologically realistic.

We first describe our simulation model of the Marcus et al task, which uses a simple combination of preexisting simulation modules, i.e., SME, MAGI, and SEQL. All of these modules have been independently tested against psychological data and independently motivated in prior modeling work. With the exception of domain-specific encoding procedures, no new processing components were created for this task. We then describe the results of our simulation of the Marcus et al data, showing that our simulation can learn the concepts within the number of trials that the infants had, without supervision and without pre-learning. We also show that the simulation can exhibit the same results with noisy input data. Finally, we discuss some of the implications of the symbolic similarity approach for models of cognitive processing.

### Modeling infant learning via structural alignment

A psychological model of the infants' learning must include the kind of input, the way the infants are assumed to encode the individual sentences, and the processes by which they generalize across the sentences. The architecture of our simulation is shown in Figure 1. We first describe our assumptions concerning the infants' processing capacities. Then we describe each component in turn.

**Processing Assumptions:** We assume that infants can represent the temporal order within the sentences (Saffran, Aslin & Newport, 1996). We further assume that the infants notice and encode identities within the sentences: for example, the fact that the last two elements match in an ABB sentence. This assumption is consistent with evidence that human infants, as well as with studies of nonhuman primates (Oden *et al*, in press), can detect identities. We also assume that infants can detect similarities between sequentially presented stimuli, consistent with studies of infant habituation, which demonstrate that infants respond to sequential sameness (e.g., Baillargeon, 1994).

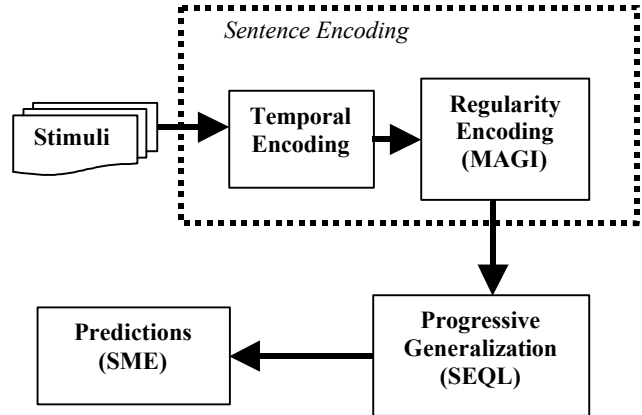


Figure 1: Simulation Architecture

**Input stimuli:** To make our simulation comparable with others, we use a representation similar to that of Elman (1999), namely, Plunkett & Marchman's (1993) distinctive feature notation. Each word has twelve phonetic features, which can be either present or absent. The presence or absence of each feature for each word is encoded by symbolic assertions. If feature  $n$  is present for word  $w$ , the assertion  $(Rn\ w)$  is included in the stimulus, and if absent, the assertion  $(Sn\ w)$  is included. Thus the acoustic features of each word are encoded as twelve attribute statements.

We modeled the Marcus et al experiment both without noise (Experiment 1) and with noise (Experiment 2). Marcus et al. used a speech synthesizer to control the pronunciation of the stimuli, but while this reduces variability, it cannot eliminate the possibility that the infant might encode something incorrectly.

**Temporal encoding:** We assume that the infant encodes the temporal sequence of the words in a sentence in two ways. First, each incoming word has an attribute associated with it, corresponding to the order in which it appears (i.e., FIRST, SECOND, or THIRD). We further assume that the infant encodes temporal relationships between the words in a sentence; to code this, an AFTER relation is added between pairs of words in the same sentence indicating their relative temporal ordering. The particular labels used in this encoding step are irrelevant – there are no rules in the system that operate on these specific predicates – the point is simply that infants are encoding the temporal order of words within sentences.

**Regularity Encoding:** We assume that the infants notice and encode identities within the sentences: for example, the fact that the last two elements match in an ABB sentence. Thus the simulation must incorporate a process that detects when words are the same. We use the MAGI model of symmetry and regularity detection (Ferguson, 1994) to automatically compute these relationships. MAGI treats symmetry as a kind of self-similarity, using a modified version of structure-mapping's constraints to guide the self-alignment process. MAGI has been successfully used with inputs ranging from stories to mathematical equations to visual stimuli,

<sup>1</sup> That structure-mapping algorithm neither subsumes, nor is subsumed by, traditional pattern matching such as unification is shown in Falkenhainer, Forbus, & Gentner (1988).

and it has done well at modeling certain aspects of visual symmetry, including making new predictions (Ferguson *et al* 1996). Here MAGI is used on the collection of words in a sentence. For any pair of words  $w1$  and  $w2$  that MAGI finds sufficiently similar, this module asserts (SIM  $w1$   $w2$ ), and a DIFF statement for every other pair of words in the sentence. (If MAGI does not find any pairs similar, DIFF statements are asserted for every pair of words.) This module also asserts (GROUP  $w1$   $w2$ ) for pairs of similar words, to mark that they form a substructure in the stimulus, and adds DIFF statements between groups and words not in the group. This use of MAGI is an example of what Ferguson (1994, in preparation) calls *analogical encoding*.

## SEQL

Once each sentence is encoded, we assume infants can detect the similarities between sequential pairs of sentences. The detection of structurally parallel patterns across a sequence of examples is modeled by SEQL (Skorstad, Gentner & Medin, 1988; Kuehne, Forbus, Gentner & Quinn, 2000), a model of the process of category learning from examples. SEQL constructs category descriptions via incremental abstraction. That is, the representation of a category is a structured description that has been generated by successive comparison with incoming exemplars. If the new exemplar and the category are sufficiently similar, the category description is modified to be their intersection -- i.e., the commonalities computed via structural alignment by a generalization algorithm. If the new exemplar is not sufficiently similar, it is stored separately and may later be used as the seed of a new category.

The structural alignment process is implemented via SME, (Falkenhainer *et al* 1988; Forbus *et al* 1994) a cognitive simulation of analogical matching. Here the base description is a category description, and the target description is the new exemplar. The structural alignments that SME computes are used in three ways by SEQL. First, the numerical structural evaluation score it computes<sup>2</sup> is used as a similarity metric, a numerical measure for deciding whether or not two descriptions are sufficiently similar. Second, the candidate inferences it computes serve as a model for category-based induction (c.f. Blok & Gentner, 2000; Forbus, Gentner, Everett, & Wu, 1997). Third, the correspondences in the best mapping SME produces serves as the basis for SEQL's generalization algorithm.

SEQL maintains a set of generalizations and a set of singular exemplars. When a new exemplar comes in, it is compared against existing generalizations to see if it can be assimilated into one of them. Otherwise, it is compared with the stored exemplars to see if a new generalization can be formed. If it is insufficiently similar to both the generalizations and the stored exemplars, it is stored as an exemplar itself.

SEQL begins with no generalizations; it simply stores its first exemplar. If the next exemplar is sufficiently close to the first, their overlap is stored as the first generalization. A

<sup>2</sup> Although SME can compute multiple mappings, we use the structural evaluation score of the best mapping, normalized by the size of the base description.

generalization consists of the overlap between the two input descriptions: that is, the shared structure found by alignment. Thus generalizations are structured descriptions of the same type as the input descriptions, although containing fewer specific features. If a new exemplar is sufficiently similar to a generalization (as determined comparing the structural evaluation score to a set threshold), then (a) the generalization is updated by retaining only the overlapping description that forms the alignment between the generalization and the exemplar; and (b) candidate inferences are projected from the generalization to the exemplar. Non-overlapping aspects of a description (e.g., phonetic features or relations that aren't shared) are thus "worn away" with each new assimilated description. (The threshold that determines when descriptions are sufficiently similar to be assimilated helps prevent descriptions from diminishing into vacuity.)

Returning now to the infant studies, we assume that babies are carrying out an ongoing process of comparing and aligning the incoming exemplars with an evolving generalization. We further assume that the relational candidate inferences from the general pattern to a new exemplar represent expectations on part of the infant.<sup>3</sup> When these expectations are violated by an incoming stimulus that does not fit the generalized pattern (e.g., an ABB test sentence after the ABA generalization has been formed), we assume the infant requires extra time to process the inconsistent stimulus.

## Simulation Experiments

In both experiments, we followed the procedure of Marcus *et al*. Each stimulus was a simple three-word sentence, encoded as described earlier. There were two sets of training stimuli, one following the ABA pattern and one following the ABB pattern. The training stimuli were (ABA) de-di-de, de-je-de, de-li-de, de-we-de, ji-di-ji, ji-je-ji, ji-li-ji, ji-we-ji, le-di-le, le-je-le, le-li-le, le-we-le, wi-di-wi, wi-je-wi, wi-li-wi, wi-we-wi and (ABB) de-di-di, de-je-je, de-li-li, de-we-we, ji-di-di, ji-je-je, ji-li-li, ji-we-we, le-di-di, le-je-je, le-li-li, le-we-we, wi-di-di, wi-je-je, wi-li-li, wi-we-we. The test stimuli in both experiments were four descriptions representing two novel ABA-type (ba-po-ba, ko-ga-ko) and two novel ABB-type sentences (ba-po-po, ko-ga-ga). The threshold value for SEQL was set to 0.85 in both experiments.

### Experiment 1

This experiment is most comparable to previous simulation models of the phenomena, in that we assume noise-free encoding of the stimuli. A simulation run consists of exposing SEQL to all of the stimuli from a particular training set (either ABA or ABB) once and then seeing the response given the four test sentences. To avoid possible biasing due to sequence effects (See Kuehne *et al.*, 2000), 20 simulation runs were made for each training set using different random

<sup>3</sup> SME can also produce attribute-level candidate inferences, and does so on these stimuli. We assume that, since these inferences concern directly perceivable features, testing them takes very little time.

orders. Identical match score and relational candidate inferences were produced for all sequences with a given stimulus set. In each case, SEQL produced a single generalization during the learning phase. For the test phase we used encodings of the corresponding stimuli used with infants, as noted above. Tables 1a and 1b show the results of this series for two generalizations paired against the four test sentences.

**Table 1a: ABA training stimuli**

Test Stimulus	Match Score	Candidate Inferences
<b>Ba-po-ba</b>	<b>0.658</b>	<b>None</b>
<b>Ko-ga-ko</b>	<b>0.689</b>	<b>None</b>
Ba-po-po	0.486	(DIFF po1 ba1) (DIFF po1 po2) (SIM ba1 po2)
Ko-ga-ga	0.455	(DIFF ga1 ko1) (DIFF ga1 ga2) (SIM ko1 ga1)

**Table 1b: ABB training stimuli**

Test Stimulus	Match Score	Candidate Inferences
Ba-po-ba	0.328	(SIM po1 ba2) (DIFF ba1 (GROUP po1 ba2))
Ko-ga-ko	0.350	(SIM ga1 ko2) (DIFF ko1 (GROUP ga1 ko2))
<b>Ba-po-po</b>	<b>0.776</b>	<b>None</b>
<b>Ko-ga-ga</b>	<b>0.753</b>	<b>None</b>

The in-grammar (bold) and out-of-grammar (plain text) matches show clear differences in their match scores. In-grammar matches are above 0.64 and do not generate relational candidate inferences. Out-of-grammar matches have match scores below 0.5, and lead to relational candidate inferences. Thus out-of-grammar test sentences lead to longer looking behavior, as predicted.

## Experiment 2

As noted earlier, we believe that noise-free stimulus encodings are unrealistic. Consequently, we used the same procedure as Experiment 1, but this time introducing noise into the representations for the training and test stimuli. For each sentence, one of the words was randomly picked, and one of its attributes (also chosen at random) was dropped or flipped, with the rest of its description being unchanged. Such changes can be significant: for example, flipping a single phonetic feature turns the word ‘de’ into the word ‘di’. Again, 20 simulation runs were made for each training set using different random orders. Naturally the match scores and, to a lesser degree, the generated candidate inferences, did vary across the individual runs. Tables 2a and 2b show the results. The scores were averaged over all 20 runs.

Although the noise affected the details of the computations, the overall pattern of results remains the same. The in-grammar (bold) match scores are far higher than the out-of-grammar (plain text) scores; and the out-of-grammar

stimuli produce relational candidate inferences while the in-grammar stimuli do not.

**Table 2a: ABA training stimuli**

Test Stimulus	Average Match Score	Candidate Inferences Min, Average, Max
<b>ba-po-ba</b>	<b>0.647</b>	<b>0, 0, 0</b>
<b>ko-ga-ko</b>	<b>0.682</b>	<b>0, 0, 0</b>
ba-po-po	0.435	2, 2.45, 3
ko-ga-ga	0.395	2, 2.55, 3

**Table 2b: ABB training stimuli**

Test Stimulus	Match Score	Candidate Inferences Min, Average, Max
ba-po-ba	0.339	2, 2, 2
ko-ga-ko	0.352	2, 2.05, 3
<b>ba-po-po</b>	<b>0.805</b>	<b>0, 0, 0</b>
<b>ko-ga-ga</b>	<b>0.783</b>	<b>0, 0, 0</b>

## Comparison with other models

The results of Marcus et al. (1999) have sparked an active debate focused on two issues: (1) Can current connectionist models (e.g., simple recurrent networks) model these results? (2) Do infants generate abstract rules that include variables?

Regarding the adequacy of simple recurrent networks, Marcus et al. state “Such networks can simulate knowledge of grammatical rules only by being consequently trained on all items to which they apply; consequently, such mechanisms cannot account for how humans generalize rules to new items that do not overlap with the items that appeared in the training.” Elman’s (1999) response describes his use of a simple recurrent network to model this task. Elman’s model requires tens of thousands of training trials on the individual syllables, and treats the problem as a supervised learning task, unlike the task facing the infants. By contrast, our simulation handles the learning task unsupervised, and produces human-like results with only exposure to stimuli equivalent to that given to the infants. Moreover, our model also continues to work with noisy data, something not true of any other published model of this phenomenon that we know of.

The learning in our model is due to the “wearing away” of non-identical phonetic attributes through subsequent comparisons. Although SEQL’s learning proceeds faster than connectionist models, it is still slower than systems that generate abstractions immediately (e.g., explanation-based learning (DeJong & Mooney, 1986)). In SEQL’s progressive alignment algorithm, the entities in the generalizations lose their concrete attributes across multiple comparisons, leaving the relational pattern of each grammar as the dominant force in the generalization only after a reasonable num-



ber of varied examples are seen.<sup>4</sup> There is considerable evidence for this kind of conservative learning (Forbus & Gentner, 1986; Medin & Ross, 1989).

Turning to the second issue, whether infants have variables and generate abstract rules, Marcus et al (1999) claims “[I]nfants extract abstract algebra-like rules that represents relationships between placeholders (variables), such as ‘the first item X is the same as the third item Y,’ or more generally that ‘item I is the same as item J.’” But our simulation does not introduce variables, in the sense commonly used in mathematics or logic. The generalizations constructed by SEQL do indeed include relational patterns that survive repeated comparisons because they are shared across the ingrammar exemplars. Furthermore, the entities (words) in the generalizations have many fewer features than the original words, as a result of the wearing away of features in successive comparisons. One could consider these patterns as a form of psychological rule, as proposed by Gentner and Medina (1998), with the proviso that the elements in the rule are not fully abstract variables, although they might asymptotically approach pure variables.

## Discussion

This paper proposes a third kind of explanation for the infant learning phenomena of Marcus et al (1999): incremental abstraction of symbolic descriptions via structural alignment. We believe our explanation is currently the best one for three reasons. First, it models the infant data with fewer extra concessions than previously published models (i.e., no pre-training, no supervision, and noisy data). Second, the processes we postulate are cognitively general; they apply to a large set of phenomena. Third, the abstraction processes we propose are consistent with research demonstrating that human learning is initially conservative (Brooks, 1987; Forbus & Gentner, 1986; Medin & Ross, 1989). Interestingly, there is ongoing research in developing symbolic connectionist models consistent with these processes (e.g., Holyoak & Hummel, 1997).

Many issues remain to be explored. For example, although our system does not introduce variables in its generalization process, there is a sense in which the entities in the generalization are on their way to becoming variables. Gentner and Medina (1998) have proposed that the process of progressive alignment can lead to rules. They further suggested that the application of rules to instances can be accomplished using the same general processes of structural alignment and projection that are used in analogy. The difference is that the base domain is an abstraction, the entities are ‘dummies’ with no features to either help or impede the match with the specific entities in the exemplar. Another issue concerns the incorporation of statistical notions in SEQL. Although SEQL is to a certain degree noise-resistant,

---

<sup>4</sup> SEQL learns with only one exposure to the 16 learning sentences, whereas Marcus’s infants received three exposures for each sentence. It is possible that the infants would have learned with only one pass; however it is also possible that the infants were less consistent in detecting the similarities than our simulation with its current parameters.

we suspect that to model large-scale learning, it will need to keep track of more statistical information than it does currently, so that properties wear away more slowly.

We note that it is common to conflate symbolic processing with rule-based behavior, and parallel processing with connectionist models. The model described here is symbolic, but it need not involve variables or rules. Further, it involves extensive parallel processing (most of SME and MAGI’s computations are parallel). Given the complexity of the phenomena, such confluations seem unwise.

The debates stirred by the Marcus et al. results bear on a critical issue in human learning and development: namely, what knowledge or mechanisms must be assumed to account for the rapid and powerful achievements demonstrated by infants in both cognition and language. Our results suggest that the general learning mechanism of structure-mapping theory may go a long way in accounting for these accomplishments.

## Acknowledgments

We thank Ron Ferguson, Ken Kurtz and Tom Mostek for valuable help and discussions. This research was supported by the Cognitive Science Division of the Office of Naval Research.

## References

- Altmann, G.T.M. and Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks, *Science* 284, 875.
- Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9(4), 321-324.
- Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science*, 3(5), 133-140.
- Blok, S. V., & Gentner, D. (2000). Reasoning from shared structure. *Proceedings of the 22<sup>nd</sup> Meeting of the Cognitive Science Society*.
- Brooks, L. R. (1987). Decentralized control of categorization: The role of prior processing episodes. In U. Neisser (Ed.), *Concepts and conceptual development: The ecological and intellectual factors in categorization* (pp. 141-174). Cambridge: Cambridge University Press.
- Christiansen, M.H. and Curtin, S.L. (1999). The power of statistical learning: No need for algebraic rules, in Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society, Erlbaum, Mahway, NJ.
- Christiansen, M.H. and Curtin, S.L. (1999). Transfer of learning: rule acquisition or statistical learning? *Trends in Cognitive Science* 3, 289-290
- DeJong, G.F. and Mooney, R.J. (1986). Explanation-based learning: An alternative view. *Machine Learning* 1(2), pp. 145-176
- Elman, J. (1999). Generalization, rules, and neural networks: A simulation of Marcus et. al, (1999). Ms., University of California, San Diego.
- Falkenhainer, B., Forbus, K., and Gentner, D. (1986). The Structure-Mapping Engine. In: *Proceedings of AAAI 86*, Philadelphia, PA, August.

- Falkenhainer, B., Forbus, K.D. and Gentner, D. (1989). The Structure Mapping Engine: an algorithm and examples. *Artificial Intelligence*, 41: 1-63
- Ferguson, R.W. (1994). MAGI: A model of analogical encoding using symmetry and regularity. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Ferguson, R.W., Aminoff, A. and Gentner, D. (1996). Modeling qualitative differences in symmetry judgments. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Forbus, K. D., & Gentner, D. (1986). Learning physical domains: Toward a theoretical framework. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2, pp. 311-348). Los Altos, CA: Kaufmann.
- Forbus, K. D., Ferguson, R. W., and Gentner, D. (1994). Incremental Structure-mapping. In: *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cognitive Science*, 7: 155-170.
- Gentner, D. and Markman, A.B. (1997). Structure-mapping in analogy and similarity. *American Psychologist*, 52, 45-56.
- Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, 65, 263-297.
- Goldstone, R.L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition* 52(2), 125-157.
- Goldstone, R.L., Medin, D.L., and Gentner, D. (1991). Relational similarity and the non-independence of features in similarity judgements. *Cognitive Psychology*, 23, 22-264.
- Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition* 70,109-135.
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3), 427-466.
- Kuehne, S.E., Forbus, K.D., Gentner, D. and Quinn, B. (2000). SEQL- Category learning as incremental abstraction using structure mapping, *Proceedings of the Twenty-second meeting of the Cognitive Science Society*.
- Marcus, G.F., Vijayan, S., Bandi Rao, S. and Vishton, P.M. (1999). Rule-learning in seven-month-old infants. *Science*, Vol. 283, 77-80
- Marcus, G.F. (1999). Do infants learn grammar with algebra or statistics?, Response to Seidenberg & Elman, Negishi, and Eimas. *Science* 284, 436-437
- Marcus, G.F. (1999). Simple recurrent networks and rule-learning: <http://psych.nyu.edu/~gary/science/es.html>.
- Markman, A.B. and Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, 25, 431-467.
- McClelland, J.L. and Plaut, D.C. (1999). Does generalization in infant learning implicate abstract algebraic rules?, *Trends in Cognitive Science* 3, 166-168
- Medin, D.L., Goldstone, R., and Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100(2), 254-278.
- Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem-solving, and induction. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 5, pp. 189-223). Hillsdale, NJ: Erlbaum.
- Oden, D. L., Thompson, R. K. R., and Premack, D. (in press). Can an ape reason analogically? Comprehension and production of analogical problems by Sarah, a chimpanzee (Pan troglodytes). In D. Gentner, K. J. Holyoak, & B. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science*. Cambridge, MA: MIT.
- Plunkett, K. and Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, 48, 21-69.
- Saffran, J., Aslin, R. and Newport, E. (1996). Statistical learning by 8-month-old infants, *Science*, 274, 1926-1928
- Seidenberg, M.S. and Elman, J. (1999), Do infants learn grammar with algebra or statistics?, *Letter, Science* 284, 434-436
- Seidenberg, M.S. and Elman, J. (1999). Networks are not hidden rules, *Trends in Cognitive Science* 3, 288-289
- Skorstad, J., Gentner, D. and Medin, D. (1988). Abstraction processes during concept learning: a structural view. In: *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Montreal: Lawrence Erlbaum Associates.

## An Optimality-Theoretic Model of Acquisition of Tense and Agreement in French

**Géraldine Legendre** ([legendre@cogsci.jhu.edu](mailto:legendre@cogsci.jhu.edu))

Department of Cognitive Science, Johns Hopkins University; 3400 N. Charles Street  
Baltimore, MD 21218 USA

**Paul Hagstrom** ([hagstrom@cogsci.jhu.edu](mailto:hagstrom@cogsci.jhu.edu))

Department of Cognitive Science, Johns Hopkins University; 3400 N. Charles Street  
Baltimore, MD 21218 USA

**Marina Todorova** ([todorova@cogsci.jhu.edu](mailto:todorova@cogsci.jhu.edu))

Department of Cognitive Science, Johns Hopkins University; 3400 N. Charles Street  
Baltimore, MD 21218 USA

**Anne Vainikka** ([vainikka@cogsci.jhu.edu](mailto:vainikka@cogsci.jhu.edu))

Department of Cognitive Science, Johns Hopkins University; 3400 N. Charles Street  
Baltimore, MD 21218 USA

### Abstract

We present a novel theoretical model of multiple stages in the acquisition of tense and agreement in Child French. First, we show that tense and agreement inflection follow independent courses of acquisition. Over the three stages of development attested in the data, tense production starts and ends at near-adult levels, but suffers a “dip” in production at the second stage. Agreement develops linearly, going roughly from none to 100% over the same time. This profile suggests a *competition* between tense and agreement at the second stage which is naturally expressed in terms of constraint violability and constraint re-ranking (Optimality Theory, Prince & Smolensky, 1993). By incorporating the further mechanism of partial rankings of constraints, our analysis successfully predicts, over three stages, the frequency with which children use tensed, agreeing, and nonfinite verbs.

### The Attested Development of Tense and Agreement in French

It is cross-linguistically well-attested that young children (around the age of 2) often produce simple sentences with a non-finite root (verb) form (NRFs), ungrammatical in the adult language, while also producing adult-like finite verbs with tense and agreement marking (Wexler, 1994, 1998, *inter alia*). What has been previously overlooked, however, is whether the distinct inflectional categories of tense and (person/number) agreement develop independently over time. A detailed analysis of spontaneous speech production data from three French children from the CHILDES Database (MacWhinney & Snow, 1985) provides strong evidence that the two categories indeed follow a different path of development.<sup>1</sup>

As a preliminary step of analysis the relevant files were analyzed by hand and classified into PLU (Predominant Length of Utterance) stages (Vainikka, Legendre & Todorova, 1999). This independent measure refines the traditional observation that children progress through one-word, two-word, and multi-word stages and has proven better suited to capture syntactic development than the well-known MLU measure (Brown, 1973).

French presents specific challenges for a study of the development of finiteness because the overwhelming majority of verbs used by young children belong to the first conjugation class (‘-er verbs’) which displays considerable homophony across morphological person inflections. In the absence of an overt subject (which is frequently omitted by young children) it cannot be determined whether a given phonetic form like [dās] *danse* ‘dance’ in the present tense carries correct agreement in person and number. However, clitic subject pronouns in French (e.g., *je* ‘I’) provide a diagnostic

---

crosslinguistically (Vainikka, Legendre & Todorova 1999). We have done in-depth work on transcripts from eight different children at this stage of development, covering English, French, Polish, Russian, and Swedish. Further analyses are underway as well. We should point out that while the conclusions drawn in this paper are made on the basis of data from (only) three children, we believe (following well-established tradition in the study of the acquisition of syntax by children) that examining a small number of subjects in detail allows us to uncover complexity that would be missed in a necessarily less detailed overview of a larger group of subjects. Furthermore, there is strong evidence to suggest that syntactic acquisition proceeds in a highly constrained and species-universal manner. Given this, we do not expect to find a great deal of variation from child to child, increasing the likelihood that the results reached on the basis of these three children will generalize across French-speaking children.

<sup>1</sup> The data we report on in this paper is part of a larger project aimed at studying the acquisition of tense and agreement

for agreement marking: Following Lambrecht (1981), Suñer (1988), Legendre (1999), and others, we take subject clitics to be an overt realization of agreement, rather than considering them to be overt subjects. Agreeing with Pierce (1992), we consider them to provide a reliable diagnostic for finiteness in child French. We also count verbs with a finite morphological shape which occur with an appropriate overt subject as agreeing.

The widespread tendency of young children to omit auxiliaries raises another issue with respect to coding. A past participle with no auxiliary has an adjectival use in adult French, and in the absence of an auxiliary it is nearly impossible to determine which use was intended by the child (adjective, main verb, or past tense). Similarly, a bare infinitive might represent either the future tense with no auxiliary, or a true NRF. We have coded only forms of the verb consisting of both the auxiliary and the participle/infinitive as instantiations of tense. Participles and infinitives used without the auxiliary were coded as non-finite forms.

We calculated the proportions of forms morphologically inflected for tense and/or agreement out of the total number of verbs produced by each child at each attested PLU stage. It is well-known that the third person singular and present tense forms are the first to appear in child productions, and for a time may be the only finite forms produced by the child. Furthermore, young children tend to overuse third person singular and present tense forms. This suggests that these serve as “default” forms, making it unclear whether a third person singular (3sg) verb is truly agreeing with a 3sg subject or whether it lacks agreement and is taking on an “elsewhere” form (see also Ferdinand, 1996). To determine the proportion of children’s verbs which actually show

agreement (and not a default form), we have counted only non-3sg and non-present forms as unambiguously showing agreement, and we present our results in these terms.

Tables 1–2 below summarize our findings relating to the use of tense and agreement, respectively, by each child. The numbers in Table 1 show the proportion of tensed verbs which had non-present forms, those in Table 2 show the proportion of verbs which appeared with non-3sg agreement. Of the verbs showing present tense or 3sg agreement, some presumably reflect a default form, while others reflect correct 3sg agreement or present tense. We will estimate the proportion of correct vs. default tense/agreement marking following the discussion below.

The combined results are graphed in Figure 1 to illustrate the development of tense and agreement across the attested PLU stages.

Because the figures in Table 1 are the percentages of overall utterances that contain *non-present* tense forms in Table 1, we do not ever expect these figures to reach 100%. To understand what level of production these percentages correspond to, we need to know what *adult-like* production of non-present tense forms is. To determine this, we ran a similar count on the adult utterances in two of the CHILDES files (Philippe 11 and Grégoire 9) in order to get at least a reasonable estimate of what adult use of non-3sg and non-present forms is. These results are given in Table 3.

Assuming that adults always produce finite verbs and given that they produce non-present tense verbs roughly 31% of the time, we can reasonably take the children’s 35% production of non-present tense (out of unambiguously tensed verbs) at stage 3b to be an adult-like level of production. On the other hand, we can also reasonably assume that the 4% production of non-3sg (of agreeing verbs) at stage 3b indicates that the children are not realizing agreement and are using a default (3sg) form.

As we can see from Tables 1–2 and Figure 1 below, tense and agreement undergo distinct patterns of development. At stage 3b, the proportion of agreeing forms in the children’s speech is negligible—it is clear that they are not yet using agreement. At the same time, the proportion of tensed forms is sufficiently high to allow us to conclude that tense is already in regular use. At the subsequent stage, 4b, agreement emerges at a significant, though not yet adult-like, level. Notice that at stage 4b, tense suffers a dip

Table 1: Verbs with non-present tense inflection (out of unambiguously tensed verbs)

Child	Stage 3b	Stage 4b	Stage 4c
G	34% (66/194)	21% (44/212)	32% (205/646)
S	37% (19/52)	10% (17/179)	25% (34/135)
P		13% (44/334)	30% (74/246)
Avg	35% (85/246)	15% (105/725)	31% (313/1027)

Table 2: Verbs with non-3sg agreement inflection (out of unambiguously agreeing verbs)

Child	Stage 3b	Stage 4b	Stage 4c
G	3% (5/156)	19% (33/172)	34% (221/650)
S	5% (2/43)	12% (13/109)	38% (51/133)
P		15% (44/303)	40% (98/246)
Avg	4% (7/199)	15% (90/584)	36% (370/1029)

Table 3: Adult usage of non-3sg and non-present tense

Adults from file	non-present	non-3sg
Grégoire 9	28% (184/661)	35% (231/659)
Philippe 11	34% (173/507)	41% (205/506)
Avg	31% (357/1168)	38% (437/1165)

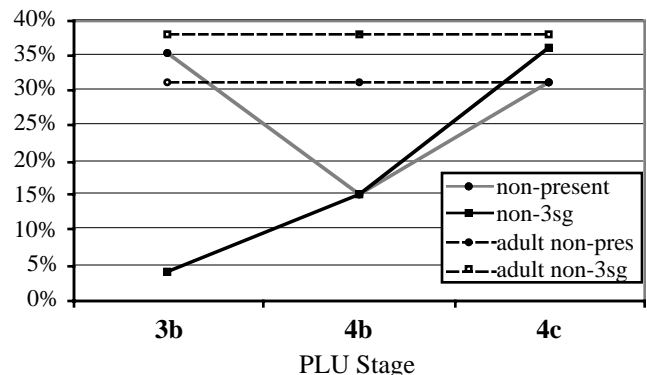


Figure 1. Tense and agreement

in production compared to stage 3b. This interesting correlation between increased use of agreeing forms and decreased use of tensed forms suggests a temporary competition between the two before they both stabilize at the subsequent stage, 4c.

The dissociation between tense and agreement is especially striking in the child production of periphrastic tenses; throughout stage 3b, Grégoire and Stéphane produce numerous instances of the past and future tenses; however, the auxiliary that appears in these utterances is always 3rd person singular: *Papa et Maman est parti* ‘Father and Mother is gone’ (Grégoire 2;0.5).

Turning now to NRFs, we found that children produce steadily fewer of these as their age/PLU stage increases. Our findings are given in Table 4 and graphed in Figure 2.

Comparing Figures 1–2, we can see that the reduction in the use of NRFs over time appears to be inversely correlated with the development of agreement: in a sense, the NRF pattern is the mirror image of the pattern we have found for agreement (Table 2). By contrast, the decrease in NRFs does not appear to correlate with the development of tense; compare Figure 2 to the previous graph of tense (Figure 1). This observation is important in light of existing claims that relate the occurrence of NRFs to the development of Tense. For example, Wexler (1994) has proposed that the underspecification of Tense is responsible for the presence of NRFs (his ‘root infinitives’) in young children’s speech. Our findings suggest at the very least that the development of agreement is also involved; the profile of NRFs is not directly linked to the profile of realization of tense.

### An Optimality-Theoretic Model of Development

Informally, the main idea behind our proposal is the following: At Stage 3b, constraints requiring realization of finiteness compete with constraints on economy of structure,

Table 4: Non-finite root forms (NRFs) out of all verbs

Child	Stage 3b	Stage 4b	Stage 4c
G	28% (83/297)	18% (51/287)	1% (7/711)
S	48% (51/106)	13% (27/205)	2% (3/152)
P		22% (105/476)	6% (14/250)
Avg	33% (134/403)	19% (183/968)	2% (24/1113)

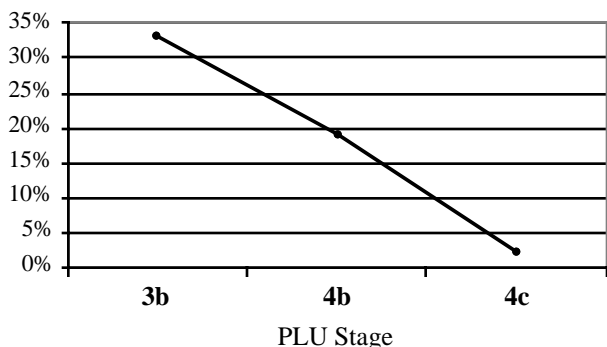


Figure 2. Non-finite root forms out of all verbs

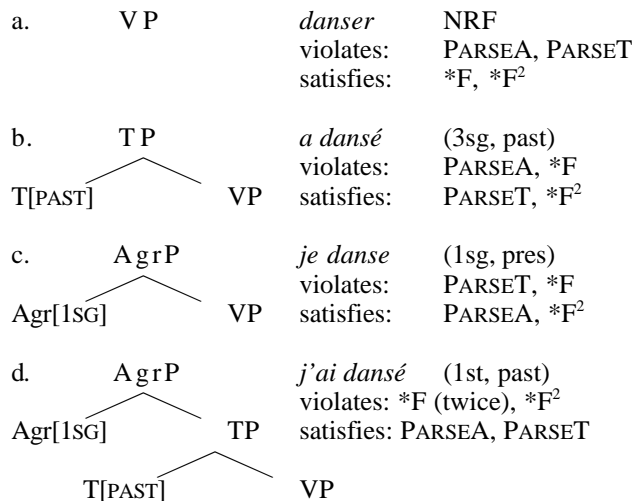


Figure 3. Candidates for input containing past tense and 1st singular agreement features

sometimes resulting in the production of finite verbs and sometimes resulting in the production of NRFs. At Stage 4b, tense and agreement compete for a single structural position; a functional projection which can realize the features either of tense or of agreement (but not both). At Stage 4c, two positions are available, allowing both tense and agreement features to be realized without competition.

Formally, the constraints which require parsing of the functional features (PARSET(ense), PARSEA(reement)) rise in the ranking relative to a fixed hierarchy of constraints penalizing structure. PARSET and PARSEA are Faithfulness constraints ensuring that what is expressed (the output of the grammar) differs minimally from what is intended (the input to the grammar).

Assuming, as is standard in generative syntax since Pollock (1989), that the presence of inflectional categories is indicative of phrase structure above that instantiating lexical categories, the constraints penalizing structure can be stated as \*F (‘No functional heads’) and \*F<sup>2</sup> (‘No pairs of functional heads’) with the invariable ranking \*F<sup>2</sup> >> \*F.<sup>2</sup>

There are four candidate structures relevant to this analysis (we assume that the input to every evaluation has tense and agreement features subject to Faithfulness constraints). They are given in Figure 3 along with examples and the constraints each satisfies and violates.

The key to our proposal is the ability of the Faithfulness constraints to “float” over a certain range in the ranking (unlike the Economy of Structure constraints \*F, \*F<sup>2</sup> discussed above, which remain fixed in their relative ranking) during the course of development. Formally, the model relies on partial constraint ranking (Reynolds, 1994; Anttila 1997, Nagy & Reynolds, 1997) and can predict not only that we see variation in outputs of the developing grammar, but also with what frequency we will see each output. As illustrated in Figure 4, a partial ordering (a) translates into a set of

<sup>2</sup> F<sup>2</sup> >> F invariably because they are part of a Power Hierarchy: F<sup>2</sup> is a local conjunction of two instances of \*F (Legendre, Smolensky & Wilson, 1998; Smolensky, 1995).

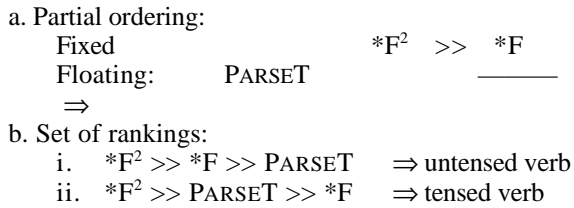


Figure 4. Partial ordering and resulting rankings

rankings (b). We see that a different candidate structure wins under each of the rankings in (b); under ranking (bi), a candidate with a nonfinite verb wins, while under ranking (bii) a candidate with a tensed verb (that is, with a functional projection to realize tense features). For any given evaluation, a grammar with the partial ordering in (a) will use one of the rankings, either (bi) or (bii), to determine the optimal candidate. Thus, in any given evaluation, either a tensed verb or an untensed verb will win the competition. We make the further assumption that either of the two rankings has an equal chance of being called upon during an evaluation (for a different assumption see Boersma, 1997). This means that there is a 50% chance that ranking (bi) will be used, yielding an untensed verb as the optimal candidate. To put it another way, we expect to see the untensed candidate 50% of the time (and to see the tensed candidate the other 50% of the time).

This example illustrates well the nature of the conflict underlying the development of finiteness. Functional features can only be parsed (satisfying the Faithfulness constraints PARSET and/or PARSEA) if the Economy of Structure constraints (\*F and possibly \*F<sup>2</sup>) are violated. The conflict is resolved by ranking. If Economy of Structure dominates Faithfulness, then functional features cannot be parsed and the optimal candidate will be a nonfinite form acting as a main verb (an NRF). If Faithfulness dominates Economy of Structure, then functional features will be parsed into a functional head, yielding a finite form as the optimal candidate (recall that *either* tensed *or* agreeing forms count as “finite” under our terminology).

This analysis, set in OT and using partial rankings, also unifies two otherwise incompatible views on the acquisition of syntactic structure. On one side, the Full Competence Hypothesis (or “Strong Continuity”; see, e.g., Poeppel & Wexler 1993) proposes that the full adult-like syntactic structures are available to the child’s grammar essentially from the outset of acquisition. On the other side, the Structure Building Hypothesis (or “Weak Continuity”; see, e.g., Vainikka 1993/4) proposes that a child initially uses syntactic structures much simpler than those of the adult language, over time adding complexity until the adult stage is reached. Our analysis shares with Full Competence the idea that the full range of the adult grammar is available to children; all of the constraints are present (but ranked in a non-adult way), the difference between a child grammar and an adult grammar is of the same type as the difference between two adult grammars (i.e., differing rankings among constraints), and the underlying representations used by children and adults draw from the same set of grammatical features. Simultaneously, our analysis shares with the Structure Build-

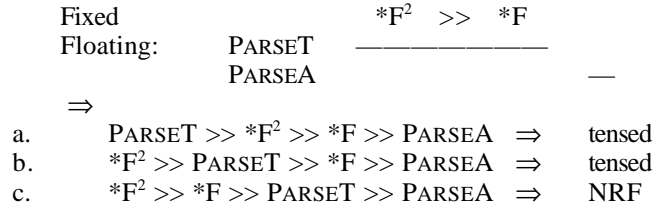


Figure 5. Stage 3b

ing Hypothesis the view that children’s representations are simpler than adult structures to begin with and become more complex over time. On our analysis, this is not due to a lack of access to adult grammatical constructs, but rather to a low ranking of Faithfulness constraints relative to constraints prohibiting structure.

The actual course of development of finiteness we propose here is an expanded version of this basic re-ranking schema. We will see that the PARSE constraints advance separately, at one point (Stage 3) with PARSET invariably outranking PARSEA with the result that the observed finite forms will be tensed, but non-agreeing.

**A Stage-By-Stage Analysis of Development**

We begin with stage 3b, where the rankings are as in Figure 5, yielding the 3 rankings given in (a–c).

At stage 3b, PARSET spans a range allowing it to sometimes outrank \*F<sup>2</sup>, and sometimes be outranked by \*F. PARSEA is always outranked by both \*F and \*F<sup>2</sup>.

Of the three rankings, only (c) results in an NRF; under this ranking, it is better not to have a functional projection (satisfying \*F) than to parse tense (which would satisfy PARSET) or agreement (which would satisfy PARSEA). This means that we expect NRFs to comprise one-third of a child’s utterances at Stage 3b.

The other two rankings yield a tensed form, but without agreement. Under these two rankings, PARSET outranks \*F, making it more important to realize tense in a functional projection than to avoid functional projections. Neither ranking yields an agreeing form because this would require two functional projections, and PARSEA is under both rankings outranked by \*F<sup>2</sup>. Thus, we expect tensed forms (without agreement) to comprise the other two-thirds of a child’s utterances at stage 3b.

What we actually observed (Table 4) was 33% NRFs and 67% finite forms, exactly the prediction. Of the finite forms, we counted only non-present forms, and found 35% such forms (Table 1). Recall that when this is compared to the adult production of 31% non-present forms (Table 3), it appears that all finite utterances the children produce at Stage 3 are tensed. Looking at agreement (Table 2), we found very few (4%) non-3sg forms, compared to an adult rate of 38%. So (with a minimal degree of idealization) we find that all finite child utterances at Stage 3 are tensed but non-agreeing, as predicted.

In Stage 4b, PARSEA advances to a position equal to PARSET; both now sometimes outrank \*F<sup>2</sup>, and can sometimes be outranked by \*F. In some rankings PARSET outranks PARSEA, while in others PARSEA outranks PARSET. These ranges yield the 12 rankings given in Figure 6.

Fixed		$*F^2$	>>	$*F$	
Floating:	PARSET				
	PARSEA				
	⇒				
a.	PARSET>>PARSEA>> $*F^2$ >> $*F$	⇒	tensed, agreeing		
b.	PARSEA>>PARSET>> $*F^2$ >> $*F$	⇒	tensed, agreeing		
c.	$*F^2$ >> $*F$ >>PARSET>>PARSEA	⇒	NRF		
d.	$*F^2$ >> $*F$ >>PARSEA>>PARSET	⇒	NRF		
e.	$*F^2$ >>PARSET>>PARSEA>> $*F$	⇒	tensed		
f.	$*F^2$ >>PARSEA>>PARSET>> $*F$	⇒	agreeing		
g.	PARSET>> $*F^2$ >>PARSEA>> $*F$	⇒	tensed		
h.	PARSEA>> $*F^2$ >>PARSET>> $*F$	⇒	agreeing		
i.	PARSET>> $*F^2$ >> $*F$ >>PARSEA	⇒	tensed		
j.	PARSEA>> $*F^2$ >> $*F$ >>PARSET	⇒	agreeing		
k.	$*F^2$ >>PARSET>> $*F$ >>PARSEA	⇒	tensed		
l.	$*F^2$ >>PARSEA>> $*F$ >>PARSET	⇒	agreeing		

Figure 6. Stage 4b

Fixed		$*F^2$	>>	$*F$	
Floating:	PARSET				
	PARSEA				
	⇒				
a.	PARSET>>PARSEA>> $*F^2$ >> $*F$	⇒	tensed, agreeing		
b.	PARSEA>>PARSET>> $*F^2$ >> $*F$	⇒	tensed, agreeing		

Figure 7. Stage 4c

Two of these rankings, (a–b), yield verb forms which are both tensed and agreeing (that is, essentially adult forms), since under those rankings it is more important to realize both tense and agreement than it is to avoid having two functional projections. Another two rankings, (c–d), yield NRFs, since under these rankings it is more important not to have any functional projections than it is to realize either tense or agreement. The rest of the rankings (e–l) yield finite forms which are either tensed (when PARSET outranks PARSEA) or agreeing (when PARSEA outranks PARSET), but not both.<sup>3</sup>

This predicts, then, that only 17% (2 out of 12) of the verb forms uttered at Stage 4b should be NRFs. We observed (Table 4) 19% NRFs, very close to the prediction. Of the remaining verbs, all finite, 17% are predicted to be adult-like (with both tense and agreement), the remaining forms having only one or the other (33% of them with only tense, 33% of them with only agreement). Again, this lines up well with the observations. Further, of the finite verbs we predict 19% non-present forms and observe 15% (Tables 1

<sup>3</sup> Note that a higher degree of constraint overlap yields a larger number of possible rankings for each evaluation, but this does not mean that the child must “exert more effort to choose” where the number of possible rankings is large. If, metaphorically, the child’s task before evaluation is to choose a random position (within its range) for each constraint, the size of this task is affected only by the number of constraints, not by the amount of overlap. Looking in from outside, we can compute which rankings could result and what the individual likelihood is of each, but this has no effect on the actual process of fixing a ranking.

and 3), and predict 23% non-3sg forms and observe 15% (Tables 2 and 3).<sup>4</sup>

Compare stage 4b to stage 3b with respect to the realization of tense. Notice that, while at stage 3b, 100% of the finite utterances were tensed, at stage 4b only 60% (6 out of 10) of the finite forms are tensed. In other words, we predict (and in fact observe) a “dip” in the child’s production of tensed forms. If children were simply “learning tense” (speaking vaguely), we would not have expected them to get worse at any point during the course of development. The proposed analysis provides an explanation for this otherwise puzzling fact. Back in stage 3b, PARSEA was ranked so low as to ensure that tense features were realized in the single functional projection allowed. What has happened at stage 4b is that the tense features and agreement features now compete for realization in the single functional projection available. Since tense sometimes (in fact, half the time) loses to agreement, we predict the observed dip in the proportion of tensed forms, which coincides with an increase in the proportion of agreeing forms.

In the last stage covered in our data, stage 4c (Figure 7 above), PARSET and PARSEA together move to a position high enough in the hierarchy that they invariably outrank  $*F^2$ . This yields 2 rankings, but both produce the same optimal candidate, a finite form which realizes both Tense and Agreement. At this stage, we predict no NRFs, and we observed only 2% NRFs in child speech (Table 4). We also expect the children’s production of non-present forms and non-3sg forms to match the proportion in adult speech, which it does quite well; we observed (Tables 1–3) 31% non-present tense forms compared with 31% for adults, and 36% non-3sg forms compared with 38% for adults.

Figures 8–9 (next page) summarize graphically how the predictions of the model match the observed child data.

### Concluding Remarks

To sum up, our research has uncovered previously overlooked properties of the acquisition of tense and agreement. We found that tense and agreement in French follow distinct courses of acquisition over the three stages studied. While the use of tense starts and ends strong, it suffers a “dip” at the intermediate stage. Meanwhile, agreement develops in a more linear way while the proportion of NRFs drops, also linearly. The distinctive profile naturally leads to an analysis in which, at the intermediary stage, tense and agreement are competing for realization. In particular, an Optimality-Theoretic analysis making use of “floating constraints” (defining partial ranking orders) allow us to predict not only the occurrence of the observed types of utterances, but their frequency as well. This result is novel; previous analyses (e.g.

<sup>4</sup> The predictions here are again scaled by the “expected” proportion of non-present forms and non-3sg forms based on what we found in the observed adult speech (Table 3). 60% of forms are predicted to be tensed by our analysis, and adults produce 31% non-present forms, so we expect to find  $60\% \times 31\% = 19\%$  of (finite) child utterances to be in a non-present form. Similarly, since adults produce 38% non-3sg forms, we expect to find  $60\% \times 38\% = 23\%$  non-3sg forms in the children’s (finite) utterances.

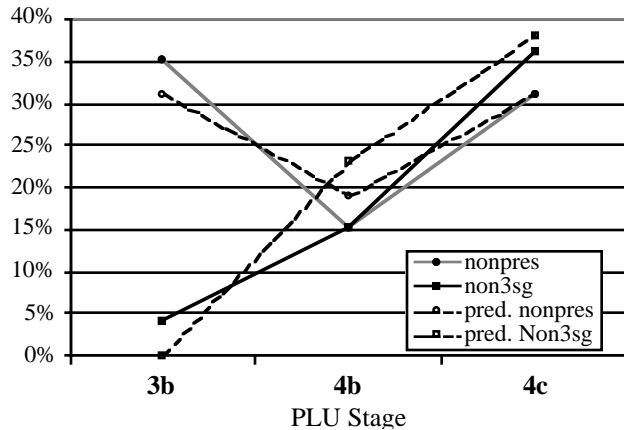


Figure 8. Predicted vs. observed tense and agreement data

Ferdinand, 1996; Pierce, 1992; Wexler 1994, 1998) have provided no clear way even to describe the facts about the changing frequencies of tense and agreement realizations over the course of acquisition. Under our proposal, the frequency predictions are a natural consequence of the re-ranking mechanism. The fundamental principle of OT, that grammars share the same constraints but rank them differently with respect to one another, requires that the acquisition process be one of re-ranking constraints. We have proposed that this re-ranking occurs not in sudden jumps but by spreading constraints across ranges in the rankings, narrowing in on the correct adult ranking. These “floating” or partially ranked constraints allow our model to make frequency predictions that seem to be borne out in child French.

### Acknowledgments

This research was supported by a National Science Foundation grant, Learning and Intelligence Systems number NSF-9720412.

### References

- Anttila, Arto (1997). *Variation in Finnish Phonology and Morphology*. Ph.D. dissertation, Stanford University.
- Boersma, P. (1997) “How We Learn Variation, Optionality, and Probability,” *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 21.
- Brown, R. (1973) *A First Language: The Early Stages*, Harvard University Press.
- Ferdinand, A. (1996) *The Development of Functional Categories, The Acquisition of the Subject in French*, Holland Institute of Generative Linguistics.
- Lambrecht, K. (1981) *Topic, Antitopic and Verb Agreement in Nonstandard French*, John Benjamins, Amsterdam.
- Legendre, G. (1999) “On the Status and Positioning of Verbal Clitics,” ms., Johns Hopkins University.
- Legendre, G., P. Smolensky, and C. Wilson (1998) “When is Less More? Faithfulness and Minimal Links in Wh-Chains,” in P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis, and D. Pesetsky, eds., *Is the Best Good Enough? Optimality and Competition in Syntax*, MIT Press, Cambridge, MA.

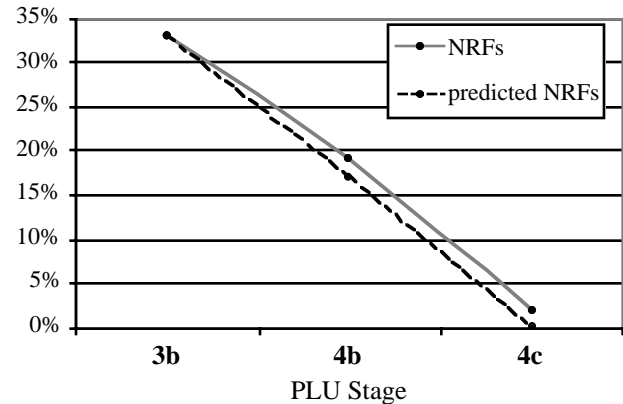


Figure 9. Predicted vs. observed NRF data

- MacWhinney, B. and C. Snow (1985) “The Child Language Data Exchange System,” *Journal of Child Language* 12, 271–96.
- Nagy, N. and B. Reynolds (1997) “Optimality Theory and Variable Word-final Deletion in Fætar,” *Language Variation and Change* 9, 37–55.
- Pierce, A. (1992) *Language Acquisition and Syntactic Theory: A Comparative Analysis of French and English Child Grammars*, Dordrecht, Kluwer.
- Poeppl, D., and K. Wexler (1993). “The Full Competence Hypothesis of Clause Structure in Early German,” *Language* 69, 1–33.
- Pollock, Jean-Yves (1989). “Verb Movement, Universal Grammar, and the Structure of IP,” *Linguistic Inquiry* 20, 365–424.
- Prince, A. and P. Smolensky (1993) *Optimality Theory: Constraint Interaction in Generative Grammar*, Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ, and Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder. To appear in *Linguistic Inquiry Monograph Series*, MIT Press, Cambridge, MA.
- Reynolds, B. (1994) *Variation and Phonological Theory*, Doctoral dissertation, University of Pennsylvania.
- Smolensky, P. (1995) “On the Internal Structure of the Constraint Component *Con* of UG,” paper presented at University of California, Los Angeles.
- Suñer, M. (1988) “The Role of Agreement in Clitic-Doubled Constructions,” *Natural Language and Linguistic Theory* 6, 391–434.
- Vainikka, A. (1993/4). “Case in the Development of Syntax.” *Language Acquisition* 3, 257–324.
- Vainikka, A., G. Legendre and M. Todorova (1999) “PLU Stages: An Independent Measure of Early Syntactic Development,” Department of Cognitive Science Technical Report, Johns Hopkins University.
- Wexler, K. (1994) “Optional Infinitives, Head Movement, and Economy of Derivation,” in N. Hornstein and D. Lightfoot, eds., *Verb Movement*, Cambridge University Press
- Wexler, K. (1998) “Very Early Parameter Setting and the Unique Checking Constraint: A New Explanation of the Optional Infinitive Stage,” *Lingua* 106, 23–79.



# Infinite RAAM: A Principled Connectionist Basis for Grammatical Competence

Simon Levy, Ofer Melnik and Jordan Pollack  
levy, melnik, pollack@cs.brandeis.edu  
Dynamical and Evolutionary Machine Organization  
Volen Center for Complex Systems,  
Brandeis University, Waltham, MA 02454, USA  
February 6, 2000

## Abstract

This paper presents Infinite RAAM (IRAAM), a new fusion of recurrent neural networks with fractal geometry, allowing us to understand the behavior of these networks as dynamical systems. Our recent work with IRAAMs has shown that they are capable of generating the context-free (non-regular) language  $a^n b^n$  for arbitrary values of  $n$ . This paper expands upon that work, showing that IRAAMs are capable of generating syntactically ambiguous languages but seem less capable of generating certain context-free constructions that are absent or disfavored in natural languages. Together, these demonstrations support our belief that IRAAMs can provide an explanatorily adequate connectionist model of grammatical competence in natural language.

## Natural Language Issues

In an early and extremely influential paper, Noam Chomsky (1956) showed that natural languages (NL's) cannot be modeled by a finite-state automaton, because of the existence of center-embedded constructions. A second and equally important observation from this work was that a minimally adequate NL grammar must be ambiguous, assigning more than one structure (interpretation) to some sentences, for example, *They are flying planes*.

The first observation led to the development of Chomsky's formal hierarchy of languages, based on the computational resources of the machines needed to recognize them. In this hierarchy, Chomsky's observation about center-embedding is expressed by saying that NL's are non-regular; i.e., they cannot be generated by a grammar having only rules of the form  $A \rightarrow bC$ , where  $A$  and  $C$  are non-terminal symbols and  $b$  is a terminal symbol.

Whether NL's are merely non-regular, belonging in the next, context-free (CF) level of the Chomsky hierarchy, or are more powerful, belonging further up in the hierarchy, became the subject of heated debate (Higginbotham 1984; Postal and Langendoen 1984; Shieber 1985). Non-CF phenomena such as reduplication/copying (Culy 1985) and crossed serial dependencies (Bresnan, Kaplan, Peters, and Zaenen 1982) suggested that a more powerful approach, using syntactic transformations (Chomsky 1957) was called for, but some researchers criticized transformations as having arbitrary power and thus failing to constrain the types of languages that could be expressed (Gazdar 1982). Further criticism of the entire formal approach came from observing that even CF grammars (CFGs) had the power to generate structures, such as a sequence followed by its mirror image, that did not seem to occur in NL (Manaster-Ramer 1986), or which placed an

extraordinary burden on the human parsing mechanism when they did occur (Bach, Brown, and Marslen-Wilson 1986).

## Connectionism and Natural Language

While debates about the complexity of NL were raging, connectionism was beginning to awaken from a fifteen-year sleep. In connectionist models many researchers found a way of embodying flexibility, graceful degradation, and other non-rigid properties that seem to characterize real cognitive systems like NL. This research culminated the publication of a highly controversial paper by Rumelhart and McClelland (1986) which provided a connectionist account of part of the grammar of English using a feed-forward neural network. The paper was soon criticized by more traditional cognitive scientists (Fodor and Pylyshyn 1988; Pinker and Prince 1988), who cited the non-generative nature of such connectionist models as a fundamental shortcoming of the entire field.

Partly in response to these criticisms, many connectionists have spent the past decade investigating network models which support generativity through recurrent (feedback) connections (Lawrence, Giles, and Fong 1998; Rodriguez, Wiles, and Elman 1999; Williams and Zipser 1989). The research we present here is an attempt to contribute to this effort while focusing as strongly as possible on the natural language issues described above. Such an attempt faces a number of challenges.

First, despite analysis of how a network's dynamics contribute to its generativity, it is often uncertain whether the dynamics can support generation of well-formed strings beyond a certain length. That is, it is unknown whether the network has a true "competence" for the language of which it has learned a few exemplars, or is merely capable of generating a finite, and hence regular, subset of the language.<sup>1</sup> Second, it is often easier to model weak, rather than strong generative capacity, by building networks that generate or recognize strings having certain properties, without assigning any syntactic structure to the strings. Third, this lack of syntactic structure inhibits the formulation of an account of syntactic ambiguity in such networks, making them less plausible as models of NL.

---

<sup>1</sup>To be fair, not all connectionists, or cognitive scientists, take seriously the notion that human language has infinite generative capacity. Though we obviously do not have the resources to argue the issue here, we are certain that a model with a provably infinite competence would be more persuasive to the cognitive science community as a whole than would a model without one.

In sum, we are concerned with formulating a recurrent network model that rigorously addresses the set of criteria that emerged from the long debate over the complexity of NL. As an candidate, the remainder of this paper presents a new formulation of RAAM (Pollack 1990), a recurrent network model that addresses the NL issues in a principled way.

### Traditional RAAM

Recursive Auto-Associative Memory or RAAM (Pollack 1990) is a method for storing tree structures in fixed-width vectors by repeated compression. Its architecture consists of two separate networks – an encoder network, which can construct a fixed-dimensional code by compressively combining the nodes of a symbolic tree from the bottom up, and a decoder network, which decompresses a fixed-width code into its two or more components. The decoder is applied recursively until it terminates in symbols, reconstructing the tree. These two networks are simultaneously trained as an auto-associator with time-varying inputs. If the training is successful, the result of bottom up encoding will coincide with top down decoding.

Following the publication of (Pollack 1990), RAAM gained widespread popularity as a model of NL syntax. Some researchers (Blank, Meeden, and Marshall 1991) found it an attractive way of “closing the gap” between the symbolic and sub-symbolic paradigms in cognitive science. Others (Van Gelder 1990) saw in RAAM a direct and simple refutation of the traditional cognitive scientists’ backlash against connectionism, and went as far as to show how traditional syntactic operations like transformations could be performed directly on RAAM representations (Chalmers 1990). As the power of the RAAM model became apparent, variants began to emerge. These included the Sequential RAAMs of (Kwasny and Kalman 1995), which showed how a RAAM could behave like a linked list, and the Labeling RAAMs of (Sperduti 1993), which encoded labeled graphs containing cycles.

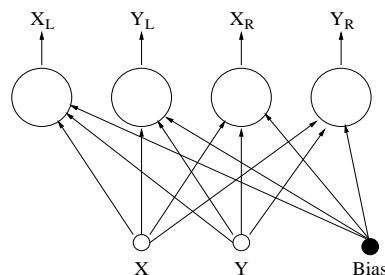
In short, RAAM seemed to hold a great deal of promise as a general connectionist solution to encoding not just NL syntax, but all sorts of structured representations.

Still, RAAM was plagued by an apparently diverse set of problems, most notably a failure to scale up to realistically large structures. We believe that these problems can be traced to the original formulation of the RAAM decoder, which works in conjunction with a logical “terminal test”, answering whether or not a given representation requires further decoding. The default terminal test merely asks if all elements in a given code are boolean, e.g. above 0.8 or below 0.2. This analog-to-binary conversion was a standard interface in back-propagation research of the late 1980’s to calculate binary functions from real-valued neurons. However, although it enabled the initial discovery of RAAM training, it led to several basic logical problems which prevented the scaling up of RAAM: 1) The “Infinite Loop” problem is that there are representations which “break” the decoder by never terminating. In other words, some trees appear “infinitely large” simply because their components never pass the terminal test. This behavior breaks computer program implementations or requires depth checking. 2) The “Precision vs. Capacity” problem is that tighter tolerances lead to more decoding er-

rors instead of a greater set of reliable representations. 3) The “Terminating Non-Terminal” problem arises when there is a “fusion” between a non-terminal and a terminal, such that the decoding of an encoded tree terminates abruptly.

In the following section of this paper we present a new formulation of RAAM networks based on an analysis of the iterated dynamics of decoding, that resolves all these problems completely. This formulation leads to a new “natural terminal test”, a natural labeling of terminals, and an inherently higher storage capacity.

### New RAAM Formulation



$$X_L = \frac{1}{1 + e^{-(w_{LX}x + w_{LY}y + w_{LB})}}$$

$$Y_L = \frac{1}{1 + e^{-(w_{LY}x + w_{LY}y + w_{LY})}}$$

$$X_R = \frac{1}{1 + e^{-(w_{RX}x + w_{RY}y + w_{RB})}}$$

$$Y_R = \frac{1}{1 + e^{-(w_{RY}x + w_{RY}y + w_{RY})}}$$

Figure 1: An example RAAM decoder that is a 4 neuron network, parameterized by 12 weights. Each application of the decoder converts an  $(X, Y)$  coordinate into two new coordinates.

Consider the RAAM decoder shown in figure 1. It consists of four neurons that each receive the same  $(X, Y)$  input. The output portion of the network is divided into a right and a left pair of neurons. In the operation of the decoder the output from each pair of neurons is recursively reapplied to the network. Using the RAAM interpretation, each such recursion implies a branching of a node of the binary tree represented by the decoder and initial starting point. However, this same network recurrence can also be evaluated in the context of dynamical systems. This network is a form of *iterated function system* or IFS (Barnsley 1993), consisting of two pseudo-contractive transforms which are iteratively applied to points in a two-dimensional space.

In the past we have examined the applicability of the IFS analogy to other interpretations of neural dynamics (Blair and Pollack 1997; Kolen 1994; Melnik and Pollack 1998; Stucki and Pollack 1992). But in the context of RAAMs the main interesting property of contractive IFSes lies in the trajectories of points in the space. For contractive IFSes the space is divided into two sets of points. The first set consists of points located on the underlying attractor (fractal attractor) of the IFS. The second set is the complement of the first, points

that are not on the attractor. The trajectories of points in this second set are characterized by a gravitation towards the attractor. Finite, multiple iterations of the transforms have the effect of bringing the points in this second set arbitrarily close to the attractor.

As noted before, the Infinite Loop and Terminating Nonterminal problems arise from an insufficient terminal test. Since some trajectories never leave the attractor and all others eventually hit the attractor. The only terminal test that guarantees the termination of all trajectories of the RAAM (IFS) is a test that includes all the points of the attractor itself.

By taking the terminal test of the decoder network to be “on the attractor”, not only are problems of infinite loops and early termination corrected, but it is now possible to have extremely large sets of trees represented in small fixed-dimensional neural codes. The attractor, being a fractal, can be generated at arbitrary resolution. In this interpretation, each possible tree, instead of being described by a single point, is now an *equivalence class* of initial points sharing the same tree-shaped trajectories to the fractal attractor. For this formulation, the set of trees generated and represented by a specific RAAM is a function of the weights, but is also governed by how the initial condition space is sampled, and by the resolution of the attractor construction. Note that the lower-resolution attractors contain all the points of their higher-dimensional counterparts (they cover them); therefore, as a coarser terminal set, they terminate trajectories earlier and so act to “prefix” the trees of the higher-dimensional attractors.

Two last pieces complete the new formulation. First, the encoder network, rather than being trained, is constructed directly as the mathematical inverse of the decoder. The terminal set of each leaf of a tree is run through the inverse left or right transforms, and then the resultant sets are intersected and any terminals subtracted. This process is continued from the bottom up until there is an empty set, or we find the set of initial conditions which encode the desired tree.

Second, using the attractor as a terminal test also allows a natural formulation of assigning labels to terminals. Barnsley (1993) noted that each point on the attractor is associated with an address which is simply the sequence of indices of the transforms used to arrive on that point from other points on the attractor. The address is essentially an infinite sequence of digits. Therefore to achieve a labeling for a specific alphabet we need only consider a sufficient number of significant digits from this address.

### Example of New RAAM Formulation

In this section, we describe how we obtain the attractor and the trees for a RAAM decoder of the sort shown in figure 1. The decoder weights in the present example were obtained by a hill-climbing search for an aesthetically appealing attractor, but the demonstration is valid for any set of decoder weights.

Recall that we are treating the decoder as an IFS that maps each input point  $(X, Y)$  in the range  $[0,1]$  to two other points  $(X_L, Y_L)$  and  $(X_R, Y_R)$  in the same range. To generate the attractor of the IFS, we first apply the two mappings (transforms) to the entire unit square at some fixed resolution. We then re-apply the transforms to the resulting set of points. We repeat this operation until the transforms do not change the

set of points any further at that resolution. Hence, we can visualize the behavior of the decoder in the unit square by examining the set of points obtained through iterated applications of the two transforms.

In figure 2, we have applied the transforms once to all points in the unit square, obtaining two large, overlapping regions, corresponding to the left and right transforms of all the original points. Note that some points are part of both the left and right regions.

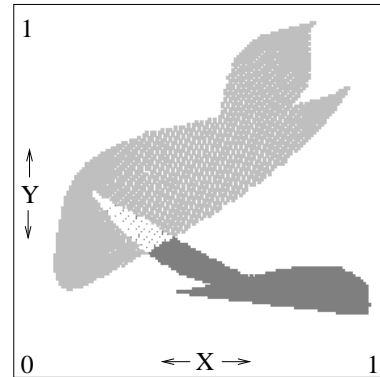


Figure 2: The unit square after one application of the transforms. The attractor is shown in gray: dark gray = points reachable from attractor on left transform, light gray = points reachable on right. The small white wedge where the gray areas overlap contains “ambiguous” attractor points reachable on both transforms.

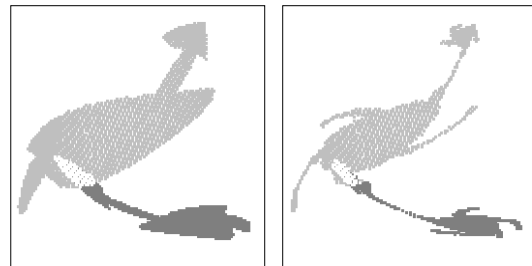


Figure 3: The unit square after two and five applications of the transforms.

Figure 3 shows the unit square after another iteration of the transforms, and after five such iterations. Figure 4 shows the final “Galaxy” attractor obtained when further iterations fail to produce any more contraction. Like any fractal, this attractor exhibits self-similarity, with the two longest arms of the galaxy ending in shapes like that of the whole attractor.

Figure 4 also shows how we derive the tree  $(1(12))$  from a point not on the attractor. Starting at a point not on the attractor (the small circle at the top of the figure), the left transform (dashed line) takes us immediately to the attractor; specifically, to an attractor region labeled 1, indicating that this region is reachable from the other attractor points on the left (first) transform only. Hence our tree so far is  $(1 \dots)$ . The *right* transform of the point at the top takes us to another point

not on the attractor, indicated by the circle in the lower left part of the figure. Like the first point, this point goes to the attractor region labeled 1 on its left transform; however, it also goes to the attractor on its right transform; specifically, to the region labeled 2, which indicates that this region is reachable from the other attractor points on the right (second) transform only. So this second point decodes the tree (1 2), and its parent tree is (1 (1 2)), completing the derivation.

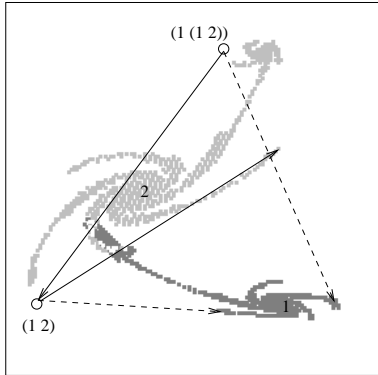


Figure 4: The final attractor, showing derivation of the tree (1 (1 2)) and its daughter tree (1 2). The left transform is shown as a dashed line, and the right transform as a straight line.

By repeating this process for every point not on the attractor, we can map out the set of all trees decoded by the RAAM at a given resolution. As described earlier, each tree in this set corresponds to an equivalence class of points that all decode to that tree. Points in the same class tend to cluster together, giving us an interesting way of laying out the RAAM’s language spatially. Figure 5 shows this phenomenon for a RAAM that we hill-climbed to decode the language  $a^n b^n$  (described in the next section), with grayscale denoting tree equivalence classes rather than attractor points. The dramatic striping pattern of the equivalence classes in this figure is not inherent in the fractal RAAM model, but derives from the comparatively elegant solution that hill-climbing produced for this language.

### Linguistic Advantages of New RAAM

As we described earlier, the new RAAM formulation thoroughly addresses the three shortcomings of the traditional RAAM model. Infinite loops and terminating non-terminals are both eliminated by making the terminal test be a test of whether or not a point is on the fractal attractor of the RAAM decoder.

Furthermore, the new formulation provides a principled account of generativity (grammatical competence). By treating the RAAM as a fractal that can be generated at any arbitrary resolution, we can increase the generative capacity of the RAAM without bound, giving us a model that scales perfectly: hence the name Infinite RAAM (IRAAM). As we have recently shown (?), it is a straightforward matter to hill-climb the weights for an IRAAM that generates all and only the strings in the language  $a^n b^n \cup a^n b^{n+1}, n \leq 5$ .

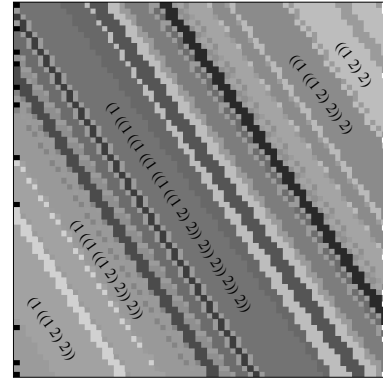


Figure 5: Tree equivalence classes for the  $a^n b^n$  system. Attractor points cluster at extreme left (colored black, labeled 1 or  $a$ ) and right (colored white, labeled 2 or  $b$ ).

Briefly, the dynamics of the network are such that for any point in the unit square, one of the two transforms of the point is guaranteed to be on the attractor. This behavior corresponds to the terminal component of a recursive grammar in Chomsky Normal Form for the language. In addition, the left transform of any point ends up on the left side of the unit square ( $x = 0$ ) and the right transform ends up on the right side ( $x = 1$ ). Hence, successive application of left/right/left... transforms leads to a zigzag dynamics that balances  $a$ ’s on the left with  $b$ ’s on the right, until a zig or zag lands on the attractor and terminates the oscillation. This behavior corresponds to the recursive component of the grammar. In (?), we provide a constructive proof for obtaining these behaviors at any resolution.

The proof gives us an exact IRAAM “competence” model for this non-regular CF language. Specifically, we show that there exists a set of weights for which a RAAM with an attractor generated at a predetermined resolution contains all and only the trees in the  $a^n b^n$  language. Performance limitations on the sizes of the trees actually produced derive from the resolution at which the non-attractor unit space is sampled, and not from an arbitrary stipulation or a breakdown of the model.

This infinite competence is not the only thing that IRAAM brings to connectionist NL modeling, however. Because IRAAM is a method of encoding and decoding *trees*, not just strings, its strong generative capacity is known. We can therefore use IRAAM as a direct model of hierarchical linguistic structure. An immediate implication of this result is that an IRAAM can be used as a parser and not just a recognizer. To the extent that real NL processing involves the assignment of meaning to strings based on structure, and not merely grammaticality judgments, this ability represents a significant advance in the application of connectionism to NL.

Finally, and perhaps most interesting, is the way in which IRAAM handles syntactic ambiguity. Consider the fractal addressing scheme that we described earlier. Each terminal point (word) on the attractor is associated with an address which is simply the sequence of indices of the transforms taken to arrive on the attractor point from other points on the attractor. Given  $K$  transforms, we would therefore assume

each digit in the sequence would fall in the range  $1, 2, \dots, K$ . For example, a binary-branching IRAAM, with two transforms, would have terminals with address digits 1 and 2. Using a one-digit address, this effectively puts each word into one of  $K$  “part of speech” equivalence classes.

This is not the whole story, though. Because there can be more than one path to a given terminal from some other terminal on the attractor, some terminals will have “ambiguous” addresses, containing digits out of the range  $1..K$ , to express the fact that more than one transform was taken to arrive at that point in the sequence. Continuing the linguistic analogy, this ambiguity corresponds to a given word’s belonging to more than one part of speech, as in Chomsky’s “flying planes” example, where *flying* can be either a verb or an adjective. For the binary-branching IRAAM example, if a given point had both a left and right inverse on the attractor, a one-digit address for that point would have to be a symbol other than 1 or 2. In general, for a  $K$ -ary IRAAM, there are  $2^K - 1$  possible one-digit addresses, consisting of  $K$  unambiguous values and  $2^K - K - 1$  ambiguous values.

This fact has great linguistic importance for IRAAM, for the following reason: typically (but not exclusively), an IRAAM decoder will favor putting the  $k$ th non-ambiguous terminal class in the  $k$ th position in a string of terminals, because the same set of weights is used to generate the attractor and the transients to the attractor. The likeliest non-terminal structure of a binary-branching IRAAM will therefore be (1 2), with structures (1 1), (2 1) and (2 2) being possible but less likely to occur. If, however, this IRAAM contains ambiguous terminals, it will very likely decode the structures (1 3), (3 2) and (3 3) as well.

Returning to the “flying planes” example, let us assign unambiguous verbs like *are* the category 1, unambiguous nouns like *planes*<sup>2</sup> the category 2, and the ambiguous *flying* the category 3. With this assignment, the natural ability of a binary IRAAM to decode the structures (1 (3 2)) and ((1 3) 2) gives us both parses of the expression *are flying planes*. Hence, we have an existence proof of a RAAM that can deal with syntactic ambiguity and non-deterministic grammars.

In short, we believe that IRAAM not only solves the problems of the earlier RAAM model, but also addresses the linguistic inadequacies of recurrent neural net models that we discussed earlier.

## What IRAAM Can’t Do

In the first section of this paper we outlined two linguistic criteria for a plausible NL model: the model should be able to handle “slightly” non-CF phenomena like copying and crossed serial dependencies and should also be incapable of handling CF phenomena absent from or deprecated in NL’s, like mirror-image constructions, or should incur a relatively high cost in producing or parsing those structures.

To investigate the latter point, we tested the ability of the IRAAM model shown in figure 1 to “learn” the context-free languages  $a^n b^n$  and  $ww^R$ ,  $w \in \{a, b\}$ . The training set consisted of the first 14 exemplars of each language (enumerated

<sup>2</sup>Readers troubled by the possibility of *planes* being a singular verb (*The carpenter planes the wood*) can substitute *cars* or some other unambiguous noun here.

in increasing order of length)<sup>3</sup>, with the fractal address 1 representing  $a$  and 2 representing  $b$ . Hill-climbing was used to learn the weights. Both the initial weights and the noise added to each weight came from a Gaussian distribution with zero mean and a standard deviation of 5.0, with the added noise’s standard deviation being scaled by the fraction of the training set missed. The resulting weights were used to generate trees on an IRAAM with a resolution of  $2^{-7}$ . The attractor was generated at that resolution and the initial starting point space was also sampled at that resolution.

Hill-climbing did not produce good results on either of these languages; the average success was six out of 14 strings covered for both languages. It is, however, instructive to look out *how* those successes were achieved. Comparing the best hill-climbed networks from each language (10 strings covered), we found that most of the strings generated by the  $a^n b^n$  network fit the general pattern of the training set: 74% of the strings fit the pattern  $a^n b^n$ . For the best  $ww^R$  network, however, only 14% fit the pattern  $ww^R$ . In other words, the  $a^n b^n$  network was actually producing mostly “grammatical” strings, whereas the  $ww^R$  network was essentially guessing.

We attribute these results to IRAAM’s aforementioned tendency to put symbols of one class ( $a$ ) on the left side of a branch and symbols of another class ( $b$ ) on the right side. In other words, trees of the form  $(a b)$ ,  $(a (a b))$ ,  $((a b) b)$ ,  $(a (a (a b)))$ ,  $((a b) b) b$ , are much more “natural” for an IRAAM than are trees of the form  $(a a)$ ,  $(b b)$ ,  $(b a)$ . But it is precisely the latter types of trees that are used as building blocks for the mirror-image language  $ww^R$ . This bias makes the mirror-image language much harder for an IRAAM to learn than the counting language  $a^n b^n$ , despite the fact that both are expressible by a simple CFG.

Although this result is by no means a proof of any sort, we consider it interesting for two reasons. First, it suggests that the languages generable by an IRAAM share an important formal property with NL, namely, the avoidance of mirror-image constructions. Second, the result illustrates how IRAAM imposes a constraint between the terminal symbol “semantics” and the nonterminal “syntax.” This constraint is absent from the definition of CFG’s (or of any grammar in the Chomsky hierarchy), where any terminal symbol can appear anywhere. To the extent that individual natural languages favor putting a given part of speech in fixed locations in a sentence or phrase (e.g., English generally has subject-verb-object, Japanese subject-object-verb), IRAAM appears to have an advantage over traditional grammars as a model of NL.

## Conclusion and Interpretations

We have demonstrated a new formulation of RAAM, which, by using a fractal attractor as a terminal test, enables the model to show competence and ambiguity, to represent a variety of tree structures, and not to represent deprecated mirror-image structures. We plan to relate this new formula-

<sup>3</sup>The number 14 was chosen because it allowed us to include all the members of  $ww^R$  for  $|w| \leq 3$ . This language has more strings of a given length than the language  $a^n b^n$ , which meant that the exemplars of the latter had to be longer in order to enumerate the first 14 of them. In effect, this makes the  $a^n b^n$  task *harder* than the  $ww^R$  task.

tion to linguistic formalisms like Tree-Adjoining Grammars (Joshi and Schabes 1997) and Categorical Grammars (Steedman 1999) having similar properties. We hypothesize that this relation may be achieved through the use of multiplicative connections to gate lexical varieties into naturally recursive dynamics.

Our work is by no means complete; nor do we mean to imply that NL grammar can be represented in four neurons with 12 weights! On the other hand, the principle of contractive maps and the emergence of fractal attractors in the limit behavior of nonlinear systems are mathematical facts, and have been used successfully in image-compression systems. Recent work by Tabor (1998) provides further evidence for the relevance of such principles to connectionist modeling of natural language. We now have reason to believe that these principles, under the right interpretation and scale, can support a neurally plausible universal grammar.

## References

- Bach, E., C. Brown, and W. Marslen-Wilson (1986). Crossed and nested dependencies in German and Dutch: A psycholinguistic study. *Language and Cognitive Processes* 1(4), 249–262.
- Barnsley, M. (1993). *Fractals everywhere*. New York: Academic Press.
- Blair, A. and J. Pollack (1997). Analysis of dynamical recognizers. *Neural Computation* 9(5), 1127–1142.
- Blank, D., L. Meeden, and J. Marshall (1991). Exploring the symbolic/subsymbolic continuum: A case study of raam. Technical Report TR332, Computer Science Department, University of Indiana.
- Bresnan, J., R. Kaplan, S. Peters, and A. Zaenen (1982). Cross-serial dependencies in Dutch. *Linguistic Inquiry* 13(4), 613–634.
- Chalmers, D. (1990). Syntactic transformations on distributed representations. *Connection Science* 2, 53–62.
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory* 2, 113–124.
- Chomsky, N. (1957). *Syntactic Structures*. Mouton.
- Culy, C. (1985). The complexity of the vocabulary of Bambara. *Linguistics and Philosophy* 8, 345–351.
- Fodor, J. and Z. Pylyshyn (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition* 28, 3–71.
- Gazdar, G. (1982). Phrase structure grammar. In P. Jacobson and G. Pullum (Eds.), *The Nature of Syntactic Representation*. Reidel.
- Higginbotham, J. (1984). English is not a context-free language. *Linguistic Inquiry* 15(2), 225–234.
- Joshi, A. and Y. Schabes (1997). Tree-adjoining grammars. In G. Rozenberg and A. Salomaa (Eds.), *Handbook of Formal Languages and Automata*, Chapter 3. Berlin: Springer Verlag.
- Kolen, J. (1994). *Exploring the Computational Capabilities of Recurrent Neural Networks*. Ph. D. thesis, Ohio State.
- Kwasny, S. and B. Kalman (1995). Tail-recursive distributed representations and simple recurrent neural networks. *Connection Science* 7(1), 61–80.
- Lawrence, S., C. Giles, and S. Fong (1998). Natural language grammatical inference with recurrent neural networks. *IEEE Transactions on Knowledge and Data Engineering*, to appear.
- Manaster-Ramer, A. (1986). Copying in natural languages, context-freeness, and queue grammars. In *Proceedings of the 24th meeting of the Association for Computational Linguistics*, pp. 85–89.
- Melnik, O. and J. Pollack (1998). A gradient descent method for a neural fractal memory. In *WCCI 98. International Joint Conference on Neural Networks: IEEE*.
- Pinker, S. and A. Prince (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28, 73–193.
- Pollack, J. (1990). Recursive distributed representations. *Artificial Intelligence* 36, 77–105.
- Postal, P. and D. Langendoen (1984). English and the class of context-free languages. *Computational Linguistics* 10(3–4), 177–181.
- Rodriguez, P., J. Wiles, and J. Elman (1999). A recurrent neural network that learns to count. *Connection Science* 11, 5–40.
- Rumelhart, D. and J. McClelland (1986). On learning the past tenses of English verbs. In D. Rumelhart and J. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 2. MIT.
- Shieber, S. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy* 8, 333–343.
- Sperduti, A. (1993). Labeling raam. Technical Report TR-93-029, International Computer Science Institute.
- Steedman, M. (1999). Categorical grammar. In R. Wilson and F. Keil (Eds.), *The MIT Encyclopedia of Cognitive Sciences*. MIT.
- Stucki, D. and J. Pollack (1992). Fractal (reconstructive analogue) memory. In *14th Annual Cognitive Science Conference*, pp. 118–123.
- Tabor, W. (1998). Dynamical automata. Technical Report TR98-1694, Computer Science Department, Cornell University.
- Van Gelder, T. (1990). Compositionality: a connectionist variation on a classical theme. *Cognitive Science* 14, 355–384.
- Williams, R. and D. Zipser (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation* 1, 270–280.

# The Acquisition of Lexical and Grammatical Aspect in a Self-Organizing Feature-Map Model

Ping Li

Ping@Cogsci.Richmond.Edu

Department of Psychology

University of Richmond

Richmond, VA 23173, USA

## Abstract

This study uses self-organizing feature maps to model the acquisition of lexical and grammatical aspect. Previous research has identified a strong association between lexical aspect and grammatical aspect in child language, on the basis of which some researchers proposed innate semantic categories (Bickerton, 1984) or prelinguistic semantic space (Slobin, 1985). Our simulations indicate that this association can be modeled by self-organization and Hebbian learning principles in a feature-map model, without making particular assumptions about the structure of innate knowledge. In line with results from Li (1999), our study further attests to the utility of self-organizing neural networks in the study of language acquisition.

## Introduction

Most linguistic theories of tense and aspect recognize two kinds of aspect: lexical aspect refers to the inherent temporal meanings of a verb, whereas grammatical aspect refers to a particular viewpoint toward the described situation. For example, whether the verb characterizes a situation as having a temporal boundary or an end result is a matter of lexical aspect, whereas whether the sentence presents a situation as ongoing (progressive/imperfective) or completed (perfective) is a matter of grammatical aspect. In English as well as in many other languages, lexical aspect is typically encoded by verb semantics, whereas grammatical aspect is encoded by morphological markers (e.g., English suffixes *-ing* and *-ed*).

Linguists have developed several systems to capture lexical aspect and grammatical aspect (see Comrie, 1976; Smith, 1997). For lexical aspect, the best-known system is Vendler's (1957) four-way classification of verbs into activities, accomplishments, achievements, and states: (1) activity verbs like *walk* and *run* encode situations as consisting of successive phases over time with no inherent endpoint; (2) accomplishment verbs like *build a house* also characterize situations as having successive phases, but differ from activities in that they encode an inherent endpoint (e.g., house-building has a terminal point and a result); (3) achievement verbs encode situations as punctual and instantaneous, e.g., *recognize a friend* and *cross the border*, and (4) state verbs encode situations as involving homogeneous states with no inherent endpoint, e.g., *know*, *want*, and *possess*. On the basis of whether the verb encodes endpoints, linguists also call activity and state verbs "atelic" (no endpoint), and accomplishment and achievement verbs "telic" (with endpoint). With respect to grammatical aspect, there are two major categories, according to Comrie (1976): imperfective

and perfective. Imperfective aspect presents a situation from an internal point of view, often as ongoing (progressive) or enduring (continuous), whereas perfective aspect presents a situation from an external perspective, often as completed. In English, the imperfective-perfective contrast is realized in the difference between the progressive *-ing* and the past-perfective *-ed*.<sup>1</sup>

Studies of language acquisition have long documented the interaction between lexical aspect and grammatical aspect in child language and in adult second language learning (for a comprehensive review, see Li & Shirai, 2000). In particular, researchers have found that young children initially tend to restrict tense-aspect morphology to specific categories of lexical aspect. This restricted or "undergeneralized" use is found in diverse languages such as Chinese, English, French, Italian, Japanese, and Turkish (see Li & Shirai, 2000 for a review). For example, English-speaking children tend to associate the use of the progressive marker *-ing* only with atelic, activity verbs, whereas they associate the past-perfective marker *-ed* only with telic verbs (accomplishments and achievements).<sup>2</sup> This strong association weakens over time, and eventually children develop adult-like competence in using both the progressive and the perfective suffixes with different lexical aspect categories.

Capitalizing on this strong association in early child language, some researchers hypothesized that children have innate semantic categories that roughly correspond to the lexical aspect distinctions of verbs. In particular, Bickerton (1984) argued that the semantic distinctions between *punctual* (e.g., *jump*) and *nonpunctual* (e.g., *walk*), and between *state* (e.g., *want*) and *process* (e.g., *walk*) are biologically programmed as part of a Language Bioprogram. Bickerton's initial claim for the proposed bioprogram was based on evidence from creole languages, but he also drew on the following evidence from early child language: (1) children treat achievement verbs (punctual) differently from activity verbs (nonpunctual) in their use of grammatical morphology; (2) children treat state verbs differently from activity (process) verbs, in that they use *-ing* only with process verbs and never with state verbs. These patterns prompted Bickerton

---

<sup>1</sup> Note that *-ed* marks both past tense and perfective aspect in English, just as *-s* marks both present tense and habitual aspect. Separate affixes are often used in other languages for tense and aspect.

<sup>2</sup> Some studies also report a third association between the habitual *-s* and state verbs, e.g., Clark (1996).

that children use tense-aspect morphology early on only to mark the bioprogrammed semantic distinctions. In a similar proposal with a somewhat different perspective, Slobin (1985) proposed that children come to the language acquisition task with a pre-structured semantic space in the Basic Child Grammar. This semantic space contains a universal, uniform set of prelinguistic semantic notions, initially independent of the child's linguistic experience, and they act like magnets to strongly attract the mapping of grammatical forms of the input language. Two contrasting categories, *process* and *result*, are in this space, and thus children would tend to map the progressive *-ing* to the process (atelic) verbs and the past-perfective *-ed* to the result (telic) verbs early on.

In this study, we entertain the same empirical results with an alternative proposal that rejects the strong version of the nativist argument on innate semantic categories.<sup>3</sup> In previous empirical studies (Li & Bowerman, 1998; Li & Shirai, 2000), we proposed that the initial lexical-morphological associations could arise as a result of the learner's analyses of the verb-morphology co-occurrence probabilities in the linguistic input, rather than innate biases. In parental speech, there are probabilistic associations between progressive markers and atelic verbs, and between perfective markers and telic verbs (see Li & Shirai 2000 for a review); children's initial undergeneralizations (restricted uses of morphology) might reflect their analyses of these probabilities. This study is a detailed implementation of this idea in a connectionist model. In previous connectionist work (Li, 1999), we explored the use of self-organizing neural networks, in particular, the self-organizing feature maps as a model of language acquisition. Our model was applied to overgeneralization and recovery phenomena in the acquisition of English reversible prefixes (*un-* and *dis-*), in connection with the acquisition of structured semantic representations (the cryptotypes of verbs). In this study, we extend this line of research to examine the undergeneralization of aspectual suffixes (*-ing*, *-s*, and *-ed*), in connection with the acquisition of semantic categories of lexical aspect. More important, we attempt to show (1) how a multiple feature-map model is able to capture the processes of semantic organization that leads to distinct lexical aspect categories that have been claimed to be innate or otherwise prelinguistic, and (2) how the model could derive child-like semantic-morphological associations on the basis of analyzing patterns in parental speech from the CHILDES database (MacWhinney, 1995). Evidence from such a study could shed light on the processes of lexical and morphological development in child language.

Several important properties of self-organizing feature maps make them particularly well suited to the study of lexical and morphological acquisition (see Li, 1999, for a discussion). First, they belong to the class of unsupervised learning networks that require no explicit teacher; learning is achieved entirely by the system's organization in response to the input. These networks provide computationally more

relevant models for language acquisition: one could argue that child language acquisition in the natural setting (especially the organization and reorganization of the lexicon) is largely a self-organizing process that proceeds without explicit teaching (MacWhinney, 1998). Second, self-organization in these networks allow for the gradual formation of structures as activity bubbles on 2-D maps, as a result of extracting an efficient representation of the complex statistical regularities from the high-dimension input space (Kohonen, 1989). This property allows us to model the emergence of semantic categories as a gradual process of lexical development. Self-organizing feature maps are also biologically plausible models: one could conceive of the human cerebral cortex as essentially a self-organizing map (or multiple maps) that compresses information on a 2-D space (Spitzer, 1999). Third, several self-organizing maps can be connected via *Hebbian learning*, a co-occurrence learning mechanism, according to which the associative strength between two neurons is increased if the neurons are both active at the same time (Hebb, 1949). In a multiple feature-map model (Miikkulainen, 1997), initially, all units on one map could be associated with all units on the other map; as self-organization takes place, the associations become more focused, so that eventually only the maximally active units on the two (or more) maps are associated. This procedure allows us to model one-to-many or many-to-many associations between forms and meanings on the basis of how often they co-occur and how strongly they co-activate in the representation. In short, self-organization and Hebbian learning are two important computational principles that aid us in the understanding of lexical representation and morphological generalization in language acquisition.

## Method

### Network Architecture

DISLEX is a multiple feature-map model of the lexicon that relies on self-organization and Hebbian learning principles (Miikkulainen, 1997). In this study, we use the basic architecture of DISLEX to model the acquisition of lexical and grammatical aspect. In this model, different feature maps dedicated to different types of linguistic information (orthography, phonology, or semantics) are connected through associative links via Hebbian learning. During learning, an input pattern activates a unit or a group of units on one of the input maps, and the resulting bubble of activity propagates through the associative links and causes an activity bubble to form in the other map. If the direction of the associative propagation is from phonology or orthography to semantics, *comprehension* is modeled; *production* is modeled if it goes from semantics to phonology or orthography. The activation of co-occurring lexical and semantic representations leads to continuous organization in these maps, and to adaptive formations of associative connections between the maps. Figure 1 presents a schematic diagram of the architecture of the model.

---

<sup>3</sup> Note that it is fundamental to the Language Bioprogram hypothesis that the semantic categories are biologically hard-wired, whereas this is left more open in the Basic Child Grammar hypothesis.



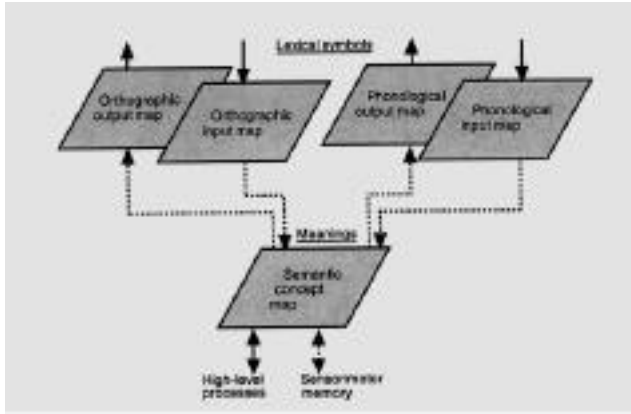


Figure 1: A multiple feature-map model of the lexicon (Miikkulainen, 1997)

In this study, we used no orthographic maps since we were modeling lexical and morphological acquisition in young children who are preliterate. We constructed two self-organizing maps, each of the size of 50 x 50 units, one for the organization of phonological input (henceforth the phonological map), and the other for the organization of semantic input (the semantic map). All simulations were run on a SUN Ultra workstation, using the DISLEX codes configured by Miikkulainen (1999).

### Input Representations

In order to model the role of linguistic input in children's acquisition of lexical and grammatical aspect, we selected as our input data the parental or caregivers' speech in the CHILDES database (MacWhinney, 1995). We extracted all the utterances of the parents, caregivers, and experimenters from the CHILDES database in over half of the English corpus (from Bates to Korman). Although not all of these utterances are child-directed, they form a representative sample of the speech that children are exposed to (e.g., dinner table talks, activities of free plays, and storytelling). A verb from this corpus was chosen as an input to the network if it occurred in the parental or caregivers' speech for five or more times in a given age period (see below). With this criterion we selected a total of 562 words (types) as input to our network. They were inputted to the network in four stages, according to the age groups at which they occurred (see below).

Previous connectionist models of language acquisition have often relied on the use of artificial input/output representations (e.g., randomly generated patterns of phonological or semantic representations) or representations that are constructed ad hoc by the modeler. Representations of linguistic information in this way are often subject to the criticism that the network works precisely because of the use of certain linguistic features. To overcome potential limitations associated with this approach, we used more realistic input data to simulate the acquisition of aspect. We represented our inputs as follows.

**Phonological representations** to our network were based on a syllabic template coding developed by MacWhinney and Leinbach (1991). Instead of a simple phonemic rep-

resentation, this representation reflects current autosegmental approaches to phonology, according to which the phonology of a word is made up by combinations of syllables in a metrical grid, and the slots in each grid made up by bundles of features that correspond to phonemes, *C*'s (consonants) and *V*'s (vowel). The MacWhinney-Leinbach model used 12 *C*-slots and 6 *V*-slots that allowed for representation of words up to three syllables. For example, the 18-slot template *CCC VV CCC VV CCC* represents a full trisyllabic structure in which each *CCCVV* is a syllable (the last *CCC* represents the consonant endings). Each *C* is represented by a set of 10 feature units, and each *V* by a set of 8 feature units.

**Semantic representations** to our network were based on the lexical co-occurrence analyses in the Hyperspace Analogue to Language (HAL) model of Burgess and Lund (1997). HAL represents word meanings through multiple lexical co-occurrence constraints in large text corpora. In this representation, the meaning of a word is determined by the word's global lexical co-occurrences in a high-dimensional space: a word is anchored with reference not only to other words immediately preceding or following it, but also to words that are further away from it in a variable co-occurrence window, with each slot (occurrence of a word) in the window acting as a constraint dimension to define the meaning of the target word. Thus, a word is represented as a vector that encodes the entire contextual history of that word in a high-dimensional space of language use. We used 100 dimensions for the encoding of each vector.

### Task and Procedure

Upon training of the network, a phonological input representation of the verb was inputted to the network, and simultaneously, the semantic representation of the same input was also presented to the network. By way of self-organization, the network formed an activity on the phonological map in response to the phonological input, and an activity on the semantic map in response to the semantic input. The phonological representations of the corresponding aspectual suffixes were also co-activated with the phonological and semantic representations of the verb, depending on whether the verb co-occurs with *-ing* *-ed*, or *-s* in the parental speech in the CHILDES database. As the network received input and continued to self-organize, it simultaneously formed associations through Hebbian learning between the two maps for all the active units that responded to the input. The network's task was to create new representations in the corresponding maps for all input words and to be able to map the semantic properties of a verb to its phonological shape and its morphological pattern.

To observe effects of the interaction between lexical and grammatical aspect in the parental input on the network's learning and representation, we designed four stages to train the network, according to the different age groups of our input data. (1) *Input Age 1;6* (13-18 months). Although parental/caregivers data in the CHILDES database are available from an age when the child is 6 months old, there are relatively few morphological markings prior to the period when the child is 12 months old. A total of 186 verbs fit

our selection criteria for the period when the child is between 13-18 months old, out of which 34 (18%) occurred with *-ing*, 9 (5%) with *-ed*, and 9 (5%) with *-s*. (2) *Input Age 2;0* (19-24 months). 324 verbs were selected, which include the new verbs as well as verbs from the previous stage, among which 76 (23%) occurred with *-ing*, 23 (7%) with *-ed*, and 24 (7%) with *-s*. (3) *Input Age 2;6* (25-30 months). 419 verbs were selected, among which 82 (20%) occurred with *-ing*, 35 (8%) with *-ed*, and 31 (7%) with *-s*. (4) *Input Age 3* (31-36 months). 562 verbs were selected, among which 123 (22%) occurred with *-ing*, 70 (12%) with *-ed*, and 61 (11%) with *-s*. These stages ensure an incremental growth of vocabulary and a coarse frequency coding: a verb or a suffix was presented to the network for the number of times it occurred across the four stages.

## Results and Discussion

We focus here on three levels of analysis of our modeling results: the role of input, the emergence of lexical aspect categories, and the formation and relaxation of strong associations between lexical and grammatical aspect.

### The Role of Input

One important rationale behind the current modeling effort is the understanding of the role of the linguistic input in guiding children’s acquisition of lexical and grammatical aspect. Earlier we emphasized the relationship between patterns observed in children’s speech and those in adult speech with respect to the interaction between lexical and grammatical aspect. But a simple correlation between children’s and adults’ patterns tells us only that the child is sensitive to the linguistic environment and is able to incorporate information from that environment into his or her own speech. It does not tell us how the child actually does this, or what mechanisms allow the child to do this. Thus, we wanted to test if a connectionist network, endowed with self-organization and Hebbian learning principles, is able to display learning patterns as the child does. If so, we can conclude that self-organization and Hebbian learning could provide the necessary kinds of mechanisms that drive the formation of patterns in children’s acquisition. In this way, our modeling enterprise provides insights into the mechanisms that underlie the learning process.

Table 1 presents a summary of the major patterns of the network’s learning according to the tense-aspect suffixes it produced at the different learning stages. It shows the results of the network’s production of three suffixes, *-ing*, *-ed*, and *-s* with three types of verbs, atelic, telic, and stative.<sup>4</sup> The results are based on the analyses of the activation of units on the phonological map that each verb in the semantic map activated, after the network had been trained for 200 epochs at each stage. The table does not include instances in which the network produced multiple suffixes with a given verb (see Table 3 for these instances).

The results as shown in Table 1 are highly consistent with empirical patterns observed in early child language: the use of the progressive aspect is closely associated with atelic

verbs that indicate ongoing processes, while that of perfective aspect is closely associated with telic verbs that indicate actions with endpoints or end results. In particular, in early child English, *-ing* is restricted to activity verbs, the perfective/past marker *-ed* restricted to telic verbs, and habitual *-s* restricted to stative verbs (see Introduction). Our network, having taken in input patterns based on realistic adult speech, behaved in the same way as children do. For example, at Input Age 1;6, the network produced *-ing* predominantly with atelic verbs (75%), *-ed* overwhelmingly with telic verbs (82%), and *-s* exclusively with stative verbs (100%). Such associations remained strong at Age 2, but gradually became weaker at later stages (the association between *-s* and stative verbs remained strong throughout).

Table 1: Network’s production of grammatical suffixes with lexical aspect categories\*

VERB	TENSE-ASPECT SUFFIXES					
	Age 1;6			Age 2;0		
	<i>-ing</i>	<i>-ed</i>	<i>-s</i>	<i>-ing</i>	<i>-ed</i>	<i>-s</i>
Atelic	75	18	0	66	16	0
Telic	25	82	0	28	84	0
Stative	0	0	100	0	0	100
VERB	Age 2;6			Age 3;0		
	<i>-ing</i>	<i>-ed</i>	<i>-s</i>	<i>-ing</i>	<i>-ed</i>	<i>-s</i>
	Atelic	64	26	0	52	9
Telic	31	74	0	44	77	10
Stative	0	0	100	4	14	80

\* Values represent percentages of verbs that occurred with the given suffix. Note that the percentages within a given column does not always add to 100%, reflecting the fact that some verbs could not be easily classified into one or the other category. This is also true for other tables.

Interestingly, when we analyzed the actual input to our network, we found similar patterns. Recall that the input to our network was based on the adult speech from the CHILDES database. Table 2 presents the percentages of use of suffixes with different verb types in the input data.

Table 2: Percentage of use of grammatical suffixes with lexical aspect categories in the input data

VERB	TENSE-ASPECT SUFFIXES					
	Age 1;6			Age 2;0		
	<i>-ing</i>	<i>-ed</i>	<i>-s</i>	<i>-ing</i>	<i>-ed</i>	<i>-s</i>
Atelic	69	22	0	74	15	17
Telic	28	77	33	24	77	0
Stative	3	0	67	2	8	83
VERB	Age 2;6			Age 3;0		
	<i>-ing</i>	<i>-ed</i>	<i>-s</i>	<i>-ing</i>	<i>-ed</i>	<i>-s</i>
	Atelic	67	23	20	67	23
Telic	25	69	20	31	65	8
Stative	8	8	60	2	12	69

This high degree of correlation between the network production and the input shows that our network was able to

<sup>4</sup> Atelic verbs correspond to Vendler’s activities, telic to accomplishments and achievements, and stative to state.

learn on the basis of the information of the co-occurrences between lexical aspect (verb types) and grammatical aspect (use of suffixes). This learning ability was due to the network's use of Hebbian learning in computing and registering (a) when the semantic, phonological, and morphological properties of a verb co-occur and (b) how often they do so.

Note that the patterns of the two tables are consistent and similar, but not identical. This is important because if the learner, child and network alike, simply mimicked what's in the input by recording each individual word and suffix and their co-occurrence, the learner would have no productive control of the relevant linguistic device and would simply produce the patterns verbatim. Our results suggest that the associations between verb types and suffixes are stronger in the network's production than in the input to the network. Our network, like the child, behaved more restrictively than what is in the input with respect to the use of tense-aspect suffixes (see Li & Shirai, 2000, for details on this point).

### The Emergence of Lexical Aspect Categories

As discussed earlier, a distinct property of feature maps is that the structures in the network's representation can be clearly visualized as activity bubbles or patterns of activity on a 2-D map. Figure 2 presents a snapshot of the network's self-organization of the semantic representations of 186 verbs at the end of the Input Age 1;6.

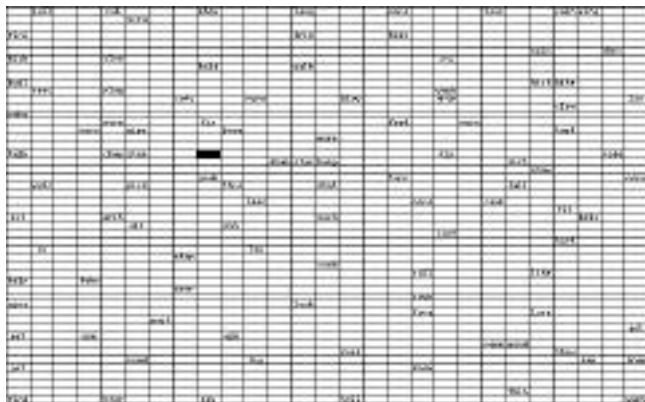


Figure 2: Network representations of verbs in the semantic map. Only the left portion of the complete map is shown due to space limit. Words longer than four letters are truncated.

An examination of this map shows that the network has clearly developed structured semantic representations that correspond to categories of lexical aspect such as telic verbs, atelic verbs, and stative verbs. Our network formed clear clusters of verbs by mapping similar verbs onto nearby regions of the map. For example, towards the lower right-hand corner, stative verbs like *feel*, *know*, *think*, *remember*, *wonder*, *love*, and *like* are mapped to the same region of the map. A second cluster of verbs occurs towards the lower left-hand corner, including verbs like *see*, *read*, *hear*, *say*, *ask*, and *tell*, all being verbs of visual or auditory activities. Still a third chain of verbs can be found in the middle-to-left portion of the map, including verbs like *catch*, *fix*, *break*, *knock*, *grab*, and *throw*, all of which are telic verbs indicat-

ing actions that lead to clear end results. Finally, a cluster of verbs can be found spanning the upper end of the map, including (from left to right) *rub*, *scrub*, *sleep*, *shout*, *laugh*, *drink*, *walk*, *kiss*, *cry*, *swim*, and *dance*, all of which are atelic activity verbs, and many of them co-occur with *-ing* early on. In contrast to this layer of verbs, the left-most columns feature primarily telic verbs, such as *finish*, *hide*, *build*, *reach*, *make*, *go*, *give*, *get*, and *find*. Of course, the network's representation at this point is still incomplete, as self-organization is still moving continuously from diffuse to more focused patterns of activity.

Crucially, on the one hand, these clusters form concentrated patterns of activity, providing the basis for semantic categories, and on the other hand, they also form focused associative pathways to the phonological and morphological representations of verbs on the other feature maps. When concentrated activities occur both horizontally (within a 2-D map) and vertically (across the maps), the semantic categories of lexical aspect will behave like magnets for the mapping of grammatical morphemes. Thus, when new verbs share enough similarities with verbs of a lexical aspect and fall within these clusters, their mapping to corresponding grammatical aspect will be readily assimilated through the existing associative pathways going from verb semantics to suffixes. This explanation provides an alternative account of the Basic Child Grammar, according to which the initial semantic categories that strongly attract grammatical mappings are privileged and pre-linguistic.

The results from our modeling offer a new way of thinking about the acquisition of categories of lexical aspect. Verbs in a lexical aspect category form complex relationships, in that they vary in (a) how many semantic features are relevant to the category, (b) how strongly each feature is activated in the representation of that category, and (c) how features overlap with each other across category members. Traditional analytical methods are much less effective, if not impossible, in dealing with these complex semantic relationships. By contrast, connectionist models that rely on distributed feature representations and nonlinear learning are ideally suited to accounting for the properties of featural overlapping and weighted featural composition (see Li & MacWhinney, 1996 for a discussion).

### From Strong Associations to Diverse Mappings

The above results suggest that the learning of grammatical suffixes is not simply the learning of a rule such as adding *-ing* or *-ed* to a verb to mark the progressive aspect or the perfective aspect, but the accumulation of associative strengths that hold between a particular suffix and a complex set of semantic features distributed across verb forms. This learning process can be best described as a statistical, probabilistic process in which the learner implicitly tallies and registers the frequency of co-occurrences (strengthening what goes with what) and co-occurrence constraints (inhibiting what does not go with what) among the semantic features, lexical forms, and tense-aspect suffixes.

This co-occurrence-and-constraint process is modeled in our network by Hebbian learning of the associative connections between forms and meanings. Hebbian learning can account for the relaxation of the associations as well as the

strong associations. Table 3 presents the same simulation results as Table 1, except here we added the multiple suffixation patterns -- a given verb was counted for multiple number of times in the table depending on the number of suffixes with which it co-occurred.

Table 3: Network's production of grammatical suffixes with lexical aspect categories (multiple suffixations)

VERB	TENSE-ASPECT SUFFIXES					
	Age 1;6			Age 2;0		
	<i>-ing</i>	<i>-ed</i>	<i>-s</i>	<i>-ing</i>	<i>-ed</i>	<i>-s</i>
Atelic	75	16	0	62	29	6
Telic	28	75	0	32	66	31
Stative	0	8	100	0	4	63
VERB	Age 2;6			Age 3;0		
	<i>-ing</i>	<i>-ed</i>	<i>-s</i>	<i>-ing</i>	<i>-ed</i>	<i>-s</i>
	Atelic	64	40	44	52	38
Telic	32	60	12	43	53	26
Stative	0	0	44	5	9	44

A comparison of this table with Table 1 reveals that for the early stages (1;6 and 2;0) the two tables are very similar; for the later stages, however, they become different, mainly with respect to the uses of *-ed* and *-s*. Detailed analyses show that over 50% of all suffixed verbs had more than one suffixes at Input Age 3;0, compared to only 5% at Input Age 1;6. These results suggest that multiple suffixation might be a driving force for the learner to break from the strong associations to more diverse associations between lexical and grammatical aspect. There was relatively little change with the *-ing* verbs, because the majority of the verbs early on were atelic verbs that took *-ing*. These same patterns were also found to be true of the parental input in the CHILDES database (see Li & Shirai, 2000, for detailed discussion).

## Conclusion

In this paper I showed that self-organizing neural networks can be used successfully to model language acquisition, following up on Li (1999). Self-organization and Hebbian learning in such networks are two important computational principles that can account for the psycholinguistic processes in the acquisition of lexical and grammatical aspect. Focused associative pathways between forms and meanings lead to particularly strong associations between lexical aspect and grammatical aspect, thereby to undergeneralized patterns of grammatical morphology as observed in early child language. Increasing associative links along with incremental vocabulary growth lead to diverse mappings. Finally, self-organization of the semantic structure of verbs leads to the formation of lexical aspect categories, on the basis of the network's analysis of the complex relationships in a high-dimensional space of language use. Our results lend insights into the mechanisms and processes of lexical-morphological acquisition, and may also generate interests in further empirical studies against which we can compare detailed patterns of modeling results.

## Acknowledgments

This research was supported by a grant from the National Science Foundation (#BCS-9975249). I am grateful to Brian MacWhinney and Risto Miikkulainen for their help, comments, and discussions at various stages of the project, and to Curt Burgess and Kevin Lund for making available their semantic vectors to our modeling.

## References

- Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and Brain Sciences*, 7, 173-188.
- Burgess, C. & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12, 1-34.
- Comrie, B. (1976). *Aspect: An introduction to the study of verbal aspect and related problems*. Cambridge, England: Cambridge University Press.
- Hebb, D. (1949). *The organization of behavior: A neuropsychological theory*. New York, NY: Wiley.
- Kohonen, T. (1989). *Self-organization and associative memory*. Heidelberg: Springer-Verlag.
- Li, P. (1999). Generalization, representation, and recovery in a self-organizing feature-map model of language acquisition. In M. Hahn & S.C. Stoness (eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp.308-313). Mahwah, NJ: Lawrence Erlbaum.
- Li, P., & Bowerman, M. (1998). The acquisition of lexical and grammatical aspect in Chinese. *First Language*, 18, 311-350.
- Li, P., & MacWhinney, B. (1996). Cryptotype, overgeneralization, and competition: A connectionist model of the learning of English reversible prefixes. *Connection Science*, 8, 3-30.
- Li, P., & Shirai, Y. (2000). *The acquisition of lexical and grammatical aspect*. Berlin and New York: Mouton de Gruyter.
- MacWhinney, B. (1995). *The CHILDES project: Tools for analyzing talk* (2nd Ed). Hillsdale, NJ: Erlbaum.
- MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, 49, 199-227.
- MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, 40, 121-157.
- Miikkulainen, R. (1997). Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language*, 59, 334-366.
- Slobin, D. (1985). Crosslinguistic evidence for the Language-Making Capacity. In D. Slobin (ed.), *The crosslinguistic study of language acquisition*. Vol.2. Hillsdale, NJ: Erlbaum.
- Smith, C. (1991). *The parameter of aspect*. Dordrecht: Kluwer.
- Spitzer, M. (1999). *The mind within the net*. MIT Press.
- Vendler, Z. (1957). Verbs and times. *Philosophical Review*, 66, 143-160.

# Irregularization: The Interaction of Item Frequency and Phonological Interference in Regular Past Tense Production

**Christopher J. Long (clong@usc.edu)**

Linguistics Department  
University of Southern California  
University Park, GFS 301  
Los Angeles, CA 90089-1693

**Amit Almor (almor@gizmo.usc.edu)**

University of Southern California  
Program in Neural, Informational, and Behavioral Sciences  
Hedco Neurosciences Building  
Los Angeles, CA 90089-2520

## Abstract

Both dual mechanism and connectionist single mechanism accounts predict that phonologically similar irregulars can interfere with regular past tense production. In dual mechanism accounts, such interference depends on irregular frequency but not on the frequency of regulars with the possible exception of high frequency regulars. Such models predict that high and low frequency regulars are equally susceptible to interference from irregulars, or that high frequency regulars would be more affected than low frequency ones. Connectionist single mechanism models, on the other hand, claim that low frequency regulars are more susceptible to interference from irregulars than high frequency regulars. We present results from two experiments that investigate interference from irregulars on the past tense production of high and low frequency regular verbs. In these experiments, high frequency regulars were less affected by interference from phonologically similar irregulars than low frequency regulars. These results support connectionist single mechanism models of past tense verb production.

The dichotomy between regular and irregular patterns permeates many levels of language. In many areas, accounting for differences between regular and irregular patterns has become the battlefield on which competing theories of language processing win or lose. One such case, which has received much attention in recent years, is the production of the past tense form of English verbs. This case has become the focus of an ongoing debate between dual mechanism and connectionist single mechanism models of word processing (MacWhinney B. & Leinbach, J., 1991; Marchman et al., 1999; Marcus, 1995a, 1995b, 1996; Pinker 1991, 1999; Pinker & Prince, 1991, Plunket & Marchman, 1993, 1996; Rumlehart & McClelland, 1986; Seidenberg 1997).

**Dual Mechanism Accounts.** In dual mechanism accounts (Marcus, 1995a, 1995b, 1996; Pinker 1991, 1999; Pinker & Prince, 1991), regular past tense is produced via a rule (add *-ed*) that applies to the root stem of the verb, which is stored in the mental lexicon. Irregulars, on the other hand, are

formed via associations between present and past tense forms, each of which is stored as a separate lexical entry. The add *-ed* rule applies as a default to verbs without a separate lexical past tense entry (regulars and non-words). The existence of a separate past tense entry, as in the case of irregulars, blocks the application of the add *-ed* rule. However, if the representation of a past tense entry is weak (due to low frequency), the add *-ed* rule could be applied erroneously, causing an over-regularization error (buy > \*bued). Such errors are well documented both under natural and experimental conditions (Berko, 1958; Bybee & Slobin, 1982; Marcus et al., 1992). In addition to irregulars being regularized, regulars can also be incorrectly produced as if they were irregulars (e.g., vie > \*vought). This phenomenon of "irregularization" has also been documented both in and out of the laboratory (Bybee and Moder, 1983; Xu & Pinker, 1995).

Pinker (1999), in his most recent version of a dual mechanism model, provides an account of irregularization. In his model, the word association mechanism responsible for irregular past tense production is a connectionist type neural network that contains both *irregular* and *high frequency regular* items. In the course of past tense production, the network attempts to compose a past tense form on the basis of the present tense. Phonological overlap between regular and irregular items can cause two types of interference during this computation. First, an incorrect form may gain enough activation to actually block the add *-ed* rule resulting in an irregularization error. Alternatively, regular past tense production could be successful, but the spurious activation caused by interference may slow down production.

By this account, interference (and past tense production) is generally insensitive to the frequency of regular verbs. One exception may result from the encoding of regular past tense forms of high frequency regular verbs in the associative network. The encoding of these past tense forms in the network can sometimes cause high frequency regulars to be more affected by interference from irregulars than low frequency regulars:

"As mentioned in note 11, sometimes high-frequency regular verbs are, paradoxically, *slower* to produce than low-frequency verbs. One explanation is that stored forms always inhibit the rule, even if they are identical to the form the rule is trying to create. Just as *broke* blocks the creation of *breaked* an entry for *walked* that is stored in memory may block the creation of *walked* by rule, slowing down the rule production (compared to, say, *stalked*, whose memory entry is too weak to slow down the rule)."

(Pinker, 1999, page 303, fn 22)

**Connectionist single mechanism Accounts.** Connectionist single mechanism models rely on a single mechanism to account for both regular and irregular past tense production (MacWhinney, B. & Leinbach, J., 1991; Marchman, Wulfeck, & Weismer, 1999; Plunket & Marchman, 1993, 1996; Rumlehart & McClelland, 1986; Seidenberg 1997). The claim is that both regulars and irregulars are represented in a single neural network. The network encodes mappings between present and past tenses as weighted links between forms. More exposure to a particular mapping strengthens the corresponding link. In this way, the strength of mappings for both regular and irregular items is determined by item frequency and the consistency of the present to past tense mapping within the neighborhood of phonologically similar verbs.

In this model, interference on regular and irregular past production is the result of a single mechanism. As activation spreads and a past tense form (regular or irregular) is being selected, interference from overlapping mappings (regular or irregular) can cause interference. In past tense production, a target form must reach a critical activation level before it is selected. In order for this to happen, activation of competing forms must be suppressed. If a competing form is not suppressed and its activation exceeds that of the correct form and reaches a critical threshold, an irregularization error occurs. However, even if correct selection is eventually successful, the activation of competing forms may cause the system to take longer to settle on the correct form and may thus result in slowed production. In this account, due to the greater strength of their mappings, high frequency regular and irregular verbs are less susceptible to interference from other items than low frequency verbs. Thus, the connectionist single mechanism account and the dual mechanism account both predict the same two types of possible interference: (1) irregularization errors, and (2) slowing of correct production. However, the two accounts differ in their prediction of how interference will affect high and low frequency regulars. The dual mechanism account predicts that interference should result in more errors and slower production for high frequency regular verbs than for low frequency regular verbs. The connectionist single mechanism account, on the other hand, predicts more errors and slowed production for the low frequency regulars than for the high frequency regulars. The following experiments tested these contrasting predictions.

## Experiment 1

This experiment investigated the effects of interference from irregulars on regular past tense production. More specifically, we tested the degree to which high and low frequency regulars are differentially affected by such interference.

One potential problem facing an investigation of this type is the subtle nature of interference effects from irregulars. In order to get around this problem, we designed the experiment so as to enhance the interfering effects of irregulars. According to both models, phonological overlap can yield irregularization errors and slowed production. We reasoned that one possible way to enhance the effect of phonological similarity is to make this similarity more salient by having participants produce the past tense form of an irregular verb (e.g., *buy*) immediately prior to producing the past tense form of a phonologically similar regular verb (e.g., *die*). This was done for both high and low frequency regulars creating the two interference conditions shown in Table 1. We decided on an all auditory presentation of stimuli and responses so as to further enhance the (possibly interfering) effect of phonological similarity. We also had a control condition in which the same regular verbs were preceded by non-similar irregulars, creating the two control conditions shown in Table 1. Thus, the complete design was 2 X 2 with factors Frequency (high vs. low) and Context (interference vs. control).

**Table 1:** Experiment 1 Conditions. Shown is the regular target (**in bold**) with the irregular from the preceding trial.

Frequency	Context	
	Interference	Control
High	Buy	Hear
	<b>Die</b>	<b>Die</b>
Low	Buy	Hear
	<b>Vie</b>	<b>Vie</b>

Both models predict that the interference condition will be slower and more error prone than the control for both high and low frequency regulars. However, the two models differ in their predictions for the interaction between context and frequency. The dual mechanism model claims that the difference between the interference and control conditions will either be equal for both high and low frequency regulars, or perhaps be greater for high than low frequency regulars. According to Pinker (1999), high frequency regulars should be more susceptible to interference when the experimental list includes a high proportion of irregular verbs, as in the present experiment (50%):

"... the harmful effects of high frequency tend to occur when the word list has a high percentage of irregular forms, encouraging subjects to go to their mental lexicons on every trial..."

(Pinker, 1999, pp. 303, f. 22)

The connectionist single mechanism account, on the other hand, predicts that the difference between the interference and control conditions will be greater for the low frequency regulars than for the high frequency regulars.

## Method

**Participants.** 61 undergraduate students from the Department of Psychology at the University of Southern California received extra credit to participate in the experiment. All were native speakers of English.

**Materials.** 20 Monosyllabic English irregular present tense verbs were matched with 20 phonologically similar monosyllabic high frequency (Mean frequency 185) and 20 phonologically similar monosyllabic low frequency (Mean frequency 3) present tense regular verbs creating 20 *high interference pairs* and 20 *low interference pairs*. The same set of items was then regrouped such that each irregular verb was matched with a non-similar sounding high and low frequency regular verb creating 20 *high control pairs* and 20 *low control pairs*. Verb frequencies were taken from the Francis & Kucera (1982) corpus.

To ensure that each participant responded to each regular item only once, the prime and control pairs were divided into four lists (1A/1B, 2A/2B) each containing one-quarter of the experimental pairs with the number, type, and frequency of pairs balanced across lists. 23 regular and 44 irregular monosyllabic present tense verbs were selected as fillers to balance the appearance of regular and irregular items on the four lists. The lists were ordered in a pseudo-random fashion. Presentation of lists was balanced across participants.

A practice list was also created consisting of 10 regular and 10 irregular present tense verbs that were not included in the experimental lists.

All items were read by a male native speaker of English and digitally recorded in 16 bit 20 MHZ format. Individual words were later excised using a digital sound manipulation program.

**Procedure.** Stimuli were presented to participants through headphones at 2000 ms intervals using the PsyScope program (Cohen, J., MacWhinney, B., Flatt, M., & Provost, J., 1993). Participants were instructed to say the past tense form of the verb they heard. In order to encourage rapid responses and reduce possible strategy effects, participants were instructed to say the past tense as quickly as possible. If participants didn't answer within a 1500ms interval, they were signaled with a beep. Responses were coded as either 'correct', 'incorrect' or 'equipment error'. All participants received the same 20-item practice list before being tested on one of the two A/B list sets. Order of presentation of the lists was alternated giving a total of four potential presentation patterns (1a>1b, 1b>1a, 2a>2b, 2b>2a) that were counterbalanced across participants. There was a short break between the lists. The total testing time was approximately 30 min. Participant responses were recorded.

Recordings were used to verify initial coding of responses and to transcribe and code responses for the error analysis.

## Results

Data were included only for responses to regular targets that followed the correct production of the preceding irregular item. Responses to regulars that followed an incorrect irregular were not included because in such cases it is not clear whether participants had processed the preceding irregular. Six participants (4% of data) and three items (2% of data) had to be removed from the analyses due to insufficient data contribution. Initial analyses of list order effects indicated that there was no interaction between list presentation order and any of the conditions of the experiment. Thus, the data from initial and second presentations were collapsed.

**Error analysis.** Responses were classified as follows:

- (1) *Correct*: the regular past tense form was correctly produced.
- (2) *Irregularization Error*: the regular past tense form was incorrect and the form of the error had a direct relationship to an existing past tense irregular form. (e.g., the past tense of 'vie' produced as 'vought').
- (3) *Miscellaneous Error*: the regular past tense form was incorrect and the form of the error did not relate to an existing irregular past tense form (e.g., vie > died).

Table 3 shows the distribution (raw numbers and percentages) of response types in the different conditions.

**Table 3:** Response types in Experiment 1

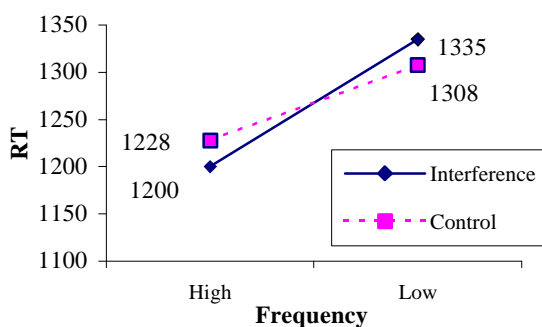
Response Type	Context	
	Interference	Control
	High	
Correct	344 (95%)	349 (96%)
Irregularization Error	7 (2%)	3 (1%)
Misc. Error	13 (4%)	10 (3%)
	Low	
Correct	256 (88%)	273 (91%)
Irregularization Error	13 (4%)	4 (1%)
Misc. Error	20 (7%)	22 (7%)
	Total	
Correct	600 (92%)	622 (94%)
Irregularization Error	20 (3%)	7 (1%)
Misc. Error	33 (5%)	32 (5%)

To assess the effect of frequency and context on response type we conducted a log-linear analysis on these responses starting with the maximal model – Context (interference vs. control) x Frequency (high vs. low) x Response Type (correct vs. irregularization error vs. miscellaneous error). The only terms that are of potential interest here are the interaction terms that included the response type and that were crucial for the model's fit. The

terms that met these criteria in this experiment were the term expressing the interaction between context and response type (LR  $\chi^2(2)=7.03$ ,  $p<.0297$ ) and the term expressing the interaction between frequency and response type (LR  $\chi^2(2)=15.17$ ,  $p<.0005$ ). The three-way interaction between context, frequency and response type was not significant (LR  $\chi^2(2)=.7698$ ). Thus, while both context and frequency had an effect on response type, these effects were independent of each other. Participants made more errors with the low frequency verbs than with the high frequency verbs but this frequency effect was comparable in the interference and control conditions. To better examine the distribution of the different response types in the interference and control conditions, we combined the responses from the high and low frequency conditions (as in the bottom part of Table 3). While in both prime and control conditions there is an equal percentage of miscellaneous errors (5%), the prime condition has more irregularization errors (5%) than the control condition (3%). This difference proved to be significant according to a chi-square test including correct responses in the analysis ( $\chi^2(1)=6.47$ ,  $p<.0394$ ), as well as an analysis of the error data alone ( $\chi^2(1)=6.62$ ,  $p<.0364$ ).

**Latencies.** Figure 1 shows RTs for correct responses in all four conditions. While low frequency regulars were produced slower in the interference than in the control condition, high frequency regulars were actually produced faster in the interference condition. An ANOVA with factors Frequency (high vs. low) and Context (interference vs. control) revealed a main effect of frequency whereby participants took significantly longer to produce the past tense form of low frequency regular verbs compared to high frequency ones (1322 ms vs.1214 ms),  $F_1(1, 54)=111$ ,  $p<.001$ ,  $F_2(1, 35)=7.244$ ,  $p<.01$ . Context had no main effect,  $F_1, F_2<1$ . The interaction between context and frequency was significant by participants,  $F_1(1, 54)=4.095$ ,  $p<.048$ , although not by items,  $F_2<1$ .

Figure 1: RT Experiment 1



## Discussion

Consistent with the predictions of both the dual mechanism and the connectionist single mechanism accounts, interference from a similar sounding irregular was found to

increase the likelihood of making an irregularization error in producing the past tense form of regular verbs. However, the error data show that, in contrast to the claims of Pinker's (1999) dual mechanism account, the production of low frequency regulars is overall more prone to errors than the production of high frequency regulars. Nevertheless, the error data did not provide strong support for the connectionist single mechanism account because high and low frequency regulars were affected equally in both the control and interference condition. This failure to detect a significant effect may be a reflection of a true lack of interaction, as the dual mechanism model may predict, or it may simply be a result of a lack of power due to the subtlety of the effect and low cell count (participants produced very few errors overall.) A more informative measure of performance that is not prone to the small cell size problem and its associated low power was provided by response latencies in correct regular past tense production.

The analysis of response latencies revealed, as in the error data analysis, a general advantage for high frequency regulars such that they were faster than low frequency regulars. Importantly, this finding precludes a speed accuracy trade-off explanation of the error data. There are two aspects of these results, however, that need to be dealt with before any further interpretation of the latency data can be made. First, the past tense form of high frequency verbs was produced faster in the interference condition than in the control condition, in contrast to the predictions of both models. Second, the fact that the interaction between frequency and context was significant only in the by-participants analysis but not in the by-items analysis suggests that items varied in some important aspect that we may have overlooked. One such aspect may be related to priming effects between the present tense forms, independent of the production of the past tense. To perform the task, participants had to, first, process the present tense form of each verb, and, second, generate the past tense form. Thus, response times in this task indicate not only the time it took participants to generate past tense forms but also the time it took them to process the present tense forms. It may be that phonological similarity, which caused interference in the production of the past tense form, facilitated the processing of the present tense form. The high frequency regulars may have thus elicited faster responses in the interference condition because, for these verbs, phonological similarity benefited the processing of the present tense forms more than it interfered with the production of the past tense forms. For the low frequency regulars, on the other hand, phonological similarity may have interfered with the production of the past tense forms more than it benefited the processing of the present tense forms. Differences between items in the relative strengths of the benefit and the interference associated with phonological similarity may also help explain the differences between the by-participants and by-items analyses. It is important to note here that this interpretation only applies for the specific irregular/regular pairs as used in



this experimental manipulation and not for regular verbs in general. In order to test this interpretation, it is necessary to confirm the facilitatory effects of phonological similarity on the processing of the present tense in the interference condition and then reanalyze the data taking these effects into account. Experiment 2 was undertaken to directly measure the effect of phonological similarity on the processing of the present forms.

## Experiment 2

This experiment was identical to Experiment 1 in materials and procedure but employed a repetition task instead of past tense generation. Thus, in Experiment 2, the interference condition of Experiment 1 became a priming condition in which the preceding phonologically similar irregular could prime the recognition of the regular target item.

### Method

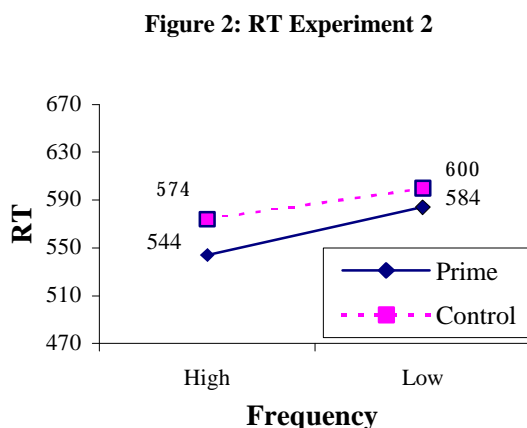
**Participants.** 25 undergraduate students (different from Experiment 1) from the Department of Psychology at the University of Southern California received extra credit to participate in the experiment. All were native speakers of English

**Materials.** Same as in Experiment 1.

**Procedure.** Same as in Experiment 1 except that participants were instructed to repeat the verb that they heard rather than produce the past tense. The presentation interval was reduced to 650 ms and the time-out interval to 600 ms.

### Results

As with Experiment 1, only responses to regular verbs that followed the correct repetition of the preceding irregular verb were included. 3 subjects (5% of data) and 4 items (7% of data) were excluded from the analysis due to insufficient data contributions. Mean RTs are shown in Figure 2. An ANOVA with factors Frequency (high vs. low) and Context (prime vs. control) revealed a main effect



of frequency whereby high frequency verbs were repeated faster than low frequency verbs (558 ms vs. 591 ms),  $F_1(1, 21)=39.59$ ,  $p<.0001$ ,  $F_2(1, 34)=3.923$ ,  $p<0.056$ . There was also a main effect of context whereby participants were faster at repeating target items in the prime condition than in the control condition, (564 ms vs. 592 ms),  $F_1(1, 21)=7.062$ ,  $p<.015$ ,  $F_2(1, 34)=6.062$ ,  $p<0.019$ . There was no interaction between context and frequency,  $F_1, F_2<1$ .

### Discussion

The fact that both high and low frequency regulars were repeated faster in the prime condition confirms the claim that the interference condition of Experiment 1 also involved facilitation of the initial processing of the present tense. Furthermore, the lack of context by frequency interaction in Experiment 2 suggests that the context by frequency interaction observed in Experiment 1 was not related to the processing of the present tenses but was only related to the generation of the past tense. Most importantly, however, the results of Experiment 2 can be used to reanalyze the results of Experiment 1 while factoring out the priming effects related to the processing of the present forms.

**Combined Experiment 1 and 2 Analyses.** One possible way to factor out the effects of present tense priming is by repeating the by-items analysis of Experiment 1 with Experiment 2 response times as covariates. Due to differences in the magnitude and variability of response times in the two experiments, response times were log transformed (Emerson, 1991). An ANCOVA of the log transformed RTs in Experiment 1, with factors Frequency (high vs. low), Context (interference vs. control), and covariates Prime and Control Log RTs from Experiment 2, found no main effect for context,  $F<1$ , a marginally significant main effect of frequency,  $F(1,32)=3.391$ ,  $p<.075$ , and finally, a significant interaction between context and frequency,  $F(1, 32)=4.176$ ,  $p<.049$ . Thus, a by-items analysis in which the processing of present tenses was controlled for, found, as did the original by-participants analysis, a significant interaction between context and frequency.

To further explore the nature of this interaction we calculated the partial correlation between the interference on each item in Experiment 1 (RT in interference condition minus RT in control condition) and their frequency (log transformed) while controlling for the item's priming in Experiment 2 (RT in prime condition minus RT in control condition). This analysis found that item frequency and the extent of interference for that item were negatively correlated ( $r= -0.35$ ,  $p<.037$ ) such that the higher the item frequency, the less was the effect of interference.

### Discussion of Combined Analyses

The reanalysis of Experiment 1 showed that once the effects of processing the present tense were factored out, the item analysis corroborated the participant analysis in showing a

significant interaction between interference and frequency. Importantly, this analysis revealed that low frequency regulars were affected by interference more than high frequency regulars. This finding is squarely incompatible with Pinker's (1999) dual mechanism account.

### General Discussion and Conclusions

Two production experiments tested the extent to which irregular verbs could interfere with the past tense production of regular verbs. Irregularization errors were more likely when regular verbs were preceded by phonologically similar irregulars than when they were preceded by phonologically dissimilar irregulars. Furthermore, overall production errors were more likely for low frequency regular verbs than for high frequency regular verbs. Finally, an analysis of latencies of correct responses showed that, once the effects of processing the present tenses are controlled for, high frequency regulars are more immune than low frequency regulars to interference from phonologically similar irregulars.

These findings are incompatible with the prediction of Pinker's *Words and Rules* dual mechanism model (Pinker, 1999) that high frequency regulars should be affected by interference more than low frequency regulars. One obvious way in which Pinker's model could be modified to account for our findings is by simply changing it to say that regular past tense production could benefit (rather than be hindered) by the existence of a form in the associative network. While such modification may help account for the current findings, it is not clear what its other consequences may be.

Clearly, frequency effects are not the only relevant evidence for understanding the mechanisms underlying past tense production. Thus, the present findings should not be viewed as the ultimate proof that the connectionist approach is right and that the dual mechanism account is wrong. Rather, the present findings should be viewed as adding one piece to a growing body of evidence that suggests that the separation of language processing into two mechanisms is buying less and less in terms of explanatory power but costing more and more in terms of unnecessary theoretical baggage.

### Acknowledgements

This research was supported by USC graduate merit fellowship (C.J.L) and NIA grant AG11773-05 (A.A.). We are grateful to Mark Seidenberg and Elaine Andersen for helpful feedback at various stages of this work.

### References

Berko, J. (1958). The Child's learning of English morphology. *Word*, 14, 150-177.

Bybee, J. L. & Moder, C. L. (1983). Morphological classes as natural categories. *Language*, 59, 251-270.

- Bybee, J. L. & Slobin, D. I. (1983). Rules and schemes in the development and use of the English past tense. *Language*, 58, 265-289.
- Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). *PsyScope*: Software developed at Carnegie Mellon University, Department of Psychology.
- Emerson, J. D. (1991). Introduction to Transformations. In D. C. Hoaglin, F. Mosteller, & J. W. Turkey (Eds.), *Fundamentals of Exploratory Analysis of Variance*. New York: Wiley.
- Francis, W. N., & Kucera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton-Mifflin.
- MacWhinney, B. and Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, 40, 121-157.
- Marchman, V. A., Wulfeck, B., & Weismer, S. E. (1999). Morphological productivity in children with normal language and SLI: A study of the English past tense. *Journal of Speech, Language, and Hearing Research*, 42(1), 206-219.
- Marcus, G. F. (1995a). The acquisition of the English past tense in children in multilayered connectionist networks. *Cognition*, 56, 271-279.
- Marcus, G. F. (1995b). Children's overregularization of English plurals: a quantitative analysis. *Child Language*, 22, 447-459.
- Marcus, G. F. (1996). Why do children say "brokeed". *Current directions in Psychological Science*, Vol. 5(3), 81-85.
- Marcus, G., Ullman, M., Pinker, S., Hollander, M., Rosen, T., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 57.
- Pinker, S. (1991). Rules of language. *Science*, 253, 530-535
- Pinker, S. (1999). *Words and Rules*. New York, NY: Basic Books.
- Pinker, S. & Prince, A. (1991). Regular and irregular morphology and the psychological status of rules and grammar. Proceedings of the 1991 Meeting of the Berkeley Linguistics Society, USA.
- Plunkett, K. & Marchman, V. A. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, 48, 21-69.
- Plunkett, K. & Marchman, V. A. (1996). Learning in a connectionist model of the acquisition of the English past tense. *Cognition*, 61, 299-308.
- Rumelhart, D. & J. McClelland (1986). 'On the learning of past tense of English verbs. Implicit rules or parallel distributed processing?' In J. McClelland, D. Rumelhart and the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.
- Seidenberg, M.S. (1997). Rules of language. *Science*, 275, 1599-1603.
- Xu, F. & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, 22, 531-556.

# A Computational Level Theory of Similarity

Bradley C. Love  
Department of Psychology  
The University of Texas at Austin  
MEZ 330  
Austin, TX 78712 USA  
*love@psy.utexas.edu*

## Abstract

Why are some pairs of objects (or events) perceived to be more similar to each other than other pairs? A computational level theory of perceived similarity is presented that extends previous geometric and set-theoretic formulations. Like previous approaches, the current account posits that the similarity of two objects is a function of the common and distinctive features of the two objects. Unlike previous approaches, similarity is also a function of higher-order compatibility relations among features (as it is in models of analogy). Objects (or concepts) are represented as directed feature graphs as opposed to feature vectors or sets. Like current accounts of human analogical processing, the approach presented here holds that representational elements are put into correspondence during the comparison processes. Correspondences are chosen in order to maximize an objective function. The function contains four terms that are motivated by theories of human comparison. The maximum of the function is monotonically related to perceived similarity. Thus, similarity is characterized as the byproduct of comparison and structural alignment. The objective function serves as a quantitative computational level theory of human comparison.

## Introduction

Since William James (1890/1950), psychologists have held that detecting the “sameness” or similarity of objects is at the backbone of cognition. Clearly, detecting similarities between novel events and previous experiences is crucial in reasoning, analogy, and object recognition. Many theories of category learning hold that similarity is the basis for categorization (see [7]). A fundamental question then is what makes two objects similar?

Almost all accounts of perceived similarity hold that similarity increases as the number of feature matches increases and decreases as the number of feature mismatches increases. In geometric models of similarity, such as multidimensional scaling (MDS) models of similarity, concepts or objects are represented as points in a multidimensional space and similarity is inversely related to the distance between points in the space [20]. Objects that match on many features will be closer together in the space than objects that mismatch on a number of features. Unfortunately, the axioms of metric spaces (e.g., minimality, symmetry, and the triangle inequality) appear to be violated by human similarity judgments (see [22]). More recently, Medin, Goldstone, and Gentner (1993) have demonstrated that an object can be rated as both more similar and more dissimilar to the same object in an object pair, which seems problematic for distance models.

Tversky’s (1977) contrast model is a non-metric set-theoretic account of perceived similarity that aims to address some of the shortcomings of the distance models. Tversky’s model is based on evaluating sets of matching and mismatching features:

$$\mathbf{sim}(x, y) = \gamma_1 \mathbf{F}(X \cap Y) - \gamma_2 \mathbf{F}(X - Y) - \gamma_3 \mathbf{F}(Y - X) \quad (1)$$

where  $\gamma_1, \gamma_2, \gamma_3 \geq 0$

where  $\mathbf{sim}(x, y)$  reads “the similarity of  $x$  to  $y$ ,”  $X$  is the set of features that represents  $x$ ,  $Y$  is the set of features that represents  $y$ ,  $X \cap Y$  is the set of features common to  $x$  and  $y$ ,  $X - Y$  is the set of features uniquely possessed by  $x$ ,  $Y - X$  is the set of features uniquely possessed by  $y$ ,  $\gamma_1$ ,  $\gamma_2$ , and  $\gamma_3$  are free parameters, and  $\mathbf{F}$  is a function over sets of features related to the features’ saliency. For simplicity and without loss of generalization, we assume here that all features are equally salient:

$$\mathbf{F}(X) = |X| \quad (2)$$

where  $|X|$  denotes the cardinality of the set  $X$ . Of course, in many situations certain features will weigh more heavily on the evaluation of similarity than other features.

Tversky’s contrast model can account for asymmetries that occur in similarity judgments. For example, “North Korea” is rated as being more similar to “China” than vice versa. The contrast model can explain such asymmetries by setting  $\gamma_2 > \gamma_3$ . Ostensibly, when comparing  $x$  and  $y$ , the focus is on first term  $x$ , which I will refer to as the *target*, and not on the second term, which I will refer to as the *base*. Both  $x$  and  $y$  will be referred to as *analogs*. In the example above, most people know more about China than North Korea. Accordingly, when evaluating how similar China is to North Korea  $|X - Y|$  will be larger than  $|Y - X|$ . Another comparison related explanation for asymmetries is that subjects prefer the base to be the object out of an object pair that allows for more analogical inferences to be projected [1]. Such asymmetries may be attributable to the similarity predicate in particular. Another alternative is that asymmetries arise from general principles related to sentence interpretation such as the figure/ground relationship between the target and base [21] or from general syntactic properties [6].

Although the contrast model can address a wide range of data, it cannot account for judgments of similarity that are relational or analogical in nature. Two analogs can be similar

even when the analogs do not have many features in common. To use Gentner’s example, one reason people judge the solar system and an atom as being similar is that our representations of these two systems share a number of higher-order relational matches (as opposed to simple feature matches). For example, electrons revolving around a larger nucleus can be put in correspondence with planets revolving around a larger sun. Although the elements of the two systems are not inherently similar (e.g., a nucleus and the sun differ in size and composition), the two analogs are judged to be similar because a mapping between the two systems exists that preserves higher-order commonalities (e.g., the sun maps to the nucleus and the planets map to the electrons). Of course, there are simpler cases of different dimensions being put in correspondence. For example, people equate high-pitched sounds with bright lights and when asked, “Which is brighter, a sneeze or a cough?” people readily answer that a sneeze is brighter [16]. The contrast model assumes that only identical features can match and does not envisage a matching process that attempts to preserve higher-order compatibilities.

More current models of comparison and analogy (e.g., [2, 13, 11]) do establish relational correspondences when comparing objects. These models tend to prefer mappings between analogs when 1) identical features can be mapped to one another, 2) there are higher-order compatibilities and structures replicate in both analogs, 3) the mapping between the two analogs is or almost is one-to-one. Although these constraints are common to all successful models of human comparison, it is not always clear how these constraints are weighted and manifested in models. In other words, different models may adopt widely different matching algorithms (e.g., [2, 11]), but can be quite similar at the computational level of analysis (in the sense of Marr, 1982). It is important to know what the commonalities and differences of competing models are in order to identify the critical issues that deserve empirical investigation.

The goal of this paper is to specify a computational level theory of comparison and similarity that is quantitative, easily understood, and falsifiable. The computational level theory takes the form of a similarity equation consisting of four terms that combine linearly (weighted by parameters). Unlike algorithmic level models where principles are often obscured within the details of the processing mechanisms, principles in the similarity equation appear as separate terms and it is clear how different principles are weighted. Best fitting parameters for a data set are interpretable and clear predictions can be made about how the best fitting parameters should change as task demands change.

Such a theory might make the common ground between models more obvious to the extent that process models conform to the same underlying computational level theory. A successful computational level theory would also make each algorithmic model’s contribution to the field clearer. For instance, a model that simply conformed to the computational level theory and made no new predictions would have no

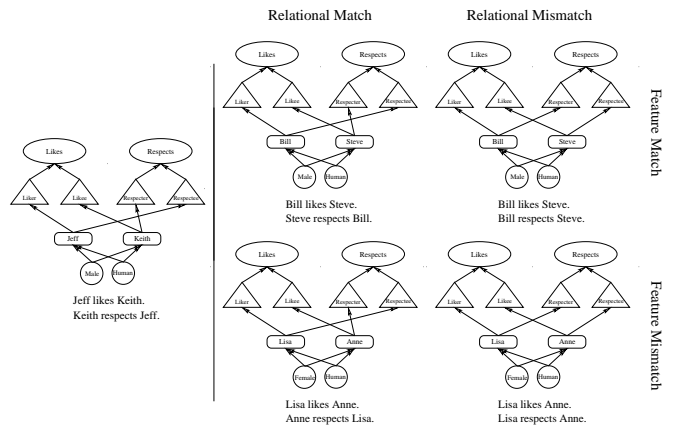


Figure 1: The target analog and its corresponding representation  $S^x$  are shown to the left of the dividing line (subjects were not shown  $S^x$ , just the text). To the right of the dividing line, an example base analog is shown for each of the four conditions (along with its  $S^y$ ).

“added value.” Many current models do have “added value.” For example LISA [11] makes predictions about how working memory limitations and discourse setting affects comparison, MAC/MAC [3] explicitly models retrieval processes in comparison, and IAM [13] focuses on the incremental nature of analogical mapping.

Although theories of comparison hold that similarity and analogical processing are deeply related [5], the linkage between similarity and comparison is even more direct in the similarity equation. In the similarity equation, the correspondences (i.e., mappings) between the two analogs are chosen so as to maximize the similarity equation, much like how an energy function is minimized in a Hopfield network when performing a computation [10]. The maximal value of the similarity equation is monotonically related to the perceived similarity of the two analogs. Thus, similarity both drives the comparison process and is an outcome of it. In the remainder of the paper, the details of the similarity equation are presented as well as some data fits.

### Mathematical Formulation

Analogs  $x$  and  $y$  are represented as directed graphs. Analog  $x$  has  $m$  different representational elements or nodes, while analog  $y$  has  $n$ .  $S^x$  is an  $m$  by  $m$  matrix that capture the connectivity of analog  $x$ . Each entry in  $S^x$  is either 0 or 1.  $S^x_{24}$  set to 1 signifies that node 2 binds to node 4. Notice that this relationship is not symmetrical — part 4 is a parent of part 2, but part 2 is not necessarily a parent of part 4. Analog  $y$  is represented in an identical fashion by  $S^y$ . Figure 1 illustrates some examples of analogs and the graphs that represent them.

In evaluating the similarity of  $x$  to  $y$ , correspondences are established between the representational elements of  $x$  and  $y$  (i.e., the nodes in  $S^x$  and  $S^y$ ). These correspondences or mappings are recorded in the  $m$  by  $n$  matrix  $A$ . Each entry in  $A$  is either 0 or 1.  $A_{ij}$  equal to 1 indicates that element  $x_i$

(of  $x$ ) maps to element  $y_j$  (of  $y$ ). The mappings are selected so as to maximize the value of the similarity equation. The idea is that perceived similarity arises out of a comparison process that establishes mappings between the two analogs. Such mappings would prove useful in analogical reasoning and inference.

For real world problems, it is impractical to try all  $(\frac{mn}{2})^2$  combinations in search of the best mapping. The theory presented here is a computational level theory of comparison and similarity and does not address this issue. The solutions to difficult real world problems can be approximated using combinatoric optimization procedures such as simulated annealing [14]. In essence, every algorithmic model of analogy solves this combinatoric optimization problem by heuristically combining mapping constraints.

The similarity equation consists of four terms that combine linearly:

$$\mathbf{E}(x, y) = \alpha_1 \Theta(x, y) + \alpha_2 \Upsilon(x, y) + \alpha_3 \Omega(x, y) + \alpha_4 \Phi(x, y) \quad \text{where } \alpha_1, \alpha_2, \alpha_3, \alpha_4 \geq 0.$$

The four terms are defined below. The terms are organized from most semantic in nature to most structurally focused.

The  $\Theta$  term is analogous to Tversky's (1977) contrast model and captures raw semantic similarity:

$$\begin{aligned} \Theta(x, y) &= \varphi_1 \mathbf{f}(X \cap Y) - \quad (4) \\ (1 - \varphi_1) &\left( \delta_1 \mathbf{f}(X - Y) + (1 - \delta_1) \mathbf{f}(Y - X) \right) \\ &\text{where } 1 \geq \varphi_1 \geq 0, \text{ and } 1 \geq \delta_1 \geq 0 \end{aligned}$$

where  $\varphi_1$  determines the relative importance of commonalities and differences in determining the similarity of two analogs. The parameter  $\delta_1$  determines how asymmetric the similarity judgment is. Given that the focus of a comparison is usually on the target,  $\delta_1$  should be greater than .5.

The second term captures semantic similarity arising from correspondences:

$$\Upsilon(x, y) = \sum_{i=1}^m \sum_{j=1}^n A_{ij} \mathbf{C}(x_i, y_j) \quad (5)$$

where  $\mathbf{C}(x_i, y_j)$  is:

$$\mathbf{C}(x_i, y_j) = \begin{cases} \varphi_1 \mathbf{F}(x_i) & \text{if } x_i \text{ is identical to } y_j, \\ \text{else } (\varphi_1 - 1) \mathbf{F}(x_i). \end{cases} \quad (6)$$

The  $\Theta$  term and Tversky's contrast model do not distinguish between commonalities and differences that arise from relational elements that are in correspondence and those that are not in correspondence. The  $\Upsilon$  term specifically addresses commonalities and differences that are in correspondence (i.e., elements linked in  $A$ ). Commonalities arising from correspondences are processed differently (i.e., have different time courses and differentially affect perceived similarity)

than commonalities (or matches) that are not in correspondence [8]. Likewise, differences that can be put into correspondence are psychologically distinct from differences that cannot be put into correspondence [15].

Humans are also sensitive to higher-level matches (i.e., compatibility relations), as in the solar system/atom example. Analogs are perceived as similar when they have a common relational structure. The  $\Omega$  term captures this type of similarity and it is high when elements from one analog map to elements in the other analog and their parents are also in correspondence.

$$\Omega(x, y) = \quad (7)$$

$$\sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^n \sum_{l=1}^n S_{ij}^x S_{kl}^y A_{ik} A_{jl} \mathbf{F}(x_i) \mathbf{F}(x_j) \mathbf{F}(y_k) \mathbf{F}(y_l).$$

The  $\Phi$  term is purely structural and ranges from 0 to 1. The  $\Phi$  term is 1 when the mapping between  $x$  and  $y$  is a bijection (one-to-one and onto). In general people prefer analogies or mappings that are one-to-one [4]. Here, we assume that complete mappings are also preferred. The  $\Phi$  term is defined as:

$$\begin{aligned} \Phi(x, y) &= \delta_1 \left( \frac{1}{1 + \sum_{i=1}^m \mathbf{F}(x_i) \left(1 - \sum_{j=1}^n A_{ij}\right)^2} \right) + \quad (8) \\ (1 - \delta_1) &\left( \frac{1}{1 + \sum_{i=1}^n \mathbf{F}(y_i) \left(1 - \sum_{j=1}^m A_{ij}\right)^2} \right) \end{aligned}$$

The data sets considered in the next section are all from controlled experiments and mappings exist that lead to maximal values of  $\Phi$ . Under these conditions, only these solutions are considered by subjects (i.e., the parameter  $\alpha_4$ , weighting  $\Phi$ , is set to a value large enough to prohibit consideration of incomplete or conflicting mappings).

## The Similarity Equation and Human Data

In this section, data from Goldstone (1994) and data from two new experiments is fit by the similarity equation. The fits are intended to illustrate how the similarity equation can be applied to human data and should not be taken as a definitive test of the equation's form. The equation's form will certainly be refined as it is applied to more data sets.

### Goldstone (1994) Experiment 2

The similarity equation was applied to data from Goldstone's (1994) Experiment 2. Subjects rated the similarity of two displays. Each display consisted of a pair of schematic butterflies. Each butterfly could be represented by four features (type of tail, type of body, type of wings, type of head). The number of matches in place (correspondences as in Equation 5) and the number of matches that were not in correspondence was manipulated, yielding fifteen conditions. Table 1 illustrates the fifteen within subject conditions. Only the  $\Theta$  and the  $\Upsilon$  terms from the similarity equation are relevant

Table 1: Various feature transformation from Goldstone’s (1994) Experiment 2. In the Target column, each of the four positions in each letter string represents a part of the target stimulus (i.e., the second position in each string could denote the head). Each letter (A, B, C, D, W, X, Y, or Z) represents a particular feature value. The base stimulus is always represented as ABCD. The next two columns list the number of  $\Theta$  and  $\Upsilon$  matches. The Rated Similarity column denotes human subjects’ rated similarity of the base and target stimuli.

Target	$\Theta$ Matches	$\Upsilon$ Matches	Rated Similarity
WXYZ	0	0	1.91
XYDZ	1	0	1.48
YZDD	1	1	3.09
XYCZ	1	1	3.12
BAYZ	2	0	3.11
YZCD	2	2	4.65
BADZ	3	0	3.60
BACZ	3	1	4.52
BADD	3	1	4.57
ABDZ	3	2	4.96
ABDD	3	3	6.56
ABCZ	3	3	6.78
BADC	4	0	4.82
BACD	4	2	6.38
ABCD	4	4	8.79

for fitting this dataset, along with the  $\alpha_1$  and  $\alpha_2$  parameters, because 1) asymmetries were not a concern ( $|X|$  is equal to  $|Y|$ ), 2) differences and commonalities in Equations 4 and 5 are perfectly correlated allowing all difference terms to be dropped (i.e.,  $\varphi_1$  is set to 1), 3) the value of  $\Omega$  and  $\Phi$  do not vary across conditions.

For simplicity, a linear relationship was assumed between the maximal value generated by the similarity equation and subjects’ similarity ratings. Of course, similarity is probably a more complex function of  $E(x, y)$  for a number of reasons, including the presence of scale effects [19]. Nevertheless, with this simplifying assumption, the similarity equation accounted for 97.0 % of the variance in the data. The similarity equation states that different sources of similarity combine additively. A modified version of the equation was fit to the data that contained a term capturing the interaction between  $\Theta$  and  $\Upsilon$ . This augmented equation did not capture significantly more variance (97.3%), supporting the stance that the terms combine additively. The fit for SIAM, a special purpose interactive connectionist model developed by Goldstone (1994), was equivalent (it accounted for 97.7 % of the variance). The similarity equation offers a simpler account of the data — perceived similarity arises from a linear weighting of the  $\Theta$  and  $\Upsilon$  terms. To SIAM’s credit, it can account for aspects of the time course data, like that subjects tend to weight matches in correspondence (i.e.,  $\Upsilon$  matches) more later in processing than  $\Theta$  matches (this is SIAM’s added value). Such data is outside the province of a computational

level theory.

## Experiment 1

In Goldstone’s (1994) Experiment 2, the  $\Omega$  term was not relevant to the data fit. To further test the predictions of the similarity equation, I collected data from subjects in a task in which higher-order relations could impact the similarity equation’s predictions (i.e., the  $\Omega$  term’s value varies across conditions). In Experiment 1, subjects rated the similarity of two situations. The number of higher-order relations shared by the two situations was manipulated (as well as the number of  $\Theta$  and/or  $\Upsilon$  matches). The main prediction the similarity equation makes in Experiment 1 is that feature and relation matches will affect rated similarity additively.

**Subjects** Twenty-one Northwestern University undergraduate students participated in the experiment for course credit.

**Design and Overview of the Experiment** The two variables (Feature Match/Mismatch and Relation Match/Mismatch) were crossed for a 2 X 2 within subjects factorial design. The design is illustrated in Figure 1. The value of each term in the similarity equation for each condition is shown in Table 2. On each trial, subjects rated the similarity of two situations. Subjects completed 20 trials in each condition for a total of 80 trials. The order of trials was randomized for each subject.

**Stimuli and Counterbalancing** Each stimulus contained the descriptions of two situations. Each situation description consisted of two sentences (see Figure 1). One situation description was displayed on the left side of the screen. The other situation description was displayed on the right side of the screen. Above the description on the left side of the screen was the label “Situation A.” Above the description on the right side of the screen was the label “Situation B.” Underneath the descriptions was a rating scale (1 indicated low similarity and 9 indicated high similarity).

Each situation contained two characters that were either both male or both female. Character names were randomly chosen (subject to constraints imposed by the trial’s condition) without replacement from the following list of names: Anne, Jennifer, Linda, Susan, Wendy, Bill, Jeff, John, Keith, and Steve. On Feature Match trials, the gender of the characters in both situations matched (i.e., all characters were male or female). Whether the common gender was male or female was randomly determined for each Feature Match trial. On Feature Mismatch trials, the genders of the characters in the two situations were different such that the two characters from one situation were both male and the two characters from the other situation were both female. On each Feature Mismatch trial, it was randomly determined whether situation A contained the male or female characters.

Both characters from a situation appeared in both sentences. Each sentence contained a predicate that linked the two characters. The same predicates appear in both situations. Two predicates were randomly chosen without

Table 2: The values of the four terms for the four conditions in Experiments 1 and 2. Notice that  $\Theta$  and  $\Upsilon$  are perfectly correlated.

	$\Theta$	$\Upsilon$	$\Omega$	$\Phi$
Feature Match/Relational Match	8	8	12	1
Feature Mismatch/Relational Match	7	7	12	1
Feature Match/Relational Mismatch	8	8	10	1
Feature Mismatch/Relational Mismatch	7	7	10	1

replacement for each trial from the following list: is taller than, respects, and likes. Which predicate appeared in the first or second sentence within a situation was random. It was also random whether or not the same character appeared first in both sentences in situation A (the character order is fixed for situation B given the character order in situation A and the trial’s condition). Again, Figure 1 illustrates an example situation pair for each condition.

**Procedure** Text was displayed in black on a white background. Trials began with a message displayed in the upper left corner of the screen alerting the subject to prepare for the next trial. After 1000 ms, this message was removed and the stimulus was displayed (i.e., the two situations along with the rating scale). Subjects then pressed a key (1 through 9) to indicate how similar the two situations were (1 indicated low similarity and 9 indicated high similarity). After subjects responded, there was a 1500 ms pause and then the next trial began.

**Results** Table 2 shows the values of the four terms for each condition. As in the previous fit, the number of relevant parameters required can be reduced to 2 (the  $\alpha_1$  parameter for the  $\Theta$  term and the  $\alpha_3$  parameter for the  $\Omega$  term). The mean similarity ratings (averaged across subjects) for each condition are shown in Table 3. The similarity equation fit 99.9% of the variance in the data. To provide a stronger test, individual subjects’ data was fit. Of course, the fit for this data will not be as good because the data for individual subjects is not as stable and each subject uses a slightly different rating scale (i.e., high similarity for one subject may result in a rating of 8, while another subject may give a rating of 7). Nevertheless, the equation fit 71.9% of the variance ( $df=81$ ). A modified version of the equation was fit to the data that contained a term capturing the interaction between  $\Theta$  and  $\Omega$ . This augmented equation did not capture significantly more variance (72.1%,  $df=80$ ), supporting the stance that the terms combine additively.

## Experiment 2

A second experiment explored how task demands affect comparison. The materials and procedure were identical to the previous experiment except that after making a similarity judgment subjects were asked to state how the people in the

Table 3: Similarity ratings for each condition (averaged over subjects) in Experiment 1.

	Relational Match	Relational Mismatch
Feature Match	8.23	4.50
Feature Mismatch	7.51	3.39

target and base analogs corresponded to one another. This judgment should force subjects to focus more on high-order relational matches and should lead to a higher weighting of the  $\Omega$  term relative to the  $\Theta$  term in the similarity equation.

**Subjects** Seventy-one Northwestern University undergraduate students participated in the experiment for course credit. The subjects were from the same population as the subjects in Experiment 1. Experiments 1 and 2 were run concurrently (though no subjects participated in both experiments).

**Design and Overview of the Experiment** The design was very close to that of Experiment 1. The main difference was that subjects made a correspondence judgment after making a similarity judgment. Another difference was that subjects performed sixty trials (fifteen in each condition) as opposed to the eighty trials performed in Experiment 1.

**Stimuli and Counterbalancing** The stimuli and counterbalancing were identical to Experiment 1 with the following addition — after making a similarity judgment, one character was randomly chosen from situation A and another character was randomly chosen from situation B and subjects were asked if they corresponded to one another.

**Procedure** The procedure was identical to Experiment 1 except that subjects made a correspondence judgment immediately after making a similarity judgment. After making the similarity judgment, a text message appeared below the rating scale. The message asked if a particular character from situation A corresponded to a particular character from situation B. Subjects were instructed to press the ‘Y’ key if they thought the two characters corresponded and to press the ‘N’ key if they thought the two characters did not correspond. The Yes/No question was displayed along with both situation descriptions and the rating scale. After making the correspondence judgment, there was a 1500 ms pause and then the next trial began.

**Results** The main predictions held. Table 4 shows the mean ratings for each condition. Feature matches had a small effect on rated similarity while relational matches had a large effect on rated similarity. The ratio  $\alpha_3/\alpha_1$  was three times larger in Experiment 2 than it was in Experiment 1. Subjects also made the correspondence judgments in the manner predicted by the similarity equation’s mapping matrix  $A$ . In terms of fit, 99.9% of the variance in the averaged data was accounted for. For individual subject fits, 66.1% ( $df=281$ ) of the variance was accounted for and adding an interaction term did not improve the fit (66.1%,  $df=280$ ).

Table 4: Similarity ratings for each condition (averaged over subjects) in Experiment 2

	Relational Match	Relational Mismatch
Feature Match	8.32	4.47
Feature Mismatch	8.02	4.08

While the fits from Experiments 1 and 2 (as well as from Goldstone’s Experiment 2) suggest different sources of similarity combine additively, I predict that after consideration of more diverse data sets (e.g., [9]) the similarity equation will be revised to make allowances for interactions between terms under certain conditions. The equation and data presented here are simply intended to motivate a new framework for approaching comparison and similarity.

## Discussion

The similarity equation presented here is a computational level theory of human comparison and perceived similarity that can account for basic findings in the similarity and analogy literatures. The equation provides clarity to the discussion of similarity because it distinguishes between a number of different factors that can affect perceived similarity. An accurate characterization of similarity is critical given its central role in theories of categorization, decision making, analogy, problem solving, and object recognition.

Twenty years after Tversky’s (1977) classic paper, many of the same questions remain. How are the representations of analogs determined, how do they change as an outcome of comparison, and how is feature saliency modulated? One possibility is that instead of static representations being compared, retrieval and comparison are interleaved such that the current mappings between the analogs direct which other information is retrieved and represented in the base and target. Analogical inference may also direct the construction of the target analog’s representation. In other words, properties or features can emerge as a result of the comparison process [18]. Hopefully this work will demarcate what is known and what is common to competing models so that researchers can wisely focus their efforts.

## Acknowledgments

I would like to thank John Hummel and Keith Holyoak for illuminating discussions. I would like to thank Dedre Gentner and Arthur Markman for their helpful comments on a previous draft. Finally, I would like to thank Rob Goldstone for providing me with his data.

## References

- [1] BOWDLE, B. F., AND GENTNER, D. Informativity and asymmetry in comparisons. *Cognitive Psychology* 34 (1997), 244–286.
- [2] FALKENHAINER, B., FORBUS, K. D., AND GENTNER, D. The structure mapping engine: Algorithm and examples. *Artificial Intelligence* 41 (1989), 1–63.
- [3] FORBUS, K. D., GENTNER, D., AND LAW, K. MAC/FAC: a model of similarity-based retrieval. *Cognitive Science* 19 (1994), 141–205.
- [4] GENTNER, D. Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7 (1983), 155–170.
- [5] GENTNER, D., AND MARKMAN, A. B. Structure mapping in analogy and similarity. *American Psychologist* 52 (1997), 45–56.
- [6] GLEITMAN, L. R., GLEITMAN, H., MILLER, C., AND OSTTRIN, R. Similar, and similar concepts. *Cognition* 58 (1996), 321–376.
- [7] GOLDSTONE, R. L. Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 20 (1994), 3–28.
- [8] GOLDSTONE, R. L., AND MEDIN, D. Time course of comparison. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 20 (1994), 29–50.
- [9] GOLDSTONE, R. L., MEDIN, D., AND GENTNER, D. Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology* 23 (1991), 222–262.
- [10] HOPFIELD, J. J., AND TANK, D. W. Neural computation of decision in optimization problems. *Biological Cybernetics* 52 (1985), 141–152.
- [11] HUMMEL, J. E., AND HOLYOAK, K. J. Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review* 104 (1997), 427–466.
- [12] JAMES, W. *The Principles of Psychology: Volume 1*. Dover, New York, 1890/1950.
- [13] KEANE, M. T., LEDGEWAY, T., AND DUFF, S. Constraints on analogical mapping: A comparison of three models. *Cognitive Science* 18 (1994), 387–438.
- [14] KIRKPARTICK, S., GELATT, C. D., AND VECCHI, M. P. Optimization by simulated annealing. *Science* 220 (1983), 671–680.
- [15] MARKMAN, A. B., AND GENTNER, D. Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language* 32 (1993), 517–535.
- [16] MARKS, L. E. Bright sneezes and dark coughs, loud sunlight and soft moonlight. *Journal of Experimental Psychology: Human Perception and Performance* 8 (1982), 177–193.
- [17] MARR, D. *Vision*. W. H. Freeman, San Francisco, 1982.
- [18] MEDIN, D. L., GOLDSTONE, R. L., AND GENTNER, D. Respects for similarity. *Psychological Review* 100, 2 (1993), 254–278.
- [19] PARDUCCI, A. Category judgment: A range-frequency model. *Psychological Review* 72 (1965), 407–418.
- [20] SHEPARD, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. Part 1. *Psychometrika* 1 (1962), 125–140.
- [21] TALMY, L. Figure and ground in complex sentences. In *Universals of human language*, J. Greenberg, C. Ferguson, and M. Moravcsik, Eds. Stanford University Press, Stanford, 1978, pp. 625–649.
- [22] TVERSKY, A. Features of similarity. *Psychological Review* 84 (1977), 327–352.



# Viewpoint Dependent Facial Expression Recognition Japanese Noh Masks and the Human Face

Michael J. Lyons<sup>†</sup>, Andre Plante<sup>†</sup>, Miyuki Kamachi<sup>‡</sup>, & Shigeru Akamatsu<sup>‡</sup>  
(mlyons@mic.atr.co.jp)

Advanced Telecommunications Research International

<sup>†</sup>Media Integration & Communications Research Lab

<sup>‡</sup>Human Information Processing Research Lab

2-2 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, Japan, 619-0288

Ruth Campbell & Mike Coleman

r.campbell@ucl.ac.uk

Dept. of Human Communication Science, University College London

Chandler House, 2 Wakefield St, London WC1 N 1 PG

## Abstract

With certain masks used in Japanese Noh drama the apparent facial expression is a function of the vertical viewing angle. Rotation in depth produces changes in the retinal image of the face which viewers may confound with the distortion of features due to muscular action. In particular, as the mask is tilted forward it appears to smile, and as it is tilted backwards it appears sad. We explored this effect in two experiments with a Noh mask and one with a 3-D model of a laser-scanned human face. Separate British and Japanese subject pools were used to investigate cross-cultural effects. The results confirmed a systematic relationship between vertical angle of view and judged affect. For the Noh mask the effect was culturally moderated, whereas for the human face there was no significant effect of culture. These results are discussed and interpreted in terms of perceptual strategies for processing familiar and unfamiliar faces.

## Introduction

A variety of visual cues inform the viewer of affect. Muscular action of the face (Ekman & Friesen, 1978), causing feature displacement and consequent wrinkling of the skin, conveys the most salient information. Visible changes in skin hue caused by modulation of blood flow are also telling signs to internal state. A further source of information about affect is delivered by body posture: positive affect is accompanied by an upright posture with the head held high and negative states may be signaled by a bowed head and crouched posture (Darwin, 1872).

Visual processing of the feature displacement and textural cues to face muscular action requires a representation sensitive to the fine metric properties of the spatial patterns on the surface on the face (Lyons *et al.*, 1999, 2000). Rotation and translation of the head in 3-D space accompanying vertical movements of the head or changes in viewpoint, distort configural relations on the face as they appear in the 2-D projection of the face on the retina. Indeed the signals from facial muscle action and head posture conflict. Affect should be judged relatively positively in a head held up and back. Under frontal viewing conditions, tilting the head backwards reduces upward curvature of the mouth in the 2-D projection of

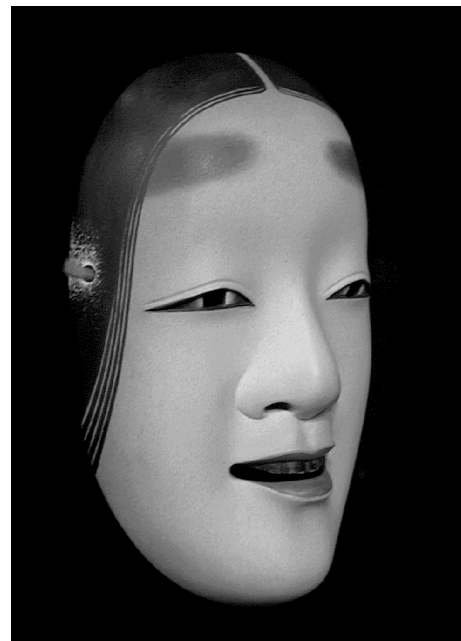


Figure 1: Magojiro mask used in the Japanese Noh drama.

the face, giving the impression of a sadder or negative expression. Tilting the head forward, a negative head posture signal, increases upward curvature of mouth - which usually accompanies the muscular action signaling a smile. It is therefore interesting to ask whether the changes due to rigid displacement of the head interferes with the interpretation of featural and textural cues due to muscular action of the face. Our interest in this question was stimulated when we learned of an illusion of facial expression perception involving masks (figure 1) used in Japanese traditional Noh drama (Komparu, 1983). It has been known for centuries in the Noh theater that certain masks, particularly those used to portray young female roles, appear to change expression as the vertical inclination of the mask changes (figure 2). Tilt the mask for-

ward and it appears to smile; tilt it backwards and it appears sad.

Is this phenomenon evidence for a lack of invariance of the facial expression recognition system under rigid transformations? Or are special techniques employed by the mask carver and Noh actor to trick the visual system into mistaking a rigid rotation of the mask for a non-rigid distortion of its internal features? We conducted three experiments to investigate the following questions (1) Do changes in vertical inclination in fact generate different perceptions of affect? (2) Is the effect culturally moderated or does it depend on familiarity with the mask? (3) Is the effect particular to the Noh mask or does it generalize to the human face?

## Materials and Methods

### Stimuli

Stimuli for **Experiment 1** consisted of photographs of a Noh mask (figure 2) at 13 inclinations, from  $-30^\circ$  to  $+30^\circ$  in equal  $5^\circ$  increments. An antique Magojiro mask, used for young female roles, dating to the Edo period (1600-1868) was photographed on a Noh stage under lighting conditions similar to what would be used during a performance. The mask was photographed from a frontal viewpoint using a digital camera (Kodak Professional DCS 460) from a distance of 7.7m with a 200 mm lens. The 3060x2036 pixel 24-bit color images were cropped and re-sampled to 300x400 pixel tiff images. Stimuli for **Experiment 2** (figure 3) comprised the same images, but cropped so as to emphasize the internal features of the face. Stimuli for **Experiment 3** (figure 4) were derived from the head and face of a 30 year old Japanese female model posing a neutral expression similar to that of the Noh mask. A Cyberware 3030 Color 3-D scanner was used to acquire shape and color information of the model's head. The 24-bit RGB color map was acquired under room light from an overhead fluorescent lamp. Screen captures were taken at 13 (virtual) head inclinations from a reconstructed 3-D model of the head, the face oriented frontally and saved as 24-bit 300x400 pixel tiff images. The vertical viewing angles varied from  $-30^\circ$  to  $+30^\circ$  in equal  $5^\circ$  increments. Inter-ocular distance and eye position were normalized for each stimulus set and matched across sets.

### Experimental Procedure

Experiments were run in separate laboratories in London and Kyoto. In each case the stimuli were displayed on a 17 inch 24-bit color computer monitor in a slightly darkened room. Viewing distance was approximately 60 cm. Following a practice trial, four epochs of all 13 stimuli were presented in succession, with presentation order randomized within each epoch. Presentation order was as follows: fixation point (500 ms) - blank (400 ms) - stimulus (300 ms). Subjects were instructed to respond whether the stimulus face appeared happy or sad by pressing the left or right shift key. Japanese subjects were instructed in Japanese using the terms “*yorokobi*” and “*kanashimi*”. Left/right assignment of response keys

was counterbalanced across subjects. The words “happy” and “sad” (in English for both subject groups) appeared on the response-appropriate side of the screen for each subject to maintain correct response orientation. All Japanese subjects were familiar with the English terms. Reaction times and decision type were recorded automatically for each subject for each trial.

### Subject Pool

Different subjects were run for each of the three experiments. There were 5 females and 5 males from each cultural group for each experiment, making 60 subjects in total. Subjects were undergraduates, graduates and staff from Doshisha University, Kyoto and University College London. Ages ranged from 18 to 50 years. All had normal or corrected-to-normal vision. All were either native to the country of testing or had first-school education in that country. The Japanese subjects were familiar with Noh masks as images or, occasionally, objects. None of the UK subjects had familiarity with Noh or had visited Japan.

## Results

All three experiments had the same mixed, 3-factor repeated measures design. There were 13 levels of the first factor (inclination), which varied within subjects, and one level of the first between-subjects factor (culture) and the second between-subjects factor (gender). Both RTs and response (as the proportion of “happy” responses over 4 trials) were examined. Following initial analysis, in which the full range of scores were examined, only the seven mid-range scores ( $-15^\circ$  to  $+15^\circ$ ) in the following treatment. Preliminary data analysis confirmed parametric, normal distributions of the scores reported here. There were no significant differences in the pattern of results when the full range was included. However, the full-range analyses may be less reliable due to non-Gaussian distribution of scores at some of the endpoints. RT (medians) showed no systematic relationship to the other variables, and are not reported here. Gender had no effect on any of the analyses and effects of gender were not considered further. Table 1 outlines the significant finding for each experiment. The relevant graphs are shown in figure 5. These analyses show a significant linear relationship between angle of inclination and rated happiness for all three experiments. The Noh mask, but not the scanned face, is classified differently by Japanese and British viewers. Further analyses and their justification in terms of individual experimental hypotheses are reported below.

### Experiment 1 - Full Mask

The experimental hypothesis was that the Noh mask would generate changes in perceived affect as a function of vertical angle. The perception of facial expressions is thought to be similar for our two cultural groups (Matsumoto, 1992), hence similar results were expected for the two subject groups.

The results confirmed the predictions in general terms, but with some important deviations. The groups differed in

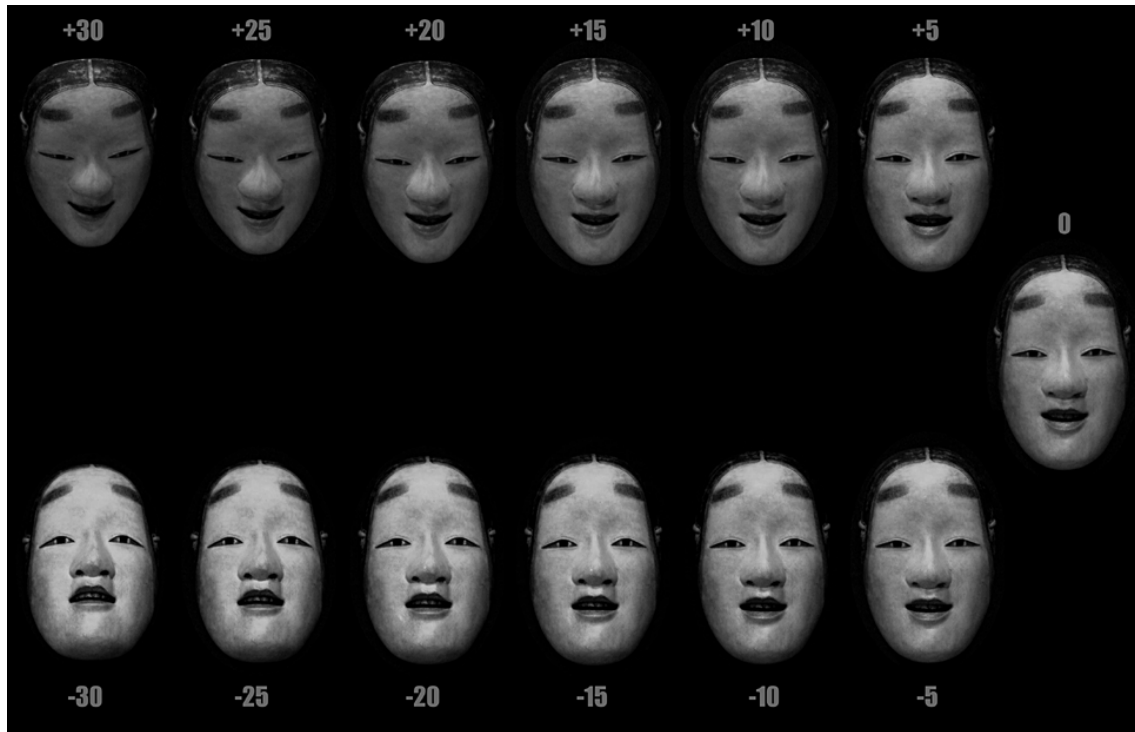


Figure 2: Stimuli for Experiment 1. Edo-period Magojiro mask at 13 different vertical inclinations.

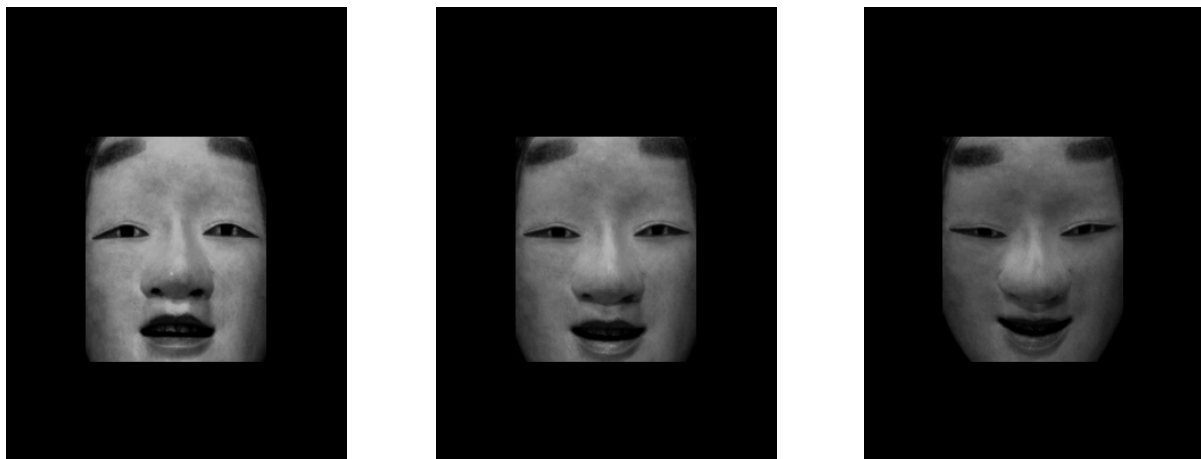


Figure 3: Sample stimuli for Experiment 2. Same images as in figure 2 but with the edges of the mask cropped to highlight internal features. Images from left to right show the mask at inclinations of  $-15^\circ$ ,  $0^\circ$ , and  $15^\circ$ .

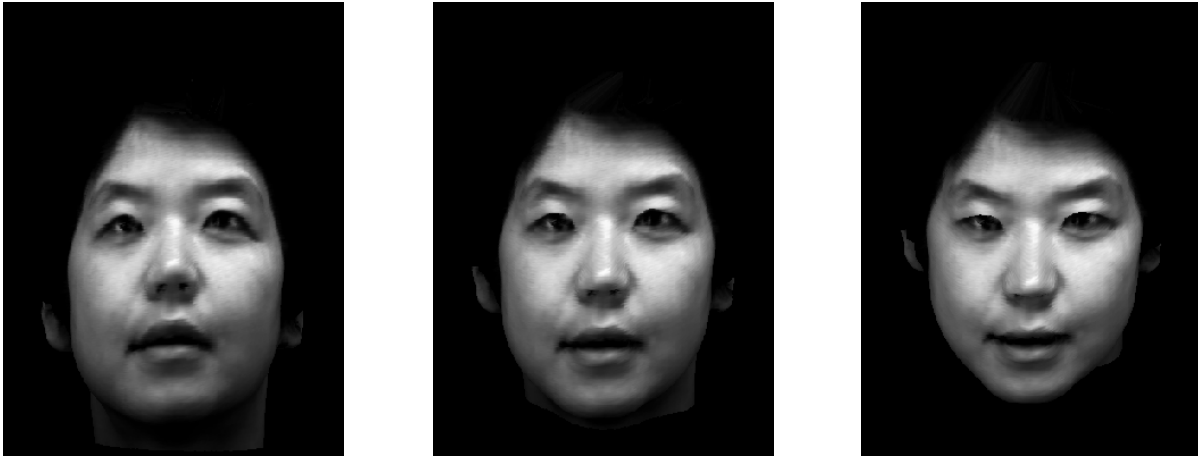


Figure 4: Sample stimuli for experiment 3. Images captured from rendered 3-D model of a Japanese female face obtained using a Cyberware laser scanner. Images from left to right show the face at inclinations of  $-15^\circ$ ,  $0^\circ$ , and  $15^\circ$ .

the inclination at which a “neutral” expression is perceived. Japanese viewers rated the back-tilted mask more positively than British viewers. This may reflect different boundaries in terms of perceived facial expression on the categories of “happy” and “sad” and their cognates in Japanese or lower rates of sadness recognition for Japanese viewers. We address this further in Experiment 3. A second unexpected finding is that the groups differed in the linearity of the relationship with inclination angle (interaction of group and inclination was highly significant). While the relationship was linear over this range for the British subjects, for the Japanese subjects, the proportion of “happy” responses peaked at  $5^\circ$ , and then dipped. At  $15^\circ$ , mean “happy” response was no greater than at  $-10^\circ$ . Why should this change in perceived expression occur? One possibility is that the Japanese viewers were more sensitive to postural cues in the images of the head. A head bowed forwards may be seen as “sadder”. Perhaps the two subjects groups weight the posture and internal features cues differently. For the Noh mask images used, pose cues are most visible in the disposition of the top of the head and the chin with change in inclination of the head.

### Experiment 2 - Cropped Mask

In this experiment, the face images were cropped to diminish cues to head pose and emphasize internal features of the face (figure 3). The experimental prediction was that this may eliminate the non-linearity in the relationship between perceived expression and vertical inclination in the Japanese viewers. The results supported this. In this study, the “dip” at greater positive inclinations was greatly reduced. Thus it appears that Japanese viewers take account of cues to head pose in ascribing expression to the image of the vertically inclined mask.

Otherwise, experiment 2 replicates the main findings of experiment 1: a linear relationship between angle of inclination and judged expression; as well as a group difference

emerged between Japanese and British viewers. The group difference may indicate that the terms “happiness” and “sadness” in English and Japanese do not share similar extensions. This would suggest that Japanese may be more willing than British viewers to ascribe “happiness” (a socially acceptable facial signal) to a relatively “unhappy” face, despite the apparent reversal of this pattern for masks at high forward tilt. If this were so, we would expect a similar disparity between groups to emerge when images of natural faces are perceived. Experiment 3 explores this possibility.

### Experiment 3 - Human Face

This experiment used stimuli generated from a 3-D laser scan of a human face (figure 4) to explore the question: do group differences in ascribing expression to a cultural artifact, the Noh mask, extend to natural face images? If they do, we may infer that cultural and linguistic interpretations of facial expression may differ between these groups. If they do not, then the Noh mask may have special perceptual status for Japanese viewers. The findings strongly support the latter conclusion. The relationship between inclination angle and happy-sad judgments was to all purposes identical in both groups. Mann-Whitney non-parametric t-tests explored group differences at each orientation point. None approached significance. We can conclude that the Noh mask effects reported in Experiments 1 and 2, including both the dip in the function at high angles of forward tilt and the “happier” classification at most other angles, reflected a cultural phenomenon - but one related to *perceptual* processing differences between the groups. Though the laser-scanned face did not replicate the lighting conditions of the naturally photographed images used in Experiments 1 and 2, the relationship between inclination and judged expression still held, suggesting that the difference in lighting differences did not contribute notably to the illusion for this set of conditions.

Table 1: Summary of F values, separate ANOVAs for each experiment (SPSS GLM).

	F(6,216) Main effect of inclination	F(1,36) Main effect of group	F(6,216) Group inclination interaction	F(1,36) Main effect of inclination Linear trend	F(1,36) Group inclination Quadratic fit
Experiment 1 Full Mask	10.5 p < 0.001	NS	4.67 p < 0.001	15.33 p < 0.01	7.56 p < 0.01
Experiment 2 Cropped Mask	23.9 p < 0.001	5.27 p < 0.05	3.47 p < 0.01	55.77 p < 0.001	3.47 p < 0.01
Experiment 3 Scanned Face	14.96 p < 0.001	NS	NS	36.88 p < 0.001	NS

## Discussion

The three experiments confirmed that the angle of vertical inclination of a face profoundly influences a simple expression discrimination task: faces tilted down have a happier cast than those tilted back. This may be understood in terms of the projection of the three-dimensional facial surface onto the image. An earlier study (Cavanagh *et al.*, 1988) noted the effect as an example of the failure of shape constancy under rotation in depth. Another study (Kappas *et al.*, 1994) looked at viewpoint dependence of facial expression recognition using video clips of posed dynamic expressions as well as a schematic wire-frame model of the face, both quite different from the Noh mask stimuli and scanned face used here. However, that work did not attempt to look at facial expression cues separately from pose cues, as in experiment 2, or study different cultural groups.

A surprising but consistent (40 subjects altogether) finding was that the Noh mask elicited different responses in the two cultural groups. The skilled processing of faces has typically been described as configural (Diamond & Carey, 1986; Young *et al.*, 1987). That is, skilled viewers take account of the various face features and their disposition in coming to a unified account of the identity or reading of the face. Their reading of the face cannot be predicted on the basis of local featural detail. One possibility is that familiarity may have delivered a greater degree of configural processing for the mask in Japanese than British viewers, for the Noh mask occasionally appears in the Japanese media, though an understanding of Noh, or interest in Noh as a tradition is no longer widespread in the Japanese population. One local feature that reliably signals “happy-sad” is the curvature of the mouth. It is possible that British viewers of the Noh mask took account of this feature alone. For Japanese viewers, other aspects of the face may have moderated the effect. Only further experiments will indicate what facial aspects these may be.

At the outset of these studies, we speculated that the three-dimensional structure of the Noh mask and the disposition of the painted features, may be intentionally designed to elicit changes of perceived expression with small changes in pose. Examination of the 3-D structure of the mask showed, for example, that the depth of the mouth region is exaggerated

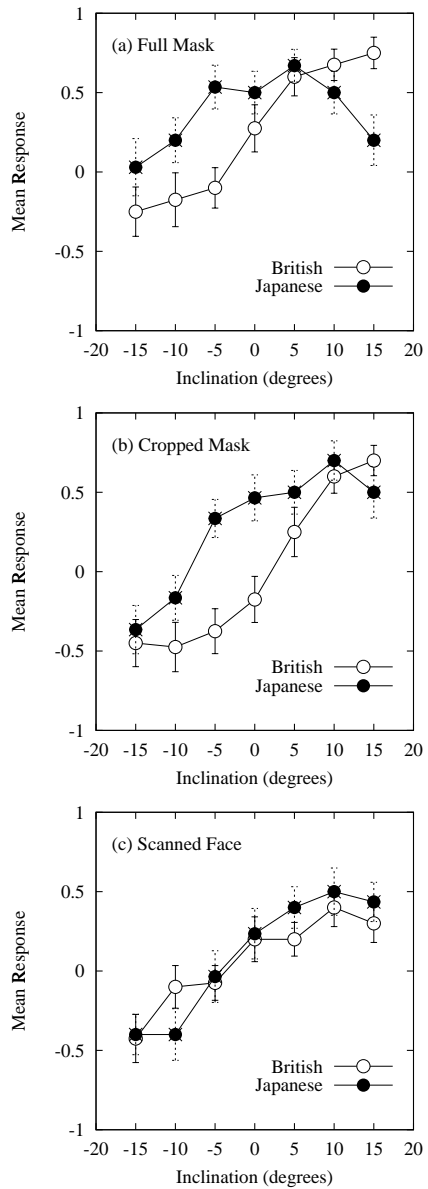


Figure 5: Mean response versus vertical inclination for the 3 experiments.

relative to the human face. Our psychological studies confirmed that small changes in pose of the mask lead to significant changes in perceived affect. A forward tilted mask appeared relatively happy and one tilted backwards, relatively sad. Paradoxically, however, in the stylized use of mask pose in Noh drama, the convention is the opposite to our findings. In one gesture known as *terasu* (shining), signifying a happy state, the mask is turned upwards. In another known as *kumorasu* (clouding), signifying a sad state, the mask is turned downwards (Komparu, 1983).

In this connection, it is notable that Zeami (1363-1443), the most influential early Noh dramatist, ranked *yugen*, or subtle profundity, as the highest aesthetic principle of Noh (Zeami, 1968). In the framework of the Noh world, a joyful pose tempered with a slightly sad mouth may be appreciated as more beautiful than a direct expression of joy. Likewise, sadness or pain masked with a smiling mouth suggests more emotional complexity than an display of pure sadness. A further interpretation is possible, not necessarily in conflict with the above. The psychometric curves (figure 5) show that small changes in inclination angle significantly affect perceived facial expression. Minor movements of the actor's head may trick viewers into thinking that the internal features of the mask are moving non-rigidly as if it were an animated living face. One of the authors (MJL) has observed this effect while watching a Noh play. In a related perceptual effect, a rigid 3-D stick man figure rocked longitudinally back and forth can appear to walk with non-rigid limb movement in 2-D projection (Sinha & Poggio, 1996).

### Acknowledgements

We are grateful to Hotaka Komparu for loan of the Magojiro mask and a helpful discussion about Noh drama; the management at Nara Shin Kokai Do for access to the Noh stage; Akiko Tohma and Katsunori Isono for their generous help; Larry Maloney and Patrick Cavanagh for stimulating discussions; and Vicki Bruce for encouraging us to pursue this study.

### References

- Cavanagh, P., Peters, S. & von Grnau, M. (1988). Rigidity failure and its effect on the Queen. *Perception*17 suppl. 27A.
- Darwin, C. (1872). *The expression of the emotions in man and animals*. London, John Murray.
- Diamond, R. & Carey, S. (1986). Why faces are and are not special: an effect of expertise. *J. exp. Psychol.* 115, 107-117.
- Ekman, P. & Friesen, W.V. (1978). *Facial Action Coding System*. Palo Alto, CA, Consulting Psychologists Press.
- Kappas, A., Hess, U., Barr, C. & Kleck, R. (1994). Angle of regard: the effect of vertical viewing angle on the perception of facial expressions. *J. Nonverbal Behav.* 18, 263-280.
- Komparu, K. (1983). *The Noh Theatre: Principles & Perspectives*. New York & Tokyo, Weatherhill/Tankosha.
- Lyons, M.J., Budynek, J., & Akamatsu, S. (1999). Automatic Classification of Single Facial Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21, 1357-1362.
- Lyons, M.J., Morikawa, K., & Akamatsu, S. (2000). A Linked Aggregate Code for Processing Faces. *Pragmatics & Cognition* 8, 63-81.
- Matsumoto, D. (1992). American-Japanese cultural differences in the recognition of universal facial expressions. *J. Cross-Cultural Psychol.*23, 72-84.
- Sinha, P. & Poggio, T. (1996). Role of learning in three-dimensional form perception. *Nature*384, 460-463.
- Young, A.W., Hellowell, D., & Hay, D.C. (1987). Configural information in face perception. *Perception*16, 747-759.
- Zeami, M. (1968). *Kadensho*. Kyoto, Doshisha Univ., Sumiya-Shinobe Publ. Inst.

# Is Lexical Retrieval in Speech Production like Recall or Recognition? The Effects of Word Frequency and Neighbourhood Size

**Siobhan B. G. MacAndrew** (s.macandrew@abertay-dundee.ac.uk)

Division of Psychology, School of Social & Health Sciences  
University of Abertay Dundee  
Dundee DD1 1NJ, UK

**Trevor A. Harley** (t.a.harley@dundee.ac.uk) & **Sheila Colgan**

Department of Psychology, University of Dundee  
Dundee DD1 4HN, UK

## Abstract

We investigate the effects of word frequency and lexical neighbourhood density on word recall and recognition. We found a three-way interaction between memory task, the size of lexical neighbourhood of a target word, and target word frequency. In particular, performance on low frequency words with many lexical neighbours was surprisingly good in the recognition condition. The results show that the number of lexical neighbours of the target moderates the word frequency effect in recognition. Large neighbourhood size always has a facilitatory effect upon performance. The findings are contrasted with those observed in lexical access in speech production.

## Introduction

To what extent is retrieving a word when speaking like accessing a fact from long-term memory? In particular, how do the language processes involved in lexical access for spontaneous speech production relate to the memory processes involved in the retrieval of word lists from long term memory? On the one hand, our intuition is that lexical retrieval is like recall. Indeed, we even talk in these terms in every day use, using constructions such as “I cannot recall that word”. On the other hand, some models of lexical access in speech production involve search through a list of phonological forms (Butterworth, 1980, 1989; Fay & Cutler, 1977). Such a search might well involve an element of recognition when the appropriate form is reached. This paper looks at two psycholinguistic variables that are well known to influence lexical access (word frequency and lexical neighbourhood size), and investigates their effect on free recall and recognition performance. We compare their effects on a memory task with their effects on a language production task.

Word frequency is an important variable in all language tasks, including speech production (Harley, 1995). Frequency always has a facilitatory effect in speech production. For example, we are faster to name high frequency words and objects with high frequency names (see Jescheniak & Levelt, 1994; Oldfield & Wingfield, 1965). Harley and Bown (1998), using a laboratory-based “tip-of-the-tongue” (TOT) induction task (Brown &

McNeill, 1965), showed that we are more likely to experience a TOT state on less common words.

The second variable employed in this study is lexical neighbourhood size. Some words (e.g. “corpse”) are phonologically unique in that there are no other words that sound like them. Other words (e.g. “cage”) have a large number of phonological neighbours (“page”, “rage”, “sage”, and “cave”, among others). Obviously we need a suitable measure of lexical similarity; we discuss this below. It is well established that a word’s lexical neighbours play in an important role in word recognition (e.g. Glushko, 1979; Grainger, 1990). It is now also becoming apparent that they play some role in word production. Harley and Bown (1998) showed that the number of phonological neighbours a word has affects lexical retrieval in the tip-of-the-tongue state. In particular, they showed that when word length and frequency are controlled for, people are more likely to have difficulty with words that have few phonological neighbours. This result showed that a large set of potential responses can in fact increase the chances of successful retrieval of the target. Harley and Bown hypothesised that structurally similar items provide supporting activation for each other. This finding also supports the “insufficient activation” hypothesis for the origin of TOTs (Burke, MacKay, Worthley, & Wade, 1991).

Although research on the effects of neighbourhood density on lexical access in speech production is at an early stage, the pattern observed is that it is easier to produce frequent words that have many neighbours. Will this pattern be observed in memory tasks? Of course, the pattern observed might well differ depending upon the exact task used. In particular, we might observe different outcomes depending on whether we use a recognition or recall memory task.

The effect of word frequency on recognition is well known, if poorly understood. The “word frequency effect” is the finding that recognition memory is better for low frequency words than high frequency words (Baddeley, 1990; Gregg, 1976; see Guttentag & Carroll, 1998, for a recent review). In recognition, we make a judgement about whether or not we have recently seen a particular item. Is

the stimulus activated because of recent exposure, or is it activated just because of an intrinsic property, such as its high frequency? There is no such conflict in the case of low frequency words, where high activation of the stimulus representation is much more likely to have come from recent exposure in the study list. Put more colloquially, frequent words are less distinctive. This line of reasoning is commonly known as the *memorability hypothesis* (e.g. Brown, Lewis, & Monk, 1977).

Less is known about how word frequency affects performance in a recall task. If frequency operates by raising the activation levels of frequently used items (e.g. Morton, 1979), then the free recall of a high frequency word should be relatively easy and that of a less frequent word relatively difficult. In summary, frequency should facilitate recall but might hinder recognition. In the light of these hypotheses, the finding that in speech production high target word frequency always has a facilitatory effect suggests that lexical access is more like recall than recognition.

The effect of a large lexical neighbourhood is to increase the number of potential responses. The existence of plausible alternative responses may have different effects on recall and recognition. Recognition is more difficult when selecting from a large set of plausible responses than a smaller set (e.g. in the long-term memory version of Sternberg, 1966). Crucially, the similarity between targets and distractors affects recognition (Dale & Baddeley, 1962). Hence large lexical neighbourhoods should hinder recognition.

It is less clear how neighbourhood size will affect recall. If free recall acts like speech production, we would expect that words with many neighbours should be relatively easy to recall. One way of conceptualising this is that a word's neighbours should act as possible retrieval cues. On the other hand, in the two-process, generation-recognition account of free recall (Anderson & Bower, 1974; Kintsch, 1970), recall contains within it an element of recognition. In this case, the recall of words with many neighbours will be either hindered, or the effects of the two processes may cancel out so that no difference is observed. We attempt to explore these issues by examining the effect of the target word's phonological neighbourhood.

It is unclear how word frequency and lexical neighbourhood size will interact. If either recognition or recall resembles speech production while the other does not, we will obtain a three-way interaction with particularly poor performance on low frequency words with few neighbours. The simplest prediction is that speech production resembles recall, and that the pattern observed in speech production should therefore also be observed in the free recall task. It is less clear what should happen in the recognition task. One possibility is that the word frequency effect should overwhelm any effects of neighbourhood size, but any prediction here is prematurely speculative.

In summary, the aim of this paper is to examine the effects of word frequency and lexical neighbourhood size on measures of memory.

## Method

### Participants

We tested 30 volunteers, who had a mean age of 34 years. They were all psychology undergraduates of the University of Dundee, Scotland. Ten females and five males took part in each of the two experimental conditions.

### Materials

All of the words used in the experiment were nouns of one or two syllables in length. The experiment required a printed list of target words used in the learning phase of the study, and a printed list of these words plus distractor words in the recognition condition.

The target items were the same as those of Experiment 2 of Harley and Bown (1998). There were 60 words in the target list, of which 30 were of high frequency (at least 100 instances per million words, with a mean of 163.7, as sampled in Francis & Kucera, 1982), and 30 of low frequency (under 9 instances per million, with a mean of 3.7). Within each list of 30, 15 of the target words had a dense lexical neighbourhood as evidenced by a mean N value of 15.1 (see Coltheart, Davelaar & Besner, 1977; our figures were taken from the MRC Database of Coltheart, 1981). The remaining 15 words had no close orthographic neighbours as evidenced by a mean N value of 0. The N value is a measure of a word's orthographic neighbourhood size: it is the number of other words that can be made from a particular word by changing one letter. Obviously the higher the N score, the larger the orthographic neighbourhood. Orthographic and phonological neighbourhood sizes are highly correlated. This issue is discussed in depth in Harley and Bown (1998), who found the same results whether orthographic or phonological neighbourhood size was used. The properties of the materials are summarised in table 1.

Table 1: Properties of materials

Condition	Frequency	N Value
High F, high N	246.7	15.1
High F, low N	225.5	0
Low F, high N	7.2	15.1
Low F, low N	5.6	0

This process yielded four sets of fifteen target words, balanced for frequency and orthographic neighbourhood size, comprising words of high frequency and high N value, high frequency and low N value, low frequency and high N value, and low frequency and low N value. The words were combined in random order to form one list.



The target words were printed in black ink, one beneath the other, in two columns on A4 paper in random order for use in the presentation phase of the experiment. Examples include “ball” and “date” (high frequency, high N), “cage” and “dove” (low frequency, high N), “growth” and “view” (high frequency, low N) and “corpse” and “tinsel” (low frequency, low N).

The recognition condition of the experiment consisted of the targets and 60 distractor items. In this particular experiment, the distractors were related in meaning to items from the target word set. Although it is clearly of interest to study other types of distractor, we wanted to make this task similar to speech production. Therefore the potential competing words were maximally plausible alternatives that were semantically similar to the targets. The target words were paired with close semantic associates. The items for the recognition task were also hand-printed in black ink, one beneath the other, in random order, on a single sheet of A4 paper.

## Procedure

All participants were given 5 minutes in which they were told to read the presentation list of 60 words and to try to remember them. This was followed by an interval of 5 minutes during which participants engaged in conversation and listened to music. Participants in the recall condition were then given 5 minutes to write down as many words as they could recall from the presentation list. Participants in the recognition condition were told they had 5 minutes to read the recognition list of 120 words and underline in pencil any words that they thought they had previously seen.

## Results

The experimental design comprised three factors. There was a between-subjects factor of memory task (with the two levels of free recall and recognition). There were two within-subjects factors, one of word frequency (with the two levels of high and low frequency) and one of lexical neighbourhood size (with the two levels of high N score and low N score).

A 2x2x2 ANOVA on the correct memory scores of the participants showed main effects of memory task ( $F(1, 28) = 14.72, p < 0.001; MSE = 143.0$ ), word frequency ( $F(1, 28) = 7.71, p < 0.025; MSE = 27.1$ ), and neighbourhood size ( $F(1, 28) = 27.9, p < 0.001; MSE = 130.2$ ).

Importantly, there was a significant three-way interaction between memory task, word frequency, and lexical neighbourhood size ( $F(1, 28) = 10.80, p < 0.01; MSE = 27.1$ ). There was also a significant two-way interaction between memory task and word frequency ( $F(1, 28) = 12.96, p < 0.005; MSE = 49.4$ ). The interaction between memory task and neighbourhood size approached significance ( $F(1, 28) = 3.95, p = 0.06, MSE = 18.4$ ), but there was no hint of any interaction between frequency and

neighbourhood size ( $F(1, 28) = 1.2$ ). Figure 1 summarises these results.

As was expected, the level of recognition performance was better than that of free recall. Performance on words with dense lexical neighbourhoods was better than that on words with sparse neighbourhoods across both the recall and recognition conditions. Performance on high frequency words was generally better than on low frequency ones. The likely source of the three-way interaction, however, is that low frequency, dense-neighbourhood words perform unusually well in the recognition task (or unusually poorly in the recall condition). Words with many neighbours are significantly easier to recognise than those with few neighbours ( $t[28] = 4.01, p < 0.001$ ). There is no difference between the corresponding conditions in the recall task ( $t[28] = 1.10$ ). Indeed, recognition performance for the less frequent words with many neighbours was the best of all conditions. A consequence of this interaction is that there is no word frequency effect for words with few neighbours in the recognition task; performance on low frequency words is in fact worse than that on high frequency words, although not significantly so ( $t[28] = 0.95$ ).

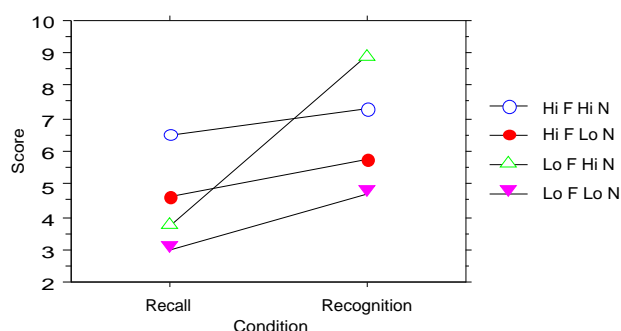


Figure 1: The effects of word frequency and lexical neighbourhood size on recall and recognition.

## Discussion

In summary, we obtained a three-way interaction between memory task, word frequency, and phonological neighbourhood size, demonstrating that these two variables have significant differential effects upon recall and recognition. Large neighbourhood size always has a facilitatory effect on both recall and recognition performance, suggesting that the neighbours of target items act as a source of support rather than interference. In the recognition task, performance on words with few neighbours was better on high frequency words than low frequency words, reversing the usual frequency effect.

In the recall task, performance on high frequency words was uniformly better than on low frequency words, confirming our predictions based on the consideration of lexical activation levels. The amount of facilitation provided by dense lexical neighbourhoods is not large, but

words with many neighbours are easier to recall than words with few.

The findings in the recognition task are more complex. The word frequency effect in recognition, whereby less frequent items are easier to recognise than high frequency words, was replicated only for words with many lexical neighbours. There was no advantage (indeed, a slight disadvantage) for low frequency words with few neighbours. This suggests that any account of the word frequency effect must take into account the role of lexical neighbourhood size. The other conditions in the recognition task are in line with those of the recall task, bearing in mind the expected generally better performance in the recognition task.

Why should low frequency words with no or few neighbours be particularly difficult to recognise? The result appears contrary to the memorability hypothesis. A word which is orthographically and phonologically unique as well as uncommon should be more noticeable and therefore memorable than one with many neighbours.

There are at least two possible explanations. The first is that during the study phase, a target word primes the words in its neighbourhood. During the test phase of the experiment, the primed items then cue the target. The more neighbours there are to act as primes in the test phase, the more likely is a correct response. Words with few neighbours do not have this advantage.

Consideration of the attention-likelihood model of Glanzer and Adams (1990) suggests another explanation. They suggested that in the study phase of a recognition experiment, people pay more attention to some items than others. In general people might redistribute effort at encoding or in rehearsal towards troublesome items (see also Fritzen, 1975; Hastie, 1975; Murnane & Shiffrin, 1975). Low-frequency words with many neighbours may strike participants as odd. They therefore pay a disproportionate amount of attention to them, in particular ensuring that the low-frequency target is not in fact one of its own neighbours. On the other hand, it is possible that participants consider low-frequency words with few neighbours to be "obvious", and therefore pay little attention to them. In the recognition phase, performance will be poor on those items that had less attention allocated to them in the study phase (the low-frequency few-neighbours words). We cannot distinguish between these two possible explanations on the basis of our current data, and of course, they may not be incompatible.

Attention-likelihood theory is one explanation of the "mirror effect". Consider an experiment with two conditions (e.g. high and low frequency items) where the items in one are better recognized than items in the other. Then the superior condition will give better recognition of previously-seen items (i.e. targets) as being old but also better recognition of new items (i.e. distractors) as being new. (See Glanzer, Adams, Iverson, & Kim, 1993; Glanzer, Kim, & Adams, 1998; Stretch & Wixted, 1998; but see also Murdock, 1998.) Consideration of lexical

neighbourhood size may be helpful in giving an account of the mirror effect.

Another surprising finding is that, counter to our intuitions and prediction, the pattern of performance observed by Harley and Bown (1998) in the TOT task is here mirrored in the recognition task, and not in the recall task. In particular, Harley and Bown found a large difference between low frequency words with dense and sparse neighbourhoods. Here we only observed this difference in the recognition task. This suggests that lexical access in speech production contains an important recognition component. Of course, some caution is necessary in making this claim; it is necessary to reproduce our findings on a task more directly oriented to speech. There are at least two possible loci for a recognition component in lexicalization. First, lexical search models such as those of Butterworth (1980) and Fay and Cutler (1977) involve search through ordered lists of lexical entries. Selecting the correct entry might involve recognition. Second, speech production might contain an element of monitoring and editing. These processes might involve recognition. There is independent evidence for the existence of monitoring processes from self-repair of speech (see Levelt, 1989) while others (e.g. Baars, Motley, & MacKay, 1975; Butterworth, 1982) postulate that it is necessary to account for characteristics of speech errors.

An important caveat to any conclusion regarding the resemblance of speech production to other memory tasks concerns what happens in the tip-of-the-tongue state. The presumption in the literature is that a TOT state is an extended form of a hesitation in normal speech (see Harley, 1995, for a review; see also Levelt, 1989). Harley and Bown (1998) suggested that strategic factors might sometimes be operative in laboratory-induced TOT states. In particular, we suggested that there might be an editor responsible for monitoring the output of the interlopers, the words that often spontaneously come to mind when in a TOT state. Others have also proposed that our potential speech output can be edited by a late-acting monitor (e.g. Levelt, 1989). This editor might sometimes discard grossly implausible candidates. The editor must be far from perfect, however, as many implausible candidates are often output; and about a quarter of the time these interlopers bear no obvious relationship to the target. If and when it operates, this post-access monitor might plausibly contain an element of a recognition process. There is no reason to suppose that this applies to either spontaneous production or the strivings to retrieve the target word itself.

If this is the case, the recognition component observed in TOT states comes from the action of post-access strategic processes, rather than the processes of lexical retrieval themselves.

In summary, we have shown that lexical access in the tip-of-the-tongue state surprisingly resembles performance on a recognition task rather than on a free recall task. We have also shown that the word frequency effect in recognition is moderated by the size of the lexical neighbourhoods of the target items. The exact way in

which neighbours exert their effects in these tasks remains to be explored.

### Acknowledgements

We are grateful to several anonymous reviewers for their helpful comments. Our data were collected by Sheila Colgan as part of an honours dissertation at the University of Dundee. We gratefully acknowledge financial assistance from the University of Abertay Dundee research development fund.

### References

- Anderson, J. R., & Bower, G. (1974). A propositional theory of recognition memory. *Memory and Cognition*, 2, 406–412.
- Baars, B. J., Motley, M. T., & MacKay, D. G. (1975). Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior*, 14, 382-391.
- Baddeley, A. (1990). *Human memory: Theory and practice*. Hove: Lawrence Erlbaum Associates.
- Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency, and negative recognition. *Quarterly Journal of Experimental Psychology*, 29, 461-473.
- Brown, R., & McNeill, D. (1966). The “tip of the tongue” phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5, 325-337.
- Burke, D., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language*, 30, 237-246.
- Butterworth, B. (1980). Some constraints on models of language production. In B. Butterworth (Ed.), *Language production, Vol. 1: Speech and talk* (pp. 423-459). London: Academic Press.
- Butterworth, B. (1982). Speech errors: Old data in search of new theories. In A. Cutler (Ed.), *Slips of the tongue*. Amsterdam: Mouton.
- Butterworth, B. (1989). Lexical access in speech production. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 108-135). Cambridge, MA: MIT Press.
- Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33A, 497-505.
- Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535-555). London: Academic Press.
- Dale, H. C. A., & Baddeley, A. D. (1962). On the nature of alternatives used in testing recognition memory. *Nature*, 196, 93-94.
- Fay, D., & Cutler, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry*, 8, 505-520.
- Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin company.
- Fritzen, J. (1975). Intralist repetition effects in free recall. *Journal of Experimental Psychology: Human Learning and Memory*, 1, 756-763.
- Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory and Cognition*, 13, 8-20.
- Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, 100, 546-567.
- Glanzer, M., Kim, K., & Adams, J. K. (1998). Response distribution as an explanation of the mirror effect. *Journal of Experimental Psychology: Human Learning and Memory*, 24, 633-644.
- Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 674-691.
- Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, 29, 228-244.
- Gregg, V. (1976). Word frequency, recognition, and recall. In J. Brown (Ed.), *Recognition and recall*. Chichester: John Wiley and Sons.
- Guttentag, R., & Carroll, D. (1998). Memorability judgements for high- and low-frequency words. *Memory and Cognition*, 26, 951-958.
- Harley, T. A. (1995). *The psychology of language*. Hove: Psychology Press.
- Harley, T. A., & Bown, H. E. (1998). What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology*, 89, 151-174.
- Hastie, R. (1975). Intralist repetition in free recall: Effects of frequency attribute recall instructions. *Journal of Experimental Psychology: Human Learning and Memory*, 1, 3-12.
- Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 824-843.
- Kintsch, W. (1970). *Learning, memory, and conceptual processes*. New York: John Wiley and Sons.
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Morton, J. (1979). Word recognition. In J. Morton & J. C. Marshall (Eds.), *Psycholinguistics series 2: Structures and processes*. London: Paul Elek.
- Murdock, B. B. (1998). The mirror effect and attention-likelihood theory: A reflective analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 524-534.
- Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 855-874.

- Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*, 273-281.
- Sternberg, S. (1966). High speed scanning in human memory. *Science*, *153*, 652-654.
- Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1379-1396.

# A Generative Connectionist Model of the Development of Rule Use in Children

Stuart Marcovitch (stuartm@psych.utoronto.ca)

Philip David Zelazo (zelazo@psych.utoronto.ca)

Department of Psychology, University of Toronto  
Toronto, ON, M5S 3G3, Canada

## Abstract

The cascade correlation algorithm (CASCOR), a generative connectionist model, was used to simulate age-related changes on the dimensional change card sort (DCCS), which has traditionally been used to evaluate the complexity of children's rule-use abilities. Like 2.5-year-olds, inexperienced networks behave as if following one rule; slightly more experienced networks (akin to 3-year-old children) behave as if following a pair of rules; and the most experienced networks (akin to 5-year-olds) behave as if following two pairs of rules. Analysis of the networks' activation levels revealed that mastery of simple rules is a necessary precondition for using higher order rules. The model also generated four novel predictions that can be tested in future research with children.

## Introduction

Since its inception, artificial intelligence has made a large impact on the field of psychology. The infusion of computer generated models into psychological research has become increasingly common. In the past decade, connectionist models have become particularly influential as a research tool in psychology. Connectionist models benefit psychology in three ways: (a) successful simulation requires formalization of the assumptions of the model, (b) analyzing the solution of a connectionist network may provide insight into the psychological mechanisms used, and (c) the model may generate novel (and often counter-intuitive) predictions. In particular, connectionist modeling used in conjunction with empirical research has the potential to shed light on patterns of development across a wide range of cognitive domains. Researchers in developmental psychology have already employed connectionist models to simulate developmental phenomenon in a variety of cognitive tasks (e.g., McClelland & Jenkins, 1991; Schultz, Schmidt, Buckingham, & Mareschal, 1995; see Elman et al., 1996, for a comprehensive review). Often, the results of these simulations call into question contemporary explanations of cognitive development.

According to Cognitive Complexity and Control theory (CCC; Frye, Zelazo, & Palfai, 1995; Zelazo & Frye, 1997), developmental improvements on tasks assessing deliberate reasoning and intentional action can be attributed to the acquisition of increasingly complex rule systems. Specifically, CCC postulates that young children (2.5 years) can use one rule, slightly older children (3 years) can use a pair of rules, while the oldest preschoolers (5 years) can use two incompatible pairs of rules. Rule-based card sorting paradigms have been employed to illustrate the number of

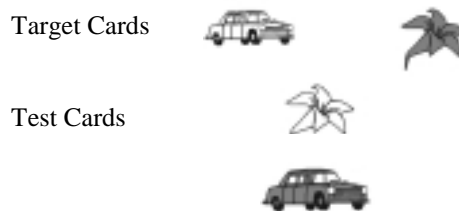


Figure 1: Stimuli for DCCS

rules that children can use. In these tasks, children are given cards that can be placed in one of two boxes based on a rule. For example, Zelazo, Reznick, & Piñon (1995) instructed 2.5-year-olds to sort pictures into categories such as things found inside the house versus things found outside. Typically, these children were able to sort the first card correctly, but then perseverated and sorted all subsequent cards in the same box. Thus, these results demonstrated that 2.5-year-old children could sort by one rule (e.g., if picture of things found inside the house then put card there), but not by a pair of rules (e.g., if picture of things found inside the house then put card here, but if picture of things found outside, then put card there).

The Dimensional Change Card Sort (DCCS; Frye et al., 1995; Zelazo, Frye, & Rapus, 1996) has also been used to reveal age-related changes in the number of rules children can use simultaneously. In the standard task, children are shown two *target cards* that differ on two dimensions, say color and shape (e.g., red car and blue flower). Children are presented with *test cards* that share one dimension with one target and the other dimension with the other target (e.g., red flower and blue car, see Figure 1). In the pre-switch phase, children are instructed to sort the test cards (i.e., match the test card to the appropriate target card) according to one rule (color or shape). After a predetermined number of pre-switch trials (e.g., 5, see Zelazo et al., 1996), children are asked to sort the same test cards by the other rule. So, the same test card will be sorted differently in the pre-switch and post-switch phases. On this task, 3-year-old children tend to pass the pre-switch phase, but fail the post-switch phase. This indicates that these children can sort by one pair of rules (e.g., in the color game, if it's red it goes here, but if it's blue it goes here), but not by two incompatible pairs of rules (e.g., if it's the color game, then if it's red it goes here, and if it's blue it goes here *but* if it's the shape game, then if it's a flower it goes here and if it's a car it goes here.) Five-year-old children tend to pass both the pre-switch and post-

switch phase, which illustrates that they can sort by two incompatible pairs of rules in the same context, and arguably requires the use of a higher order rule for selecting between pairs of rules.

The goal of the present study was to simulate the development of rule use in children using a generative connectionist model. Our study had three objectives: (a) to capture the age-related changes that are observed in children's sorting between the ages of 2.5 and 5 years, (b) to generate novel predictions, and (c) to explore what the internal structure of the connectionist networks reveals about the structuring of dimensions and features within the dimensions *vis à vis* success on the task.

In the present study, we used the cascade correlation learning algorithm (CASCOR; Fahlman & LeBiere, 1990) to simulate children's performance on the DCCS. Some researchers (e.g., Shultz, 1991) have suggested that CASCOR is appropriate in simulations of cognitive development because it embodies Piaget's principles of assimilation and accommodation. CASCOR is a generative algorithm that begins with connections between all the inputs and the output, but no hidden units. The model attempts to learn the training set in the constraints of this architecture, a phase akin to the Piagetian concept of assimilation. However, if the training set cannot be learned within a specific network architecture, hidden units are recruited as needed to increase computational power. Each hidden unit receives connections from all input units and all previously recruited hidden units. The restructuring of the network to create a more adaptive architecture is akin to the Piagetian concept of accommodation. One advantage of CASCOR is that the hidden unit chosen for recruitment is the one that will produce the lowest overall error. Consequently, the modified network is poised to solve the task at hand, and will do so more efficiently (using fewer hidden units) than networks with fixed architectures.

### Training Phase

Age-related changes in the DCCS were simulated using CASCOR. The networks had 15 inputs. The first input determined the game that was to be played (color or shape). The next 12 input units determined the color and shape of the stimulus cards. Each card was coded across 4 attribute units (red, blue, car, flower). A value of 1.0 indicated the presence of an attribute while a value of 0.0 indicated the absence of the attribute. For example, the values {1.0, 0.0, 0.0, 1.0} indicated a red flower. The test card and the two target cards were each represented by specific configurations across the 12 units. The 14<sup>th</sup> and 15<sup>th</sup> units were context units, which determined if the network was learning in the *training* context {1.0, 0.0} or the *test* context {0.0, 1.0}. These context units were necessary to distinguish learning that occurred in the natural environment (training) from the laboratory environment (test). There was one output unit that returned a value ranging from -0.5 to 0.5. Matching to the first target card was assigned an output value of -0.5, whereas matching to the second target card was assigned an output value of 0.5. The target value that

was closest to the actual output value was considered the matching target.

In the training set, the network received a set of simple rules. The network was presented with the relevant game (e.g., color), a bidimensional test card (e.g., red flower), and two bidimensional target cards (e.g., a red car and a blue flower). For all the examples in the training set, the context units were set to the training context (i.e., 1.0, 0.0).

The network updated its weights based on a supervised learning algorithm. The network's output was compared to the expected output (i.e., in the color game, a red flower should be matched to the red car), and the weights were updated using the quickprop algorithm (Fahlman, 1988) and batch learning (i.e., the weights were updated after each epoch, as opposed to each example). Quickprop is a weight adjustment algorithm that is much quicker than backprop because it uses second-order (curvature) information as well as first-order (slope) information when adjusting weights, whereas backprop is restricted to slope information. Slope information indicates the direction of change; curvature information provides an index of the change in slope, which is used to determine the magnitude of weight change (Mareschal & Shultz, 1996; also see Fahlman, 1988, for more details).

In the training phase of the simulation, all possible training combinations were used. That is, 2 games (color or shape) X 4 test cards (red flower, blue flower, red car, blue car) X 4 target combinations (red flower, blue flower, red car, blue car for target 'A'; target 'B' differed from target 'A' on both dimensions), which yielded 32 training examples. Because the preliminary goal was to simulate data that were averaged over groups of children, a cross-sectional design was implemented as per previous studies of the DCCS with children (e.g., Zelazo et al., 1996). Twenty networks were trained in each of 5 conditions that differed on the number of epochs of training that the network experienced. The conditions were 50, 75, 100, 150, and 225 epochs.

### Test Phase

After various amounts of exposure to the training set, training was halted so that the network could be tested. Testing consisted of changing the training set to five examples (pre-switch trials) that correspond to the five trials of the pre-switch phase of the DCCS. In all five trials, the network was presented with the same game (i.e., shape), the same two target cards (i.e., target 'A' was a red flower, target 'B' was a blue car), and the context nodes were set to the test context (i.e., 0.0, 1.0). The two possible test cards were presented (i.e., red car and blue flower) on alternate trials with one test card presented three times and the other test card presented twice. The network updated its connection weights after each pre-switch trial. After the fifth pre-switch trial, the network was tested on two post-switch trials. These were equivalent to the pre-switch trials, except now the network was asked to sort by the other dimension (e.g., color). The output revealed how the network sorted each of the two test cards. Because weights were *not*

updated in the post-switch phase, two post-switch trials were sufficient for the appropriate categorization of the network.

The network outputs were categorized into one of four categories based on criteria used with children (e.g., Zelazo et al., 1996):

- (1) Fail Pre-Switch - The network incorrectly sorted on two or more pre-switch trials.
- (2) Fail Post-Switch (same box) – The network passed the pre-switch phase, but incorrectly sorted on one of the two test trials in the post-switch phase (i.e., the network put all of the cards in the same box).
- (3) Fail Post-Switch (perseveratively) – The network passed the pre-switch but incorrectly sorted both test cards in the post-switch phase (i.e., the network perseverated on the two original rules).
- (4) Pass Post-Switch – The network correctly sorted both test cards in the post-switch phase.

## Results

The CASCOR network began with the 15 input units and the one output unit. Although the network did not initially contain hidden units, these were recruited as needed through the progression of the simulation. The number of hidden units recruited was noted. The number of networks in each of the four classifications is displayed in Table 1.

Table 1: Performance of CASCOR networks on DCCS

No. of Epochs	Categorization of Network			
	FPre	FPost Box	Fpost Pers	Pass
50	12 (1*)	2 (1*)	0	6 (1*)
75	5 (5*)	7 (7*)	5 (5*)	3 (2*)
100	10 (10*)	2 (2*)	5 (5*)	3 (3*)
150	2 (2**)	1 (1*)	4 (2*, 2**)	13 (3*, 10**)
225	1 (1**)	0	0	19 (19**)

*Note.* FPre = Fail Pre-Switch; Fpost Box = Fail Post-Switch (same box); FPost Pers = Fail Post-Switch (perseveratively); Pass = Pass Post-Switch. The number of hidden units recruited by the networks is represented by asterisks (\*). For example, 3\* means three networks recruited one hidden unit, while 10\*\* means 10 networks recruited 2 hidden units.

In the 50-epoch condition, 12 out of 20 (60%) of the networks failed the pre-switch phase. For the slightly more experienced network in the 75-epoch condition, 15 out of 20 (75%) of the networks passed the pre-switch phase. Furthermore, 12 out of 15 (80%) of those networks went on to fail the post-switch phase. In the 225-epoch condition, 19 out of 20 (95%) of the networks passed the pre-switch phase. All of those networks (100%) went on to pass the

post-switch phase. Overall, this pattern of results mirrored the pattern found in the empirical literature. Namely, the youngest children tend to fail the pre-switch phase, indicating failure to use a single pair of rules systematically. The slightly older children pass the pre-switch phase but fail the post-switch phase. Finally, the oldest children tend to pass both the pre-switch and post-switch phases, arguably indicating that they were capable of using a higher order rule for selecting between two incompatible pairs of rules.

Table 2: Number (and row percentages) of networks in each classification based on the number of hidden units.

No. of Hidden Units	Categorization of Network			
	FPre	FPost Box	Fpost Pers	Pass
0	11 (61%)	1 (6%)	0	6 (33%)
1	16 (33%)	11 (23%)	12 (25%)	9 (19%)
2	3 (9%)	0	2 (6%)	29 (85%)

*Note.* FPre = Fail Pre-Switch; Fpost Box = Fail Post-Switch (same box); FPost Pers = Fail Post-Switch (perseveratively); Pass = Pass Post-Switch.

The number of hidden units the network recruited seems to be related, albeit imperfectly, to performance on the DCCS. Table 2 displays the classification of networks across all five conditions based on the number of hidden units. A chi-squared analysis revealed a relation between the number of hidden units and the DCCS classification,  $\chi^2(6, N = 100) = 49.40, p < .01$ . The majority of networks with no hidden units fail the pre-switch phase, while the majority of networks with two hidden units pass both the pre-switch and post-switch phases. Networks with one hidden unit tend to be transitional and distributed across all four conditions. Thus, it can be argued that by acquiring more sophisticated internal representation (measured by the number of hidden units), more complex rules can be solved.

The current findings are congruent with Siegler's (1996) notion that cognitive development is driven by changes in strategy selection. According to this notion, children typically have a number of strategies available to them to solve any task. With age, the likelihood of selecting more appropriate strategies increases. However, even at older ages, children sometimes select inappropriate strategies. In the current simulations, increases in the number of hidden units may correspond to increases in the likelihood of selecting a more appropriate strategy. For example, networks with two hidden units usually adopt the most appropriate strategy (85% of the time), but occasionally adopt a less-appropriate strategy.

In addition to capturing the general pattern of age-related changes on the task, the simulations offer several predictions that raise interesting questions for future empirical work:

- (1) In networks that passed the pre-switch phase but failed the post-switch phase, there was a developmental increase in the proportion that failed perseveratively (as opposed to sorting cards in the same box). In the four network conditions where these types of errors occur, the proportions that failed perseveratively were 0%, 42%, 71%, and 80%, for 50, 75, 100, and 150 epochs respectively. We expect a similar increase with children.
- (2) The proportion of networks that passed the pre-switch phase followed a U-shaped developmental trajectory. The proportions in the network conditions were 40% at 50 epochs, 75% at 75 epochs, **50%** at 100 epochs, 90% at 150 epochs, and 95% at 225 epochs respectively. It is predicted that children will follow a similar U-shaped trajectory.
- (3) The unexpected decrease in the proportion of networks that pass the pre-switch phase occurred in the same condition (100 epochs) as when the networks began to fail the post-switch phase perseveratively as opposed to putting the cards in the same box. Arguably, this occurred because the networks are beginning to categorize both dimensions simultaneously. This will lead to a decrease in performance in the pre-switch phase (sorting is more likely to be based on the wrong dimension), and an increase in perseverative errors in the post-switch phase (more likely to sort the cards according to the dimension that was previously correct). It is predicted that careful analyses of children's performance will reveal similar trends.
- (4) Although 60% of the networks at 50 epochs failed the pre-switch phase, those that passed tended to pass the post-switch phase (6 out of 8, 75%). It is predicted that the youngest children (2.5-year-olds) who are able to pass the pre-switch phase will succeed in the post-switch phase. Perhaps these children have learned to sort a pair of rules, but fail to link the rules in the pre-switch to the rules in the post-switch. As a result, the post-switch phase is treated independently of the pre-switch phase, with a consequent absence of proactive interference.

### Analysis of Network Activations

A primary benefit of connectionist simulations to cognitive psychology is the ability to analyze the internal representations of the networks. To that end, cluster analyses were carried out on the activations of the hidden units and the output node in the networks for each of the training examples. Figure 2 displays graphically the results from the analysis of one randomly selected network in the 225-epoch condition<sup>1</sup> (i.e., after the network had learned to

sort successfully on both pre-switch and post-switch trials). Each training example is represented by a string of seven letters. The first letter denotes which game the network is required to play. The next six letters denote the test card, the first target and the second target respectively. Training examples that are clustered together elicit similar activation levels from the hidden units and the output. Because the features of the first target card necessarily determine the features of the second target card (e.g., red flower is always paired with blue car), only the first target card is discussed in the analysis.

As can be seen from Figure 2, group A contains all of the examples that have flowers both in the test card and in the target card. In contrast, all training examples that have cars in the test card and in the target card are in group B. Thus, the network appears first to discriminate, at least partially, on the basis of the shape dimension.

Group A (the flower group) can further be separated into 2 subgroups, C and D. Of all the test cards in group A, subgroup C contains all of the blue test cards, whereas subgroup D contains most of the red test cards (75%). Similarly, group B (the car group) can be further separated into subgroups E and F. Of all the test cards in group B, most of the blue test cards (80%) are in subgroup E, whereas most of the red test cards (75%) are in subgroup F. Therefore, once the shape dimension is established, the network appears to discriminate on the basis of color.

Correct performance on the DCCS requires more than successful categorization of the stimuli by the appropriate dimension. It is also necessary to categorize the stimuli by the type of game that is to be played. In Figure 2, all branches labeled G indicate the six places where this occurs. Based on the network's activation levels, we can speculate that success on the DCCS may first involve categorizing the stimuli by one dimension. Once this categorization has been established, the stimuli are then categorized by the other dimension. Only when both dimensions are appropriately categorized can a higher order rule that discriminates between the two dimensions, such as the type of game, be considered. This interpretation is consistent with CCC theory (Frye et al., 1995; Zelazo & Frye, 1997). For example, Zelazo (1999) suggested that success on the pre-switch phase of the DCCS requires the conjunction of two simple rules into a contrastive pair of rules. Each pair of rules must then be mastered before a higher order rule controlling their selection can be evoked. Without this higher order rule, children will select the rule that is most strongly associated with the given context (i.e., fail perseveratively on the post-switch phase).

### Conclusions

In conclusion, the CASCOR simulations were successful in its three goals. First, the age-related changes on the DCCS task were simulated. Namely, inexperienced networks failed the pre-switch phase, slightly more experienced networks

---

<sup>1</sup> Cluster analyses on less experienced nets revealed similar patterns as the 225-epoch condition. However, the results were

---

more variable. It appears that experience stabilizes the clustering structure.



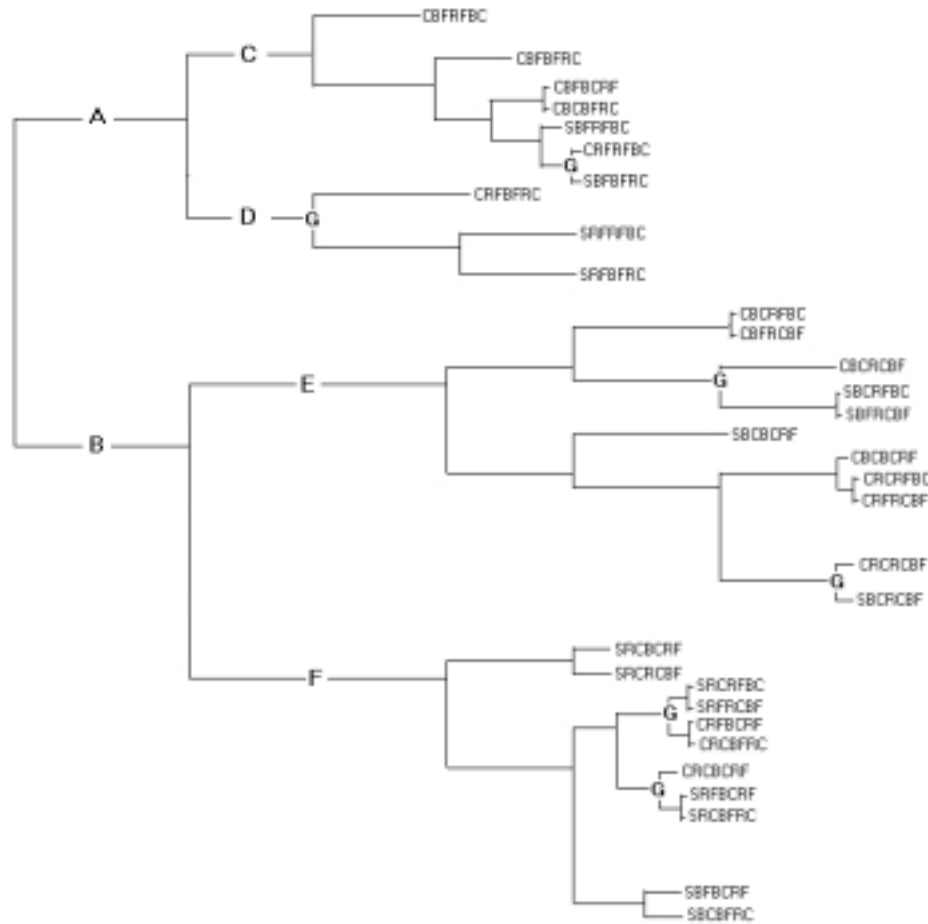


Figure 2: Cluster analysis on hidden and output unit activations of a randomly selected network in the 225-epoch condition.

passed the pre-switch phase but failed the post-switch phase and the most experienced networks passed both the pre-switch and post-switch phases. Second, novel predictions were generated and will be tested in future research. These include (1) an age-related increase in the number of children who fail the post-switch phase perseveratively (as opposed to sorting all the test cards in the same box), (2) a U-shaped developmental curve depicting performance on pre-switch trials, and (3) those very young children who pass the pre-switch phase will also pass the post-switch phase due to a relative lack of proactive interference. Third, cluster analyses on the hidden and output unit activations suggest that the formation of a higher order rule requires that the stimuli can be appropriately categorized by the appropriate dimensions. Further empirical research, coupled with modifications to modeling, hopefully will lead to an increased understanding of the mechanisms involved in the development of children's flexible rule use.

### Acknowledgements

This research was supported in part by a research grant from NSERC of Canada to Philip David Zelazo. We thank Ulrich Müller for providing constructive comments on an earlier draft of this manuscript.

### References

- Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.
- Fahlman, S. E. (1988). Faster-learning variations on back-propagation: An empirical study. In D. S. Touretzky, G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School*. Los Altos, CA: Morgan Kaufmann.
- Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in neural information processing systems*, Vol. 2. Los Altos, CA: Morgan Kaufmann.
- Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development*, 10, 483-527.
- Mareschal, D., & Shultz, T. R. (1996). Generative connectionist networks and constructivist cognitive development. *Cognitive Development*, 11, 571-603.
- McClelland, J. L., & Jenkins, E. (1991). Nature, nurture, and connections: Implications of connectionist models for cognitive development. In K. VanLehn (Ed.), *Architectures for intelligence*. Hillsdale, NJ: Erlbaum.

- Shultz, T. R. (1991). Simulating stages of human cognitive development with connectionist models. In L. Birnbaum & G. Collins (Eds.), *Machine learning: Proceedings of the eighth international workshop*. San Mateo, CA: Morgan Kaufmann.
- Shultz, T. R., Schmidt, W. C., Buckingham, D., & Mareschal, D. (1995). Modeling cognitive development with a generative connectionist algorithm. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling*. Hillsdale, NJ: Erlbaum.
- Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.
- Zelazo, P. D. (1999). Language, levels of consciousness, and the development of intentional action. In P. D. Zelazo, J. W. Astington, & D. R. Olson (Eds.), *Developing theories of intention*. Mahwah, NJ: Erlbaum.
- Zelazo, P. D., & Frye, D. (1997). Cognitive complexity and control: A theory of the development of deliberate reasoning and intentional action. In M. Stamenov (Ed.), *Language structure, discourse, and the access to consciousness*. Amsterdam & Philadelphia: John Benjamins.
- Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development, 11*, 37-63.
- Zelazo, P. D., Reznick, J. S., & Piñon, D. E. (1995). Response control and the execution of verbal rules. *Developmental Psychology, 31*, 508-517.

## Transfer Along a Continuum: Differentiation or Association?

**I.P.L. McLaren** (iplm2@cus.cam.ac.uk)

Department of Experimental Psychology; Downing Street  
Cambridge, CB2 3EB UK

**M. Suret** (mbs22@cam.ac.uk)

Department of Experimental Psychology; Downing Street  
Cambridge, CB2 3EB UK

### Abstract

There has been a revival of interest in the question of the optimal training schedule for a difficult discrimination. McClelland (personal communication) argues that the optimal schedule is one which starts with a much easier discrimination on the same dimension as the difficult one, arranged so that the easy problem can be made to gradually converge on the more difficult one. He further argues, in agreement with Saksida (1999), that the reason for this is that representations are more easily formed during acquisition of the easy problem, which can then be put to use in solving the difficult discrimination. As associative learning theorists we are more familiar with another account - that initiated by Lawrence (1952) - which agrees that the optimal schedule is one which employs a strategy of transfer along a continuum. Where the accounts differ is in the mechanism for transfer; rather than appealing to representation formation, this account explains the benefits of training on the easier problem in terms of the dimensional features / elements that acquire associative strength and their ability to generalize appropriately to the hard problem. In this paper we report experiments that attempt to distinguish between these two accounts by manipulating / deconfounding stimulus exposure and training. We demonstrate the basic effect, and show that pre-exposure to the stimuli that comprise the easy problem is less effective than pre-exposure to the stimuli that make up the more difficult discrimination. Our conclusion is that this latter result is not what one would predict from the non-associative account given above, but that it fits well with McLaren, Kaye, and Mackintosh's model of perceptual and associative learning.

### Introduction

Lawrence (1952) demonstrated that it was possible for training on an easy perceptual discrimination to transfer to a more difficult problem on the same dimension. An example of the type of problem that he studied would be a brightness discrimination between two rectangles in similar shades of gray (hard version) or black and white rectangles (easy version). Groups of pigeons trained on these problems for the same number of trials can reach a point where the group trained on the easy problem have solved it, i.e. they have learned to peck at one rectangle for grain and to ignore the other, whereas the group trained on the hard problem have made little progress. If both groups are now trained on the hard problem for a further number

of trials, i.e. the group previously trained on the easy version of the problem are now switched to the harder version, then the result of interest is that the group switched from easy to hard acquires the hard problem much more rapidly than the group trained on the hard problem from the outset. This result holds despite the fact that the total amount of training is the same for both groups, and that the group that acquires the problem more slowly is the one that has received more training on that specific problem. This is the phenomenon of transfer along a continuum (TAC), and is the subject of the research reported in this paper.

The standard associative account of this phenomenon appeals to the notion of generalization. The stimuli for the easy problem become associated with reward and non-reward respectively, and then generalize to the stimuli for the hard problem (e.g. see Mackintosh, 1983). This is more effective than training on the hard problem itself because it is so difficult to learn, which is taken to be because the stimuli are so similar to one another. Figure 1 can be used to illustrate one possible instantiation of this explanation. On this approach, a stimulus is represented by a set of activated elements or units, a distributed representation. Variation along a stimulus dimension such as brightness will, for the most part, be represented by different elements corresponding to different values on the dimension, rather than the activation level of an individual element being the primary indicator of value on the dimension (c.f. chapter by Thompson in Mostofsky, 1965). Each element has a 'tuning curve' such that it responds most strongly to a certain value on the dimension and this response drops off fairly rapidly with 'distance' from this optimal value. Note that many elements will be active when any stimulus with value on that dimension is present, the coding is via a pattern of activation. Learning will proceed via association between the elements activated by a stimulus and other units representing reward. We are now in a position to explain Lawrence's results. In the case of the easy problem (shown top in the figure) the stimuli are well separated on the dimension and there is relatively little overlap between the patterns of activation that represent them. Learning proceeds rapidly, favoring those elements which are most active on a trial, and there is little generalization between stimuli to slow acquisition of the problem. In the case of the hard problem (shown bottom in the figure)

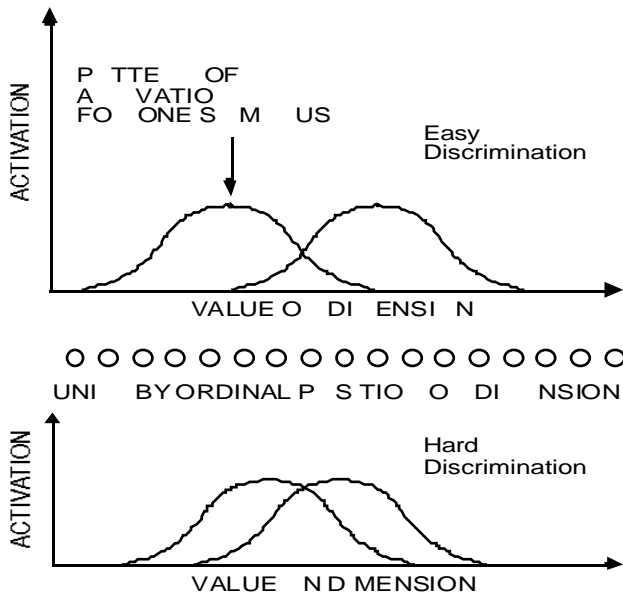


Figure 1: Stimulus representation on a dimension for easy (top) and hard (bottom) discriminations.

the situation is somewhat different, in that the large degree of overlap between the stimulus representations results in considerable generalization between the stimuli, and this is what makes the discrimination difficult. The elements that are most active, and so dominate learning, are not those that best discriminate between the stimuli. As a result acquisition is slow. Now consider the case where the easy problem is first acquired, and then the subjects are switched on to the hard problem. The training on the easy problem will result in exactly the elements that are the most predictive of reward or non-reward in the hard problem gaining considerable associative strength because they are highly activated by the easy stimuli. Thus the learning will transfer well to the hard problem, and will be more than an equivalent amount of training on the hard problem would have provided (because of the relatively large activations of the elements concerned).

There is another tradition in psychology, however, that appeals to quite different, non-associative processes to explain the phenomenon of transfer along a continuum. It can be traced back at least as far as the work of Eleanor Gibson (1969), who conjectured that a process of differentiation, contingent on exposure to the stimuli in question, resulted in representations of the stimuli that better enabled discrimination between them. Gibson's thesis is perhaps most naturally captured in terms of competitive learning coding schemes that require no explicit instruction to develop representations that capture the structure of a stimulus set that they are exposed to. Our example of such a system is that due to Saksida (1999), which is explicitly designed to deal with phenomena of the kind under consideration here. Figure 2 allows us to contrast Saksida's model with the standard associative account. Instead of stimulus elements being directly associated with reward representations (shown top), there is a non-associative pre-processor prior to association to reward representations (shown bottom in the

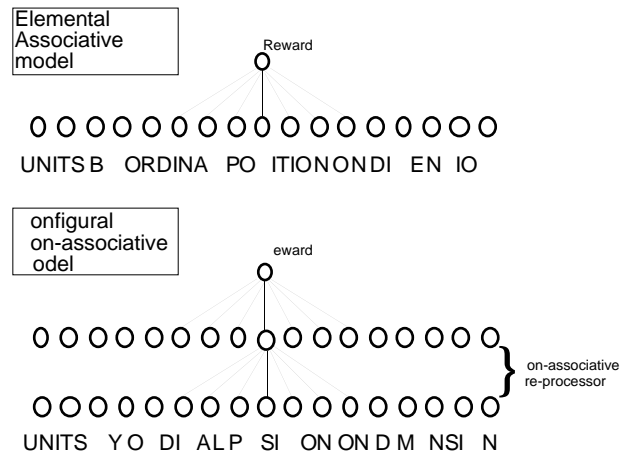


Figure 2: Architecture for associative and non-associative models of TAC.

figure). The model develops a representation of the input at this intermediate, competitive layer, and, in Saksida's model, it does so in a way that drives initially overlapping representations for two stimuli apart so that they become more discriminable (i.e. differentiation). The interesting possibility raised by such a model is that the explanation for TAC might be quite different to that generated by the associative account given earlier. Instead of appealing to generalization of associative strength from the easy problem to the hard version, it could be that training on the relatively easy problem could develop a coding at the competitive layer that meant that the hard problem was no longer as difficult as would have been the case. Whereas before the hard problem would have (initially at least) given rise to highly overlapping patterns of activation at the competitive layer, now the patterns of activation are better separated because of the coding scheme developed for the dimension whilst solving the easier problem. In a sense what happens is that the process of developing discriminable representations for the easier problem drags apart the representations for the harder problem as well. Saksida herself is quite definite on this..."One clear prediction of the current model is that exposure to a pair of similar stimuli will facilitate discrimination of stimuli that are even more similar along the same dimension." and..."pre-exposure to two stimuli will facilitate discrimination of other stimuli whose representations fall between them on the competitive layer" Saksida (1999). The only provisos being that the easier discrimination should itself employ relatively similar stimuli, and should be studied long enough for the stimuli to become discriminable (i.e. for the competitive layer to develop the necessary representations).

The strategy adopted in this paper is to contrast these two accounts with specific reference to the issue of whether or not TAC is best characterized as due to elementally-based generalization or rather to perceptual

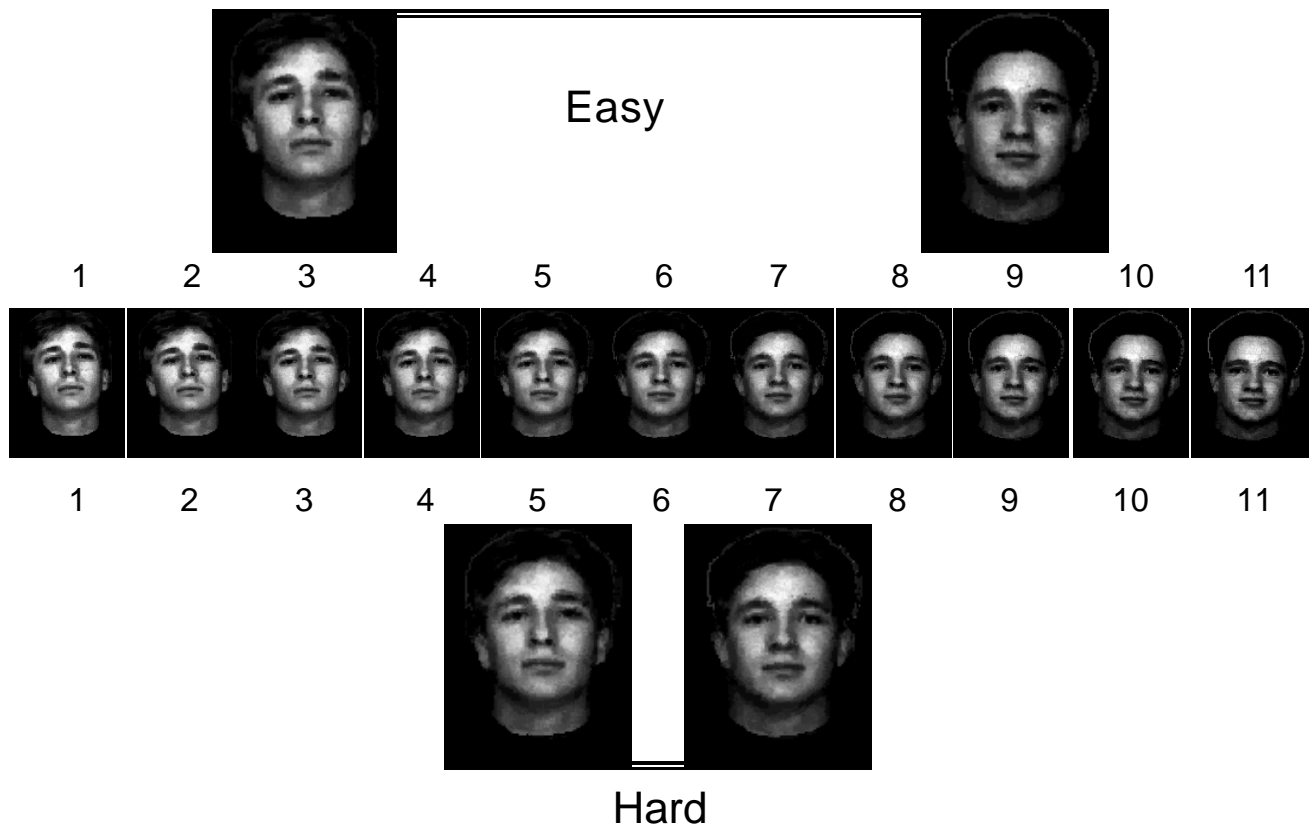


Figure 3: One of the four morphed face dimensions used in these experiments.

differentiation as a result of representation formation. Experiment 1 demonstrates TAC using stimuli that meet Saksida's criteria for her models application. Experiment 2 then assesses whether the effect can be explained predominantly in terms of representation development by looking at the effects of pre-exposure to the stimuli used in the easy and hard problems. The logic here is that if TAC is mainly due to representation development, then the training phase is effectively equivalent to a pre-exposure phase, and so explicit pre-exposure should generate the same pattern of results.

### Experiment 1

#### Stimuli and Apparatus

In all phases of the experiment, pictures of faces were shown in the form of gray-scale images. These had been created from standard passport photographs of university undergraduates, which had been scanned into the computer. These stimuli were presented on an Apple Macintosh computer running Microsoft Basic. They were 3.5 cm by 4.5 cm and subjects sat approximately 50 cm from the screen. The face stimuli for this experiment were constructed by taking pairs of faces and morphing from one to the other in 10 equal steps, giving a dimension with 11 values in all. The faces in a given pair are chosen to be similar (which aids the morphing process in keeping the transitions smooth) so that neighboring stimuli on the dimension are very similar indeed. Figure 3 illustrates the morphed face dimension for one pair of faces, there were

four pairs of faces in total and the faces at 3 and 9 on the dimension always constituted the easy problem and those at 5 and 7 the hard version. All four dimensions were used concurrently for every subject, with the assignment of the face dimensions to the conditions of the experiment counterbalanced appropriately. Pilot testing revealed that the discriminations were difficult (even for the 3 vs. 9 case) but possible under the conditions of this experiment, and subjects reported that their performance was hard to characterize in terms of rules based on features (desirable if performance is to be associatively driven).

#### Subjects and Design

Subjects were 40 Cambridge undergraduates and graduates with an age range of 18 - 30. They were randomly assigned to two equal groups, one of which (Group Easy) was pre-trained on the easy problem for all four dimensions concurrently for a fixed number of trials (40 trials in total, five for each face), the other (Group Hard) was pre-trained on the hard version for all four dimensions for an equal number of trials. After the pre-training phase both groups were then trained on the hard problem for all four dimensions concurrently (again 40 trials in total, five for each individual face). This was followed by a final test phase in which performance on the hard problem for each dimension was assessed without giving the feedback used in pre-training and training. In this phase each face is also shown 5 times. The data of interest are the responses to the stimuli in this final test phase. If the discrimination between 5 and 7 is better

learnt after pre-training on 3 and 9 then this would be evidence of TAC.

## Procedure

In both the pre-training and training phases of the experiment, subjects were told that once they pressed the space bar, a constant stream of stimuli in the form of faces would appear on the screen, and that their task was to sort these stimuli into two categories. They were to do this by pressing one of two keys ('x' on the left or '.' on the right) and would receive immediate feedback as to the correctness of the response. If they did not respond within a few seconds (4.25 sec) they would be timed out. The subjects were told that the faces were randomly and equally allocated to either left or right key and that their task was to simply find out and remember which ones were 'right' and which ones were 'left'. Once the subject initiated the experiment, trials were continuous. Stimuli were presented singly, and each trial started with a '+' for 0.7 sec which was then replaced by a rectangular frame for 0.2 sec. Each face appeared and stayed in the screen for a maximum of 4.25 sec and disappeared once a response or time-out was made. Feedback was then given for 1 sec, either 'correct' displayed in the center of the screen or 'error' and a beep if the wrong key was chosen.

After they had completed the pre-training and training phases, subjects progressed to the test phase of the experiment. Subjects were told to categorize the stimuli into the two categories based on the judgments they had made in the training phase. So, if a face had been 'a left key stimulus' in the training phase, it was to be allocated again to the 'left key' category in the testing phase. This time no feedback was given. The procedure of stimulus presentation was as before with the exception that feedback was replaced in this phase by a 1 sec pause between the subject's response and the proceeding stimulus.

## Results

The results of Experiment 1 are shown in Figure 4. One key, e.g. the left key, is designated the negative category (a press scores -0.5 for that stimulus) and the other right key the positive category (scores +0.5) during test.

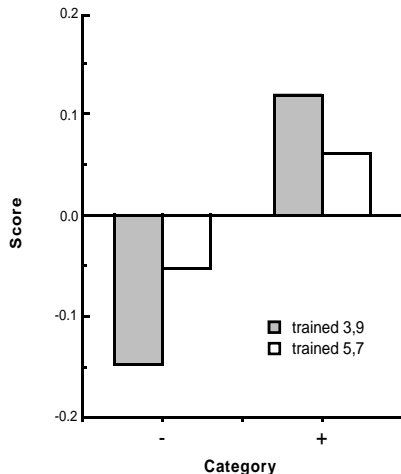


Figure 4: Results for Experiment 1.

Key assignments were counterbalanced across subjects so that the positive category has equal numbers of left and right key responses (at least by design). The test score indicates the average of the key presses across subjects, and would be zero if subjects were indifferent to which stimulus went with a given key, and ranges from +0.5 to -0.5. The group pre-trained on the easy problem (3,9) shows much better performance on the hard discrimination than the group pre-trained on that discrimination (5,7) itself. That is, the 3,9 group has more positive scores for its positive stimulus on test, and more negative scores for its negative stimulus.

These impressions are borne out by statistical analysis, in which all probabilities are two-tail unless otherwise specified. ANOVA on the results with a between subjects factor of type of pre-training (3,9, vs. 5,7) and a within subjects factor of type of stimulus (- vs. +) gave an  $F(1,38) = 27.75$ ,  $p < .001$  for the main effect of type of stimulus and  $F(1,38) = 4.39$ ,  $p < .05$  for the interaction between the two factors. The first effect refers to the fact that the positive stimulus is, overall, given a more positive score than its negative counterpart, the interaction reveals that the difference in score between positive and negative stimuli was significantly greater for the 3,9 group who were pre-trained on the easy problem. This demonstrates transfer along a continuum with these stimuli. Planned comparisons on the positive and negative stimuli for each group separately reveal that both groups are significantly better than chance on the test discrimination,  $F(1,19) = 24.9$  and  $5.52$ , both  $p < .05$ . Thus both groups can be said to have learned the discrimination.

## Discussion

Experiment 1 provides a convincing demonstration of transfer along a continuum in human subjects using an artificial dimension constructed by morphing between similar faces. Performance on the hard problem after pre-training on the easy problem is much better than if pre-training had been on the hard problem used during training and test. Nevertheless, both groups were able to acquire the discrimination under the conditions of the experiment.

We are now in a position to ask if this TAC effect is simply due to exposure to the stimuli used in the easy problem, or if instead it requires that subjects be trained on the easy problem for the effect to occur. Experiment 2 seeks to answer this question by pre-exposing subjects to the stimuli of either the easy or hard problem instead of pre-training them.

## Experiment 2

In this experiment the stimuli are the same as in Experiment 1, and two new groups of 20 subjects from the same population are assigned to two different pre-exposure conditions. These are equivalent to the pre-training conditions of Experiment 1 except that a) no response is required as it is pre-exposure and b) each stimulus is shown for a fixed duration of 2 sec. This

duration was chosen to ensure that subjects in this experiment received the same or greater total time of exposure to the stimuli compared to all the subjects in Experiment 1. Thus subjects were pre-exposed to the stimuli that constituted either the easy or the hard problem, then trained on the hard problem exactly as in Experiment 1, then tested exactly as in Experiment 1. If the results of Experiment 1 were predominantly due to exposure to the stimuli of a given problem, then the results of this experiment should resemble those of the previous experiment. If, on the other hand, they were strongly dependent on the training element during pre-training then we might expect the results to differ in that evidence for any TAC effect should disappear.

## Results

The results of Experiment 2 are shown in Figure 5.

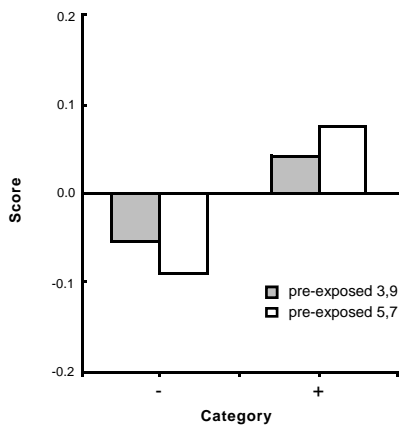


Figure 5: The results of Experiment 2.

Once again an ANOVA with one between factor of type of pre-exposure (3,9, vs. 5,7) and one within factor of stimulus type (5 vs. 7) was conducted which gave an overall main effect of stimulus type,  $F(1,38) = 9.84$ ,  $p < .005$ , but no significant interaction between the two factors ( $F < 1$ ). Thus there is good evidence for acquisition of the discrimination, but no significant evidence that pre-exposure to either the easy or hard problem stimuli had any differential effect. Contrary to expectations on a differentiation account of TAC, the group pre-exposed to the hard problem stimuli was actually numerically better on test. Planned comparisons on the two groups revealed that the group pre-exposed to the hard problem was significantly better than chance on test,  $F(1,19) = 7.26$ ,  $p < .05$ , whilst the other group was only marginal,  $F(1,19) = 2.89$ ,  $p(1\text{-tail}) = .052$ .

As the two experiments are highly comparable in their stimuli, apparatus, procedures and subject populations we can compare them in a single analysis. When this is done there is a three way interaction ((pre-trained vs. pre-exposed) x (problem, 3,9 vs. 5,7) x (stimulus on test, 5 vs. 7) that indicates that the effect of pre-exposure in Experiment 2 is significantly different to the effect of pre-

training in Experiment 1. Finally, pre-exposure to the easy problem in Experiment 2 was significantly less effective than pre-training on the easy problem in Experiment 1  $F(1,38) = 4.83$ ,  $p < .05$ . This is despite the fact that pre-exposure in Experiment 2 was at the maximum level observed in Experiment 1 (where the fluctuations were due to different speeds of response during pre-training).

## Discussion

The results of Experiment 2, taken in conjunction with those of Experiment 1, do not support a differentiation account of transfer along a continuum. The effect of pre-exposure to the problem stimuli is seen to produce the converse pattern of results to pre-training, that is, with regard to learning the hard problem during the training phase, pre-exposure to the easy problem is less effective than pre-exposure to the hard problem, whereas pre-training on the easy problem is much more effective than pre-training on the hard problem. This pattern of results strongly suggests that the advantage that accrues as a result of pre-training on the easy problem is due to generalization of the associations acquired during that pre-training.

One loose end in this experiment concerns the extent to which pre-exposure can be said to have an effect at all, given that the two groups do not differ significantly. Some light can be cast on this issue by considering the data from a previous series of experiments (McLaren, 1997) which used the same procedures and stimuli, but merely trained the face discrimination (as here) but without any pre-training or pre-exposure. Under these circumstances the hard problem was not solved (mean difference between the positive and negative stimuli was only .037,  $F < 1$ ), and a comparison between these results and those of Experiment 2 reveals that the group pre-exposed to the hard problem is better than the group simply trained on the hard problem,  $F(1,43) = 2.64$ ,  $p(\text{one-tail}) = 0.055$ . This means that pre-exposure to the hard problem has had a near significant beneficial effect (i.e. we have some evidence for perceptual learning), though this is not true of pre-exposure to the easy problem ( $F < 1$ ).

## General Discussion

In this paper we have contrasted two classes of model of TAC, associative and non-associative versions. We should make it clear that while we have found no evidence that supports the non-associative account relative to the associative version, nor do our results falsify the non-associative position adopted by Saksida and others. What is needed to rescue this account is a parametrisation of the non-associative model that allows the generalization from pre-training to dominate any effects of representation formation and differentiation. In these circumstances the two types of models would in some sense be different instantiations of the same psychological theory.

Taking Saksida's account first: the perceptual learning effect seen in Experiment 2 would be due to representation

formation and differentiation, as the initially overlapping patterns of activation for the two to-be-discriminated stimuli were pushed apart by competitive learning, and the TAC effect would be due to generalization associations to reward and non-reward formed during pre-training to training and testing. The model would be constrained so that the latter effect would be stronger than the former, which is not a typical feature of this class of model. The more natural account would attribute the difficulty of the hard problem to the need to establish well differentiated representations of the stimuli, a process that was aided by training on the easy problem. It may well be that the need both to model TAC and perceptual learning within this class of model may impose unsustainable constraints on its ability to function effectively, but this is a question for future research.

The associative account offered here follows McLaren *et al's* (1989) theory of association and representation. The explanation of TAC is the standard associative account given earlier in this paper, but the explanation of the perceptual learning effect seen in Experiment 2 may bear further exposition. On this theory, exposure to two similar stimuli that will be represented as overlapping patterns of activation results in a decrease in the salience (in this case this can be understood as the degree of activation) of the elements representing those stimuli. This occurs to the extent that they become predicted by associations from other elements. The reduction in salience will be greatest for the elements shared by the two stimuli (the overlap) because they are encountered, and hence engage in learning, twice as often as the elements unique to either stimulus. The effect is that the elements that make the discrimination difficult (because they are shared by the stimuli and lead to generalization between them) become relatively less salient than those that enable discrimination between the stimuli. This consequence of pre-exposure leads to the discrimination between the stimuli becoming easier, as the distinctive features (represented by the unique elements) of each stimulus are now able to preferentially engage in learning. The result is perceptual learning, in that the discrimination is learned faster after pre-exposure. The effect is predicted to be greater for more similar stimuli, which fits well with the greater pre-exposure benefit for the harder problem. This is because the more similar stimuli are taken to have a higher proportion of shared elements, and so the effect of reducing these elements' salience relative to the unique, distinctive elements of the stimuli is proportionately greater.

## Conclusion

Associative theories of representation development and learning are adequate to model the transfer along a continuum effect reported here. Non-associative theories that appeal to competitive learning or some other mechanism for representation formation are probably able to instantiate the same psychological theory, but offer nothing new in modelling these data. The challenge is to find data that require this type of theory rather than an associative account.

## Acknowledgments

This research was supported by an ESRC grant to IPL McLaren

## References

- Gibson, EJ (1969). *Principles of perceptual learning and development*. New York: Appleton-Century Crofts.
- Lawrence, D.H. 1952. The transfer of a discrimination along a continuum. *Journal of Comparative and Physiological Psychology*, 45, 511-516.
- McLaren, I.P.L. 1997. Peak-Shift on a Novel Stimulus Dimension. Paper given to the Conference on Associative Learning at Gregynog, Wales, UK.
- McLaren, I.P.L., Kaye, H. and Mackintosh, N.J. 1989. An associative theory of the representation of stimuli: applications to perceptual learning and latent inhibition. In R.G.M. Morris (Ed.) *Parallel Distributed Processing - Implications for Psychology and Neurobiology*. Oxford. OUP.
- Mackintosh, NJ (1983). *Conditioning and associative learning*. Oxford: Oxford University Press.
- Saksida, LM. 1999. Effects of similarity and experience on discrimination learning: A nonassociative connectionist model of perceptual learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 25, 308-23.
- Thompson, RF (1965). The neural basis of stimulus generalization. In DI Mostofsky (Ed.). *Stimulus Generalization*, pp.154-178. Stanford: Stanford University Press.



# Regularity and Irregularity in French Inflectional Morphology

**Fanny Meunier** (fanny.meunier@mrc-cbu.cam.ac.uk)

Medical Research Council Cognition and Brain Sciences Unit;  
Cambridge, UK

**William Marslen-Wilson** (william.marslen-wilson@mrc-cbu.cam.ac.uk)

Medical Research Council Cognition and Brain Sciences Unit;  
Cambridge, UK

## Abstract

Can regular and irregular verb forms be accommodated by a single representational mechanism or is a dual mechanism account required? In a first experiment, we used a cross-modal repetition priming paradigm to investigate the mental representation of regular and irregular verb forms in French. Subjects heard a spoken prime (such as aimons) immediately followed by lexical decision to a visual probe (such as aimer). We contrasted four types of French verbs, varying in the phonological and morphological regularity of their verb form inflection. These were (i) regular verbs (aimons/aimer) (ii) verbs that undergo predictable phonological changes (sèment/semmer) (iii) verbs to which sub-rules apply (teignent/teindre) and (iv) irregular verbs with idiosyncratic alternations (vont/aller). The infinitive forms of these verbs were presented as target in three prime conditions: preceded either by a regular form, an irregular/modified form (except for the regular verbs), or a control unrelated prime. Morphologically related primes, whether regular or irregular, significantly facilitated lexical decision responses for all four verb classes. The same pattern of results was observed in a second experiment using a masked priming paradigm. These results contrasted with English, where regularly inflected verbs prime their stems but irregular verbs do not. We argue that the pattern observed in French reflects the decomposability of French irregular forms.

## Introduction

Psycholinguistic models have proposed a distinction between information that can be obtained through rules and information that must be recalled from a list. On the one hand, distributed approaches argue for a single mechanism underlying the representation and processing of both regular and irregular items (Plunkett & Marchman, 1993) and, on the other hand, symbolic approaches argue for a dual mechanism account, where regular forms are generated by rule but irregular forms are stored as rote-learned whole forms (e.g. Pinker, 1991). Many studies tackle this issue by trying to determine whether the co-occurrence of regular and irregular verb forms in a given language can be accommodated by a single representational mechanism or whether a dual mechanism account is required.

In English, verbs have only three types of morphological processing contexts: 3rd person singular, past tense and progressive forms (jumps, jumping, jumped). This inflec-

tional system offers a sharp contrast between a single, dominant, regular process of past-tense formation (adding the regular affix -ed to an unchanged stem) and a small, heterogeneous group of irregular past-tense forms (mostly of an idiosyncratic nature).

Several sources of evidence suggest that the linguistic differences between regular and irregular forms lead to differences in the way these forms are represented in the English mental lexicon. A major source of evidence is research using repetition priming tasks, where a test word is preceded by a related prime word. The target word walk, for example, is preceded either by a morphologically related word (e.g. walked), or an unrelated word (e.g. goal). Previous research done in English shows diminished or absent priming between irregular tense and the stem (drove/drive) versus a strong priming effect between regular pairs such as walked/walk (Kempey & Morton, 1982; Napps, 1989; Stanners, Neiser, Herson & Hall, 1979). Pinker (1991) claimed that these results support the dual mechanism dichotomy. Convergent results have been observed using the cross-modal paradigm, where the prime is presented auditorily (Marslen-Wilson, Hare & Older, 1995). Again significant priming is only observed for regular inflected forms (such as walked/walk) and not for irregular ones (such as dug/dig).

In this framework, priming is explained as reflecting the fact that regular forms share a representation with their stem, and both inflected and non-inflected forms of a given verb map directly onto the representation of the stem at the level of the lexical entry. The morphological priming effect results from the repeated activation of the same morpheme by prime and target. On the contrary, an irregular form will have a separate form representation from the stem to which it is related and this may lead to a reduction of priming between the two items, under specific testing conditions. This may be due either to competition between the two representations (stem and irregular form) or as a consequence of the blocking function assigned to the listed irregular form (the presence of a lexical entry for the irregular form will prevent the application of the default suffix).

One problem with English, however, as a basis for generalisations about regularity and irregularity, is that the English past-tense forms do not differ simply in regularity, but also along a number of dimensions, including contrasts in

basic morphological procedure (suffixation versus stem change), the absence versus presence of phonological constraints on morphological processes, and high versus low type frequency of classes of past forms. In order to disentangle potential evidence about the general properties of morphological systems from the possible idiosyncrasies of English past tense formation, it is necessary to conduct parallel experiments in other languages which exhibit comparable but cleaner contrasts between regular and irregular procedures.

One language that we have looked at already in this light is Italian. This is a much richer inflected language (with many different types of tense and person suffixes) where there are a number of irregular past-tense forms that obey similar criteria for irregularity as the English irregulars, but where they occur in a morphologically more structured and phonologically more predictable linguistic environment. Using a cross-modal priming paradigm, Orsolini and Marslen-Wilson (1997) observed the same amount of priming when the prime was regular and when it was irregular. They suggest a possible account that attempts to capture the sub-regularities of the verb forms through an explicit system of rules rather than relying on an analogical network to represent them implicitly.

Here we report an extension of this research to French, which, like Italian, has a richer inflectional system than English, and which allows us to explore a wider range of types of irregularity. In French, verbs are organised into three basic morphological classes, called conjugations. These distinctions use as first criteria the infinitive form and as second the imperfect form. The major class is conjugation 1, containing verbs with infinitives ending in *-er* (such as aimer, voler...). This is the most productive class and fully regular. Conjugation 2 is formed by verbs that have an infinitive in *-ir* and imperfect in *-iss-* (such as finir, salir...). It is a smaller class than conjugation 1 and it is no longer productive, but it is fully regular. Conjugation 3 contains verbs with infinitives ending in *-ir* (and that do not have an imperfect in *-iss-*), *-oir*, *-re* (such as dormir, boire, peindre...) and the verb aller. Verbs contained in this group are highly irregular.

In our experiment we used four types of verbs. The first condition was wholly regular verbs from the first conjugation such as aimer; the second condition was regular verbs from the first conjugation but that in a few forms have a phonologically triggered surface change, such as amener-amène. These types of phonological changes (reflecting a high/low alternation) are also observed in the case of gender marking (fermier-fermière). We will call this condition the morphophonological constraint group. The third group consisted of irregular verbs from conjugation 3 but where the irregularities were common to at least 10 verbs such as teindre-teignent, peindre-peignent. Verbs in this group are closest to the ones used in the Italian experiment; we will refer to it as the sub-regularity group. The fourth group, more similar to the type of irregularity found in English was made up of highly idiosyncratic suppletive alternations such as aller-vont.

If the patterns of results observed in English and in Italian are not language specific but are due to the type of irregu-

larities then in French we should observe the same amount of priming when the prime is regular and in the morphophonological and sub-regularity irregular conditions. On the contrary, when the prime is an idiosyncratic form, we may, as in English, observe no or less priming than with a regular form. Priming effects in this experiment are evaluated by comparing reaction times when the prime is related versus unrelated to the target, and also by comparing response latencies when the related prime is regular and when it is irregular.

## Experiment 1

### Method

**Material and Design** We used a cross-modal paradigm. The prime was auditorily presented and immediately followed by a visual presentation of the target-item. Subjects made a lexical decision response to the visual target, which was preceded by a regular or irregular related or unrelated prime.

Ninety-six verbs falling in four categories were selected, as described earlier, and examples are listed in Table 1 below. We used as the target the infinitive form of the verb. We chose for each verb of each category, three types of prime (verb forms): A regular form, an irregular form and a control (or baseline) word matched on the regular form. To keep the design balanced, regular verb targets were preceded by two different regular targets. Targets were between 4 and 11 letter long.

Table 1: Examples of stimuli.

Verb Type	Infinitive Target	Forms Regular	Irregular
Regular	<i>aimer</i>	<i>aimerons</i> <i>aimons</i>	<i>n/a</i>
Morpho-phonologic constraints	<i>semer</i>	<i>semons</i>	<i>sème</i>
Sub regularity	<i>teindre</i>	<i>teindra</i>	<i>teignent</i>
Idiosyncratic	<i>aller</i>	<i>allons</i>	<i>iront</i>

For each of the 96 regular primes, we selected a control word that was matched to the regular experimental prime for surface frequency, number of syllables and tense and person of the verb form. None of the neutral condition words were morphologically, semantically or phonologically related to the target. We also constructed filler pairs in order to reduce the proportion of related pairs within the list. We added 64 pairs in which the target was a word (such as calculons/partir), and 160 pairs in which the target was a non-word (such as marchera/enteler). Each prime list was composed of 96 experimental words (of which 64 were related to the target and 32 were not), 64 words with an unrelated target word, 160 words with a nonword target (64

pairs in which prime and target shared formal features and 96 primes followed by a nonword target which was unrelated). To sum up, we had 160 word-word pairs and 160 word-nonword pairs.

In order to avoid the repetition of a given target for a subject, we constructed 3 experimental lists of 320 items each. A given target appeared only once in each list: with a regular related prime in one list, an irregular related prime in the second list and a control prime in the third one. In each list, 2/3 of the experimental prime-target pairs were morphologically related (64 pairs). The number of pairs of each experimental condition was equal (8) in each list. Each subject heard only one list so that each saw a third of the items with a regular related prime, a third with an irregular related prime and a third with a control prime. The list of targets was the same for all subjects, only prime lists varied. To give a break to the subjects we split up each list. Experimental pairs of each condition were equally distributed in each segment of the list. Each part of the list started with 10 items that were not experimental ones. Before starting to hear the list itself, the subject had training with 20 prime-target pairs. The experimental session lasted 25 minutes.

**Procedure** A French female native speaker recorded primes on a DAT. Each prime was then digitized at a rate of 22kHz and stored on computer hard disk. Each word was isolated in a single independent file. This allowed us to control the time between the end of the prime and the presentation of the target. The prime was binaurally presented to the subject and was immediately followed (ISI 0ms) by the presentation of the target. This latter was written on a CRT screen in front of the subject. The target stayed on the screen until the subject made a response. The task of the subject was to push one of the two buttons on a response box (one for word, the other for non-word), as fast as he or she could. Subjects were alone in the testing room.

**Participants** Thirty-six students of Psychology at the University Paris V - René Descartes took part to the experiment. All were native French speakers and they were between 18 and 30 years old.

## Results

Reaction times higher than 1500 ms were eliminated from the statistical analyses; less than 1% of reaction times were suppressed with this criterion. There were 2% of errors on experimental words. Analyses of variance were conducted on the inverse reaction time data. This allowed outliers to be included without unduly affecting the estimates of condition mean (Ratcliff, 1993; Ulrich & Miller, 1994). Two analyses have been run: one across subject (F1) and the other across item (F2). Reaction times per conditions are presented in Figure 2. This also gives the priming effects and their associated significance values.

First, comparing regular conditions and control conditions, we observed an effect of morphological priming ( $F(1,35)=103.17$ ,  $p<.000$ ;  $F(1,92)=119.45$ ,  $p<.000$ ) and an effect of type of verbs per subjects ( $F(3,105)=9.19$ ,

$p<.000$ ;  $F(1,92)=2.05$ , n.s.) but no interaction between these two factors ( $F(1,35)<1$ ;  $F(1,92)<1$ ). Comparing irregular conditions with control conditions, we observed an effect of morphological priming ( $F(1,35)=80.43$ ,  $p<.000$ ;  $F(1,92)=142.33$ ,  $p<.000$ ) and an effect of type of verb ( $F(3,105)=7.04$ ,  $p<.000$ ;  $F(1,92)=2.26$ ,  $p=.09$ ) but no interaction between the two ( $F(3,105)=1.31$ , n.s.;  $F(3,92)=1.83$ , n.s.). Comparing regular and irregular conditions, we observed no effect of type of priming ( $F(1,35)=2.98$ , n.s.;  $F(1,92)=1.01$ , n.s.), an effect of verb types per subject ( $F(3,105)=4.31$ ,  $p<.007$ ;  $F(1,92)=1.13$ , n.s.) but again no interaction ( $F(1,35)<1$ ;  $F(1,92)<1$ ).

These results show that irregular and regular verb forms prime their infinitive form equally, and that these priming effects do not vary with the type of verb (irregular vs. regular).

Table 2: Results of Experiment 1

Type of verbs	Primes	Targets	RT(ms)	Priming effect
Regular	<i>aimerons</i>	<i>aimer</i>	523	44**
	<i>aimons</i>		530	37**
	<i>porterons</i>		567	
Morpho-phonologic constraints	<i>semons</i>	<i>semer</i>	539	57**
	<i>sème</i>		545	51**
	<i>votons</i>		596	
Sub regularity	<i>teindra</i>	<i>teindre</i>	553	60**
	<i>teignent</i>		551	62**
	<i>nichera</i>		613	
Idiosyncratic	<i>allons</i>	<i>aller</i>	544	49**
	<i>irons</i>		545	48**
	<i>tenons</i>		593	

Note: \*\*  $p < .05$

## Discussion

This cross-modal priming experiment presented a pattern of results which was very clear cut: a massive morphological priming effect and no interaction between this effect and the type of primes (regular vs. irregular) or the type of verbs. These results show that in French there is no difference in the amount of priming produced by a regular verb form and the one produced by an irregular verb form on the identification of their infinitive.

A major concern in cross modal experiments is to determine if the priming effects observed for morphologically related pairs are due to shared morphemes in a morphologically structured mental lexicon, or if they are due to the semantic relationships between the morphologically related pairs. Given the across-the-board priming effects in Experiment 1, and given that all the primes and targets were highly semantically as well as morphologically related, we decided to run the same materials in a second experiment using a masked priming technique developed by Forster and Davis (1984). The masked priming technique has been shown to be highly sensitive to overlap at the level of form (Forster, Davis, Schoknecht, & Carter, 1987; Forster and Taft, 1994), but not of meaning. Although masked priming

effects for associatively related pairs have been observed (Sereno, 1991), no pure semantic effect had been reported. In masked priming a forward pattern mask is presented immediately before the prime and the prime is then covered by the target item: this latter is used as a backward mask. The temporal interval between the onset of the priming stimulus and the subsequent target stimulus is very brief (47 ms in our experiment). At these short prime durations, the combination of forward and the backward masking prevents the subject from consciously seeing the prime. The consequence of this is that the participant's responses are not influenced by a conscious appreciation of the relationship between the prime and the target. This reduces the possibility that any priming effect is due to the fact that the participant realises that the prime and the target often share a common morpheme.

## Experiment 2

### Method

**Material and Design** Our second experiment used a masked-priming paradigm with the same stimuli as the previous experiment. We added two additional controls: a semantic condition, where the prime and the target were semantically related, to check that the masked priming paradigm was not picking up semantic effects; and an orthographic condition where the prime and the target orthographically overlapped to the same degree as the related pairs but had no semantic or morphologic relationship. We selected 24 target-words. For each target word in this condition (such as *mâcher*), one prime was semantically related to the target (*broyait*), one prime was phonologically related to the target (*machine*) and the third type of prime was an unrelated control (*progrès*). As a consequence of these changes we removed 24 word/word filler pairs to keep the balance between word and nonword answers. As a result this gave us the same number of items in each list as for the previous experiment.

**Procedure** The same hardware and software were used as in the previous experiment. Each trial consisted of three visual events. The first was a forward pattern mask consisting of a sequence of '#'. The second event was the display of the prime word for 47 ms. The third event was the presentation of a target word or nonword for 500ms. The prime was in lower case and the target in upper case to make sure that the former was appropriately masked. Subjects were asked to make a quick and accurate lexical decision about the target by pressing a 'word' or 'nonword' key. The experiment lasted about 30 minutes and started with 10 practice trials followed by 10 warm-up pairs and then the experimental trials. There were breaks as in the previous experiment. No subjects reported any awareness of the presence of a prime.

**Participants** Another 42 native French speakers of the same age and from the same population as before took part in the experiment.

### Results

Reaction times higher than 1500 ms were eliminated from the statistical analyses; less than 1% of reaction times were suppressed with this criterion. There were 2% of errors on experimental words. Analyses of variance were conducted on the inverse reaction time data both across subject (F1) and item (F2). Reaction times per condition are presented in Table 3. This also gives the priming effects and their associated significance values.

Table 3: Results of Experiment 2

Conditions	Primes	Targets	RT ms	Priming effect
Regular	<i>aimerons</i>	<i>aimer</i>	551	19**
	<i>aimons</i>		552	18**
	<i>porterons</i>		570	
Morpho-phonologic constraints	<i>semons</i>	<i>semer</i>	569	19**
	<i>sème</i>		566	22**
	<i>votons</i>		588	
Sub regularity	<i>teindra</i>	<i>teindre</i>	564	32**
	<i>teignent</i>		578	18**
	<i>nichera</i>		596	
Idiosyncratic	<i>allons</i>	<i>aller</i>	560	32**
	<i>irons</i>		578	14**
	<i>tenons</i>		592	
Semantic and Orthographic Controls	<i>broyait</i> <i>machine</i> <i>progres</i>	<i>macher</i>	587 599 592	5 -7

Note: \*\*  $p < .05$

Comparing first regular conditions and control conditions, we observed an effect of morphological priming ( $F(1,41)=36.74$ ,  $p<.000$ ;  $F(1,92)=59.06$ ,  $p<.000$ ) and an effect of type of verb per subject ( $F(3, 123)=4.73$ ,  $p<.004$ ;  $F(2<1)$ ) but no interaction between the two factors ( $F(1<1$ ;  $F(2<1)$ ). Comparing irregular conditions with control conditions, we observed an effect of morphological priming ( $F(1,41)=22.03$ ,  $p<.000$ ;  $F(1,92)=30.96$ ,  $p<.000$ ) and an effect of type of verb per subjects ( $F(1(3, 123)=7.84$ ,  $p<.000$ ;  $F(2(1, 92)=1.152$ , n.s.) but no interaction ( $F(1<1$ ;  $F(2<1)$ ). Comparing regular and irregular conditions, we observed no effect of the type of priming ( $F(1(41)=2.16$ , n.s.;  $F(2(1,92)=2.09$ , n.s.), an effect of type of verb per subject ( $F(1(3, 123)=4.81$ ,  $p<.003$ ;  $F(2<1)$ ) but no interaction ( $F(1(3,123)=1.31$ , n.s;  $F(2(3,92)=1.18$ , n.s.). In the control condition we found no effect of semantic priming ( $F(1(41)=1.8$ , n.s.;  $F(2(1,21)=1.05$ , n.s.) and no effect of orthographic overlap ( $F(1(41)<1$ ;  $F(2<1)$ ), allowing us to rule out accounts of the results in terms of simple form overlap between prime and target.

These results confirmed the results observed in the cross-modal experiment and show that irregular and regular verb forms prime their infinitive form equally, and that these

priming effects do not vary with the type of verb (irregular vs. regular). The fact that these effects are found in a task which is generally insensitive to semantic relations between prime and target - and where the semantic control condition showed no priming - is good evidence that these are genuinely morphological effects, reflecting repeated access to the same underlying morpheme. This morpheme seems to be accessed equally effectively, regardless of the degree or type of irregularity on the prime word.

## General Discussion

The question asked here was whether French regular and irregular inflected forms show different priming patterns, in the same way as English. The dual mechanism hypothesis postulates a rule-based symbolic processor that supports the representation and generation of regular forms, while an associative rote-memory system is required to account for irregular forms. Pinker (1991) claimed that the different priming effect observed in English for regular and irregular forms support the dual mechanism dichotomy. Using French we found no such difference. The priming generated by regular inflected words did not differ from the priming generated by irregular forms. The facilitatory effects of morphologically related primes are just as strong whether they involve the same or different underlying roots as their targets. Pairs like *aimons/aimer* prime just as well as pairs like *buvons/boire*. These findings seem inconsistent with the predictions of the dual mechanism hypothesis for the processing behaviour of listed forms in a repetition priming task. In the framework of the dual mechanism account, because Conjugation 3 verbal forms are completely idiosyncratic and unpredictable, they will use rote-learning of irregular stems and they will be stored as independent but linked forms in a pattern-associative memory. For a priming task, this predicts reduced priming between prime/target pairs involving different underlying roots, a prediction confirmed in earlier research in English. The results obtained in French contrast with those obtained in English.

The pattern of results observed in French could be explained in terms of connectionist distributed networks, operating sub-symbolically and without syntax (Rumelhart & McClelland, 1986; MacWhinney & Leinbach, 1991; Plunkett & Marchman, 1993). Indeed, the absence of interaction between priming effects observed with regular forms and priming effects observed with irregular forms seems to go against the dual mechanism hypothesis. We could argue that English speakers use rules because the contrast between regular and irregular verbs is sharp, which is not the case in French. In French irregular verbs may often have regular forms in many cases and irregular forms only for particular tenses and persons. As an example table 4 presents the different forms of the irregular verb *aller* for three different tenses and all persons. So for the verb *aller*, while imperfect forms are fully regular (such as *allais*), future forms are all irregular (such as *irai*) and the present forms are both regular and irregular, depending of the person (such as *allons* and *vont*). This complexity and the lack of clear-cut distinction between regular and irregular verbs could discourage

the system from relying on rules. However, the idea that the cognitive system would not be able to use regularities because of the complexity of the verbal system lacks plausibility given the complexity of other processes involved in language understanding.

Also, the result profile observed in French might not, by itself, be an insuperable problem for the dual mechanism account. One possibility is that the processing architecture of French differs from that of English in ways which allow listed allomorphs to prime each other. The idea would be that even if regular forms are retrieved by rule decomposition and not irregular ones, the behavioral output observed (in this case, the priming effect) would be the same even if the underlying processes are different.

Table 4: Verbal forms of the verb *aller* for the three indicative tenses and the three singular and plural persons.

<b>aller</b> <b>(infinitive form)</b>	<b>Present</b>	<b>Imperfect</b>	<b>Future</b>
je (1 sing.)	<i>vais</i>	<i>allais</i>	<i>irai</i>
tu (2 sing.)	<i>vas</i>	<i>allais</i>	<i>iras</i>
il/elle/on (3 sing.)	<i>va</i>	<i>allait</i>	<i>ira</i>
nous (1 plur.)	<i>allons</i>	<i>allions</i>	<i>irons</i>
vous (2 plur.)	<i>allez</i>	<i>alliez</i>	<i>irez</i>
ils/elles (3 plur.)	<i>vont</i>	<i>allaient</i>	<i>iront</i>

Note: sing.: singular; plur.: plural; 1: first person (I or we); 2: second person (you); 3: third person (he/she or they).

Perhaps a more important difference between the idiosyncratic verbal forms in French and the irregular forms in English is that French forms are decomposable while English forms are not. English irregular forms such as *drove* or *gave* are not only irregular but also cannot be further morphologically decomposed. They must be learnt and represented as unanalysable whole forms. In French, even idiosyncratic irregular forms like *buvait* (from *boire*) undergo a regular suffixation procedure: '-ait' is the regular affix for the imperfect third person form. Irregular forms in French are composed of a changed stem plus a regular affix. The irregularity is in the choice of the stem used but the regular suffix procedure applies anyway.

Marslen-Wilson et al. (1995) explain the English results in terms of the inhibitory consequences during acquisition of having to learn, for each irregular stem, to block the application of the default regular suffix. If indeed in English the two possible stems compete with each other during identification in order to block the decomposition process (in case the form that has to be identified is the irregular one), such a process would not be necessary in French. For French irregular verbs, two types of stem would be possible but even if the form presented is irregular there would not be the same type of competition because in both cases decomposition would be necessary to reach identification of the verbal form. Both regular and irregular forms would follow the same processing pathways - which is arguably not the case for regular and irregular forms in English (Marslen-Wilson & Tyler, 1998). If regular and irregular forms can co-exist in this fashion, then both can be linked to the

underlying verbal morpheme without competition from the other - and without the requirement to postulate distinct types of computational procedure to support the generation and analysis of each type of form.

### Acknowledgments

This research was supported by the Medical Research Council (U.K). The research of Fanny Meunier has been made possible by fellowships from the Fyssen Foundation and the French Foreign Office (Bourse Lavoisier). We thank Juan Segui for his hospitality during testing and Matthew Brett for his comments.

### References

- Forster, K.I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 680-698.
- Forster, K.I., Davis, C., Schoknecht, C., & Carter, R. (1987). Masked priming with graphemically related forms: Repetition or partial activation? *Quarterly Journal of Experimental Psychology*, 39, 211-251.
- Forster, K.I., and Taft, M. (1994). Bodies, antibodies, and neighborhood density effects in masked form-priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 844-863.
- Kempey, S.T., & Morton, J. (1982). The effects of priming with regularity and irregularity related words in auditory word recognition. *British Journal of Psychology*, 73, 441-454.
- Marslen-Wilson, W.D., Hare, M. & Older, L. (1995). Priming and blocking in the mental lexicon: The English past tense. *Paper presented at the Meeting of the Experimental Psychology Society, London, January.*
- Marslen-Wilson, W.D., & Tyler, L. K. (1998) Rules, representations, and the English past tense. *Trends in Cognitive Sciences*, 11, 428-435.
- MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, 40, 121-157.
- Napps, S. (1989). Morphemic relationships in the lexicon : Are they distincts from semantic and formal relationships ? *Memory and Cognition*, 17, 729-739.
- Orsolini, M. & Marslen-Wilson, W.D., (1997). Universals in Morphological Representation: Evidence from Italian. *Language and Cognitive Processes*, 12(1), 1-47.
- Pinker, S. (1991). Rules of language. *Science*, 253, 530-535.
- Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, 48, 21-69.
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114(3), 510-532.
- Rumelhart, D.E., & McClelland, J.L. (1986). On learning the past tenses of English verbs. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2. Psychological and biological models* (pp. 216-271). Cambridge, MA: MIT Press.
- Sereno, J. A. (1991). Graphemic, associative, and syntactic priming effects at a brief stimulus onset asynchrony in lexical decision and naming, *Journal of Experimental Psychology : Learning, Memory and Cognition*, 17(3), 459-477.
- Stanners, R.F., Neiser, J.J., Hernon, W.P. and Hall, R. (1979). Memory representation for morphologically related words. *Journal of Verbal Learning and Verbal Behavior*, 18, 339-413.
- Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, 123(1), 34-80.

# Neighborhood and Position Effects Interact in Naming Latency

**Jeanne C. Milostan** (JMILOSTA@CS.UCSD.EDU) **Victor Ferreira** (FERREIRA@PSY.UCSD.EDU)  
**Garrison W. Cottrell** (GARY@CS.UCSD.EDU)  
Computer Science and Engineering Department 0114  
University of California San Diego  
La Jolla, CA 92093 USA  
Department of Psychology  
University of California San Diego  
La Jolla, CA 92093 USA

## Abstract

Naming latency studies have recently shown a position-of-irregularity effect (words with early irregularities seem slowed compared to those with late irregularities), for which Dual-Route models of reading can account. Milostan & Cottrell (1998) showed that the initial studies contained a confound between irregularity position and friend/enemy ratio, and that the statistical confound could be captured by connectionist networks which then show the supposed position effect. This paper presents work to disentangle the position/regularity confound through a subject study and additional connectionist explorations. The latency data show that, once friend/enemy ratios are controlled for, the supposed position effect is driven entirely by high-enemy words in the first position. Further, connectionist network simulations show that network error at the first phoneme position only is a better match for naming latency, while overall network error produces a better match to subject error counts.

## Introduction

A major component of the task of learning to read is the development of a mapping from orthography to phonology. In a complete model of reading, message understanding must play a role, but many psycholinguistic phenomena can be explained in the context of this simple mapping task. A difficulty in learning this mapping is that in a language such as English, the mapping is *quasiregular* (Plaut et al., 1996); there are a wide range of exceptions to the general rules. As with nearly all psychological phenomena, more frequent stimuli are processed faster, leading to shorter naming latencies. The regularity of mapping interacts with this variable, a robust finding that is well-explained by connectionist accounts (Seidenberg and McClelland, 1989; Taraban and McClelland, 1987).

In this paper we continue consideration of a recent effect that seems difficult to account for in terms of the standard parallel network models. Coltheart & Rastle (1994) have shown that the amount of delay experienced in naming an exception word is related to the phonemic position of the irregularity in pronunciation. Specifically, the earlier the exception occurs in the word, the longer the latency to the onset of pronouncing the word. Table 1, adapted from (Coltheart and Rastle, 1994) shows the response latencies to two-syllable words by normal subjects. There is a clear left-to-right ranking of the latencies compared to controls in the last row of the Table. Coltheart *et al.* claim this delay ranking cannot be achieved by standard connectionist models. Earlier work (Milostan and Cottrell, 1998) showed that the origin of the effect seen in the Coltheart study lies in a statistical regularity of English, related to

Filler	Position of Irregular				
	1	2	3	4	5
Nonword					
Irregular	554	542	530	529	537
Regular	502	516	518	523	525
Difference	52	26	12	6	12
Exception					
Irregular	545	524	528	526	528
Regular	500	503	503	515	524
Difference	45	21	25	11	4
Avg. Diff.	48.5	23.5	18.5	8.5	8

Table 1: Naming Latency vs. Irregularity Position

the number of “friends” and “enemies” of the pronunciation within the word’s neighborhood.<sup>1</sup> The human subject study and network simulations presented in this paper attempt to tease apart the effects of phoneme position and neighborhood ratio.

## Background

Computational modeling of the reading task has been approached from a number of different perspectives. Advocates of a dual-route model of oral reading claim that two separate routes, one lexical (a lexicon, often hypothesized to be an associative network) and one rule-based, are *required* to account for certain phenomena in reaction times and nonword pronunciation seen in human subjects (Coltheart et al., 1993). Connectionist modelers claim that the same phenomena can be captured in a single-route model which learns simply by exposure to a representative dataset (Seidenberg and McClelland, 1989).

In the Dual-Route Cascade model (DRC) (Coltheart et al., 1993), the lexical route is implemented as an Interactive Activation (McClelland and Rumelhart, 1981) system, while the non-lexical route is implemented by a set of grapheme-phoneme correspondence (GPC) rules learned from a dataset. Input at the letter identification layer is activated in a left-to-right sequential fashion to simulate the reading direction of English, and fed simultaneously to the two pathways in the

<sup>1</sup>Friends are words with the same pronunciations for the ambiguous letter-to-sound correspondence; enemies are words with different pronunciations.

model. Activation from both the GPC route and the lexicon route then begins to interact at the output (phoneme) level, starting with the phonemes at the beginning of the word. If the GPC and the lexicon agree on pronunciation, the correct phonemes will be activated quickly. For words with irregular pronunciation, the lexicon and GPC routes will activate different phonemes: the GPC route will try to activate the regular pronunciation while the lexical route will activate the irregular (correct) pronunciation. Inhibitory links between alternate phoneme pronunciations will slow down the rise in activation, causing words with inconsistencies to be pronounced more slowly than regular words. This slowing will not occur, however, when an irregularity appears late in a word since the lexicon will try to activate *all* of a word's phonemes as soon as the word's lexical node becomes active. If an irregularity is late in a word, the correct pronunciation will begin to be activated before the GPC route is able to vote against it. Hence late irregularities will not be as affected by the conflicting information. This result is validated by simulations with the one-syllable DRC model (Coltheart and Rastle, 1994).

Several connectionist systems have been developed to model the orthography to phonology process (Seidenberg and McClelland, 1989; Plaut et al., 1996). These connectionist models provide evidence that the task, with accompanying phenomena, can be learned through a single mechanism. In particular, Plaut *et al.* (henceforth PMSP) develop a recurrent network which duplicates the naming latencies appropriate to their data set, consisting of approximately 3000 one-syllable English words (monosyllabic words with frequency greater than zero in the Kuçera & Francis corpus (Kuçera and Francis, 1967)). Naming latencies are computed based on time-to-settle for the recurrent network, and based on mean squared error (MSE) for a feed-forward model used in some simulations. The structure of the feed-forward network is shown in Figure 1. In addition to duplicating frequency and regularity interactions displayed in previous subject work, this model also performs appropriately in providing pronunciation of pronounceable nonwords. This provides an improvement over, and a validation of, previous work with a strictly feed-forward network (Seidenberg and McClelland, 1989). (Milostan and Cottrell, 1998) then showed that the serial position effect proposed by Coltheart & Rastle could be accounted for by a statistical regularity in English, as measured by the Enemy Ratio (# of enemies in a word's neighborhood divided by the total size of the word's neighborhood). (Milostan and Cottrell, 1998) showed that, for the words used in (Coltheart and Rastle, 1994), words with earlier irregularities had higher enemy ratios, and that the parallel connectionist model of PMSP, exposed to the same statistical regularities, also shows the same left-to-right effect that (Coltheart and Rastle, 1994) claimed it would not.

## Experiment

Intuition suggests that, since English is read from left to right, left-to-right phenomena such as the serial position effect might be seen, independent of statistical confounds. However, as with all assumptions, such effects must be verified through careful testing, and the source of such effects must be carefully delineated within the model hypothesized for the system at hand.

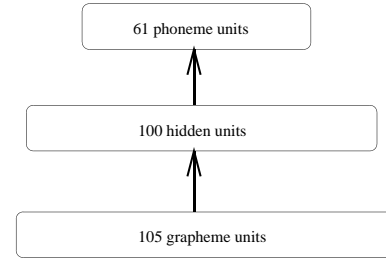


Figure 1: Single Syllable Ortho-to-Phono Network

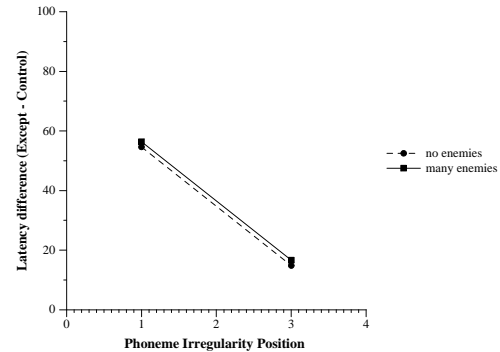


Figure 2: Hypothetical Position-Only Effect

In a serial system such as the DRC, which by design processes input orthography from left to right, any observed left-to-right irregularity effect is a direct result of the GPC operation. On the other hand, for a parallel model such as the PMSP system, which produces the output phonology all at once, effects of irregularity are driven by neighborhood enemy/friend measures, and serial effects should disappear once these enemy ratios are controlled.

The serial position effect seen by Coltheart & Rastle could be the result of a confound between the position of the irregularity and the statistics of English. Earlier positions appear to have more irregularities. It would be productive, then, to retest the Coltheart & Rastle hypothesis, this time controlling for amount of consistency. If the serial position effect does hold regardless of the enemy ratio of the test words, an effect similar to that shown in Figure 2 would be expected. If, however, the effect is due to enemy ratio alone, the results should be similar to that of Figure 3. The subject experiment and network simulation presented here are an attempt to adjudicate between these options, and stimuli will vary in both position of irregularity, and in enemy ratio, in order to determine the source of the effects.

## Difficulties of GPC rules

One of the major discrepancies between the PMSP work and DRC model is the latter's assumption of the existence of a pronunciation rule system. This rule system *defines* whether a word is regular or not. Thus, all irregular stimuli chosen for experiments on the DRC model are chosen according to the GPC rules. Experiments which attempt to refute the DRC model at any level must also take these rules into consideration when choosing stimuli.

Ideally, the same words that the DRC system uses should



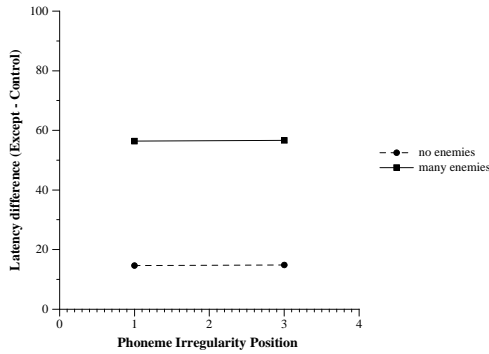


Figure 3: Hypothetical Enemy-Ratio-Only Effect

be addressed. Thus, irregular words for this experiment were identified using the Australian GPC used in the DRC system, and the Australian pronunciations from the MRC database, from which the GPC rules were initially derived. Neighborhood Enemy Ratios were also calculated using the MRC pronunciations. American English would then be used only for identifying errors in subjects' pronunciations.

A program was written implementing the GPC rules of the DRC system as listed in (Rastle and Coltheart, 1999). A word was considered irregular if the pronunciation generated by the rules did not match the pronunciation provided by the MRC database. From the list of identified exception words, homographs where one generated pronunciation was correct were excluded (/wind/ vs. /wInd/), as were Australian words with spellings not commonly used in the United States (gaol). Also excluded were words marked as irregular by the GPC rule which states that word-final /s/ is *always* converted to /z/. This rule causes all words ending in -ace or -ice (face, mice) to be considered irregular.

Overall, the words which were identified by this procedure using the Australian English were also found to be irregular through a similar procedure using the rules of American English pronunciation from (Venezky, 1970). The details of that investigation are reported with a companion study in (Milostan et al., 2000).

### Neighborhoods

Neighborhoods are defined using an extension of the (Taraban and McClelland, 1987) neighborhood rules described in (Milostan and Cottrell, 1998), summarized for single syllable words here:

- Consonant neighborhoods consist of orthographic clusters which correspond to the same location in the word. For one-syllable words, this results in 2 consonant cluster locations: onset and coda.
- Each vowel group is considered within the context of its coda. In order for a word to be in the neighborhood of a test word, it must have the same vowel group ('E' is considered separately from 'EE') and be followed by the same consonant cluster ending that syllable. As an example, the 'OO' neighborhood in 'BOOK' are all those words ending in 'OOK', with the first syllable coda containing only 'K'.
- Consonant cluster neighborhoods include the preceding

Enemy Ratio	Position	
	Front	Back
High	aunt	plaid
Low	earl	fluke

Table 2: Sample Experiment Words

vowel for coda consonants, and the following vowel for onset consonants. As expected, consonant irregularities are by far the minority, and are limited to 'CH', 'TH', 'G', 'C', 'Q', and the silent instantiations such as 'T' and 'H'.

## Methods

### Subjects

Subjects were 23 undergraduate psychology students from University of California San Diego. All subjects had normal or corrected-to-normal vision, and were native North-American-English speakers. They were given course credit for their participation.

### Materials

Sixty-four words with irregular grapheme-to-phoneme correspondences (according to the GPC rules of the DRC model) were chosen. Each target was uninflected and monosyllabic, and had between 3 to 6 letters with Kuçera-Francis frequency between zero and twenty-two.

The chosen words had an irregular grapheme-to-phoneme correspondence in either the first ("front") or third ("back") phoneme position, and were divided into 2 lists on that basis. Each list was further divided into two sublists, based on whether the word had only friends in the neighborhood based on the regularity (Enemy Ratio  $En = 0.00$ ) or mostly enemies at that location (Enemy Ratio  $0.6 \leq En < 1.0$ ). Since a word's neighborhood by our measure includes itself, words with a neighborhood size of one ("loners") were excluded from consideration. These words correspond to Colheart's categorization of "irregular consistent".

Of the eligible words, the front-enemy condition had only 16 candidate words. Each of the other three conditions were randomly pruned down to size 16 in order to balance the conditions. The resulting average word frequency did not differ significantly between conditions ( $M = 4.8281, F(3, 60) = 0.476, p = 0.700$ ). Each irregular word was then matched with a regular control word. Control words were matched to their irregular partners based on initial phoneme (since different phonemes take longer to trigger the microphone) and number of letters. The controls were also in the zero to 22 Kuçera-Francis frequency range.

An example test word from each of the four conditions is shown in Table 2.

## Results

Of the original 25 subjects, data from 2 were unusable (in one case the latency data were accidentally deleted; in the other case the audio recording did not function so errors could not be scored). For the remaining 23 data sets, latencies associated with voice key failures were discarded; if the stimulus was either a test word or a control the associated (control or test) word was similarly disregarded (13 pairs total

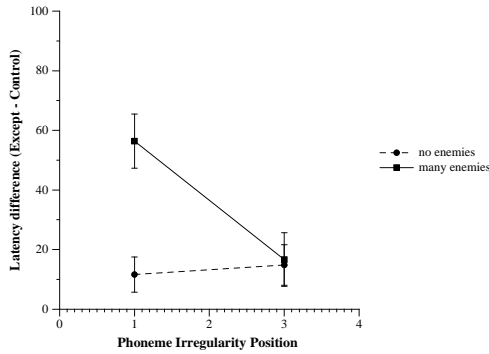


Figure 4: Naming Latency Results overall

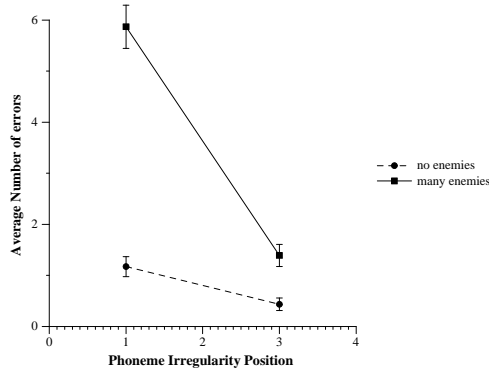


Figure 5: Error Count Results overall

over all subjects). Latencies for all nonword fillers were also discarded. Words which were pronounced incorrectly, along with the associated match, were removed for separate error analysis.

Naming latency differences were then calculated by subtracting the control word latency from the associated test word latency. Analysis of variance (ANOVA) was then performed on these values. Words in the high enemy ratio condition had significantly greater latency differences than the words in the friend condition ( $F(1, 22) = 18.16, p < 0.0005$ ), and there was a significant interaction between enemy ratio and position of irregularity ( $F(1, 22) = 8.419, p = 0.008$ ). Latency differences for first and third position irregularities, combining both enemy ratio conditions, approached but did not reach significance ( $F(1, 22) = 3.766, p = 0.065$ ). The latency data is shown in Figure 4.

Subjects made a total of 22 errors on control words, and 248 errors on irregular test words. Control words are not considered in the error analysis. Subjects made significantly more errors for front position irregulars than for back position irregulars ( $F(1, 22) = 32.922, p < 0.0005$ ), and more errors for high-enemy words than for low-enemy words ( $F(1, 22) = 326.549, p < 0.0005$ ). Position and enemy ratio also had a significant interaction in number of errors made ( $F(1, 22) = 12.415, p = 0.002$ ). These error data are shown in Figure 5.

## Discussion

From the data collected in this experiment, there is a slight effect of irregularity position, but this appears to be completely

driven by the words with high enemy ratios (see Figure 4). First-position-irregular words with high numbers of enemies in their neighborhood take longer to name than similar words with friends only. This effect has mostly disappeared for those words with third position grapheme-phoneme irregularities.

This makes sense from a cascaded information processing point of view (McClelland, 1979), since it is possible that any (potential) errors late in a word can be resolved by the time the third phoneme is ready to be produced. This difference in time delays can be considered an effect of the temporal nature of the speech process, and the time available to make online corrections. Words with later irregularities have, by definition, *regular* grapheme-phoneme correspondences at the beginning. The subject can begin pronouncing those phonemes immediately, even if she must then make accommodations later. Thus, the initial phoneme in an irregular (high enemy ratio) word may be produced with the same latency as a completely regular word, while the phoneme at the irregular mapping itself may actually be delayed internal to the word. However, there is currently no way of measuring the latencies of each internal phoneme using only the voice key.

## Feed-Forward Network Performance

The feed-forward network of PMSP does not contain a temporal component. Since all phonemes are calculated simultaneously, the irregularity position may not play a part in the latencies calculated from the network as these are actually a measure of the difference between the correct target pronunciation and the network's actual output across the word. Thus, the feed-forward pronunciation network should be affected by enemy ratio alone, as those words with many contradictory spelling-sound mappings will receive less total reinforcement for the correct mapping.

Five feed-forward connectionist networks were trained on 3015 single syllable words as described in (Plaut et al., 1996; Milostan and Cottrell, 1998). This data set is the 2998 words used in PMSP plus 17 additional words used in the current subject experiment. These words were not included in the PMSP data set as they are of frequency rating zero.

Naming latency was then calculated for each test word by using the sum squared error at the output layer, producing the results shown in Figure 6. Unexpectedly, it appears that the back position irregulars take *longer* to name than the front irregulars, regardless of enemy ratio. Remember, however, that naming latency in these feed-forward networks is a measure of error, not of time directly. The representation used for the output layer is a sparse coding of the output phonemes. Of the 62 units, only a small number will be on for any particular word. Thus, the network is exposed to a training set where the majority of the output units are off most of the time. These networks learn how to turn off units very well, and thus there will be less discrepancy between the target and actual output which should be off and is, than for a target which should be on and the related output unit which is actually on. This means that, everything else being equal, training pairs with more units on in the target will inherently produce more error than for those with fewer on targets.

As an example, consider a hypothetical network with 10 output units, and compare the results of two targets, one of

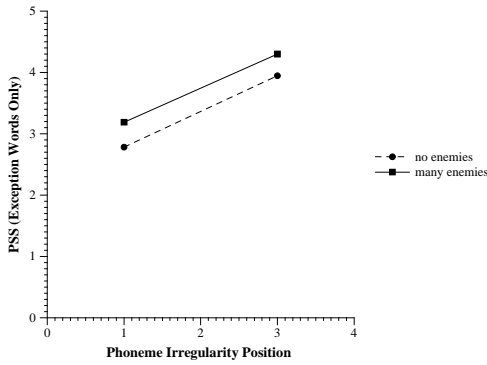


Figure 6: Network Sum Squared Error

which has one unit on and the other of which has two units on. Since the “off” units receive more training, assume that any units off in the target have an activation of 0.1 in the actual network during testing, while the on-units are activated at 0.8. Both of the hypothetical training sets accurately produce the intended output, but in the case where the target has one unit on the network shows an error of  $(0.1^2 * 9) + (0.2^2 * 1) = .13$ , while the target with 2 units on has an error of  $(0.1^2 * 8) + (0.2^2 * 2) = .16$ . This discrepancy becomes more exaggerated as the number of off-units increases. Thus, if there is a systematic difference in the expected number of on-units among the conditions, those targets with more on-units may be unduly penalized. Examination of the output targets for the various test categories reveals that indeed, those in the back position conditions have more on-units than the front position targets, as shown in Figure 7. This means that the words in the back position systematically have one more phoneme than the front-position test words.

In parallel connectionist models, output error is associated with naming latency under the assumption that the more error the output shows, the longer it takes for the system to then converge on a veridical representation for a further stage which will begin the actual production of the speech signal. If the output for each of the ON-bits in the representation can be cleaned up in parallel, then the time required before the next stage may begin is more a measure of the average amount of time required to make the cleanup. Thus, the average ON-bit error provides a more realistic measure of naming latency.

To correct for the bias in number of bits on between first- and third-position words, the total output error for each word was divided by the number of ON-units in that word’s output representation. These results are shown in Figure 8. As in the human data, the networks show a significant effect of enemy ratio ( $F(1,4) = 478.669, p < 0.0005$ ) and a significant interaction between enemy ratio and position ( $F(1,4) = 621.335, p < 0.0005$ ). Unlike the human subjects, however, the networks also show a significant effect of position ( $F(1,4) = 683.588, p < 0.0005$ ). Again, this appears to be mostly driven by the high enemy ratio words. This is actually a bit surprising since the networks produce output in parallel, and thus would not have any time to “correct” for later-position errors. The network results instead reflect the finding that English words are more consistent in endings than they are in onsets (Treiman et al., 1995).

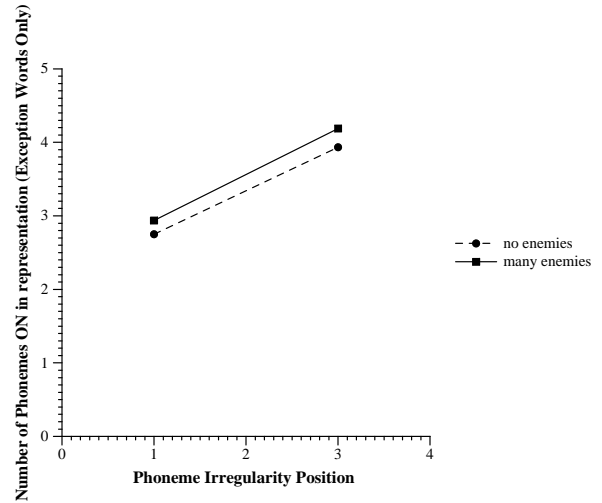


Figure 7: Number of Units ON in Output Representation

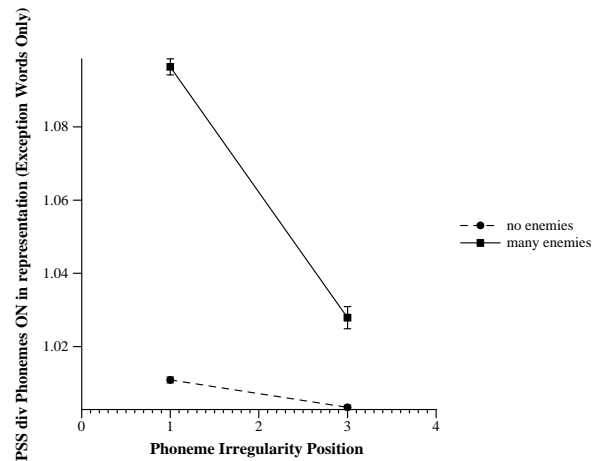


Figure 8: Network Error / Number of Phonemes ON

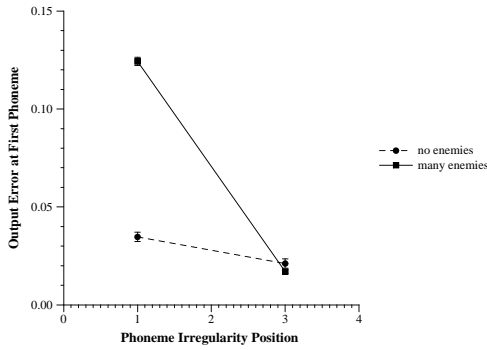


Figure 9: Study #1 Network Error at First Phoneme Only

The corrections made through dividing network output error by the number of ON-bits is reasonable enough, assuming the networks represent a parallel stage which leads to a temporal phoneme-output system. If the motor system were reading off the outputs of this network sequentially, then naming latency would be measured by the error at the first phoneme produced, just as naming latency in subjects is measured as the time to begin the first phoneme. For the networks developed here, then, an even better measure of the naming latency would be the network output error at the first phoneme only. This data is shown in Figure 9. Here there is a significant main effect of both position ( $F(1,4) = 1071.117, p < 0.0005$ ) and Enemy-Ratio ( $F(1,4) = 242.419, p < 0.0005$ ), and a significant interaction between position and ratio ( $F(1,4) = 612.797, p < 0.0005$ ). As can be seen, the data here produce a better match to the human subject latency data, while the total output error (Figure 8) is a better match to the human subject error data (Figure 5). The total network error (driven by the Enemy-Ratio of the words in question) is reflective of the probability that the subject will make an error on the word, while the network first-phoneme-error represents the amount of time it takes for the subjects to begin producing that first phoneme.

### Summary

Subject performance in the naming latency task is driven mostly by the high-enemy-ratio words. Words with irregularities in the first position in the “many enemies” condition are greatly slowed, while there is not much difference in naming latencies for words in any of the other conditions.

When using connectionist networks to model word naming latencies, it is traditional to equate overall output error with latency. The network simulations of this subject study show that, rather, the output error of the first phoneme only is a better model of the subject naming latencies. This implies that a recurrent network which produced the phonemes one at a time, perhaps using a feature-based representation, would result in a better model, defining the output error for the first phoneme as the naming latency.

This study seems to show that the traditional measure of network latency is instead a better predictor of subject error. If the system produces a large overall error, chances of settling into an incorrect attractor basin are increased. Continued experimentation into this idea is currently being undertaken.

The subject study performed here indicates that the enemy

ratio at the first position, not the phoneme position *per se*, is the driving force behind word naming latency.

### References

- Coltheart, M., Curitis, B., Atkins, P., and Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, 100(4):589–608.
- Coltheart, M. and Rastle, K. (1994). Serial processing in reading aloud: Evidence for dual route models of reading. *Journal of Experimental Psychology: Human Perception and Performance*, 20(6):1197–1211.
- Kučera, H. and Francis, W. (1967). *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.
- McClelland, J. (1979). On the time-relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, 86:287–330.
- McClelland, J. and Rumelhart, D. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, 88:375–407.
- Milostan, J. C. and Cottrell, G. W. (1998). Serial order in reading aloud: Connectionist models and neighborhood structure. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems 10*, pages 59–65, Denver, Colorado. MIT Press : Cambridge, MA.
- Milostan, J. C., Cottrell, G. W., and Ferreira, V. S. (2000). The role of statistical regularities in reading and connectionist modeling. In Preparation.
- Plaut, D., McClelland, J., Seidenberg, M., and Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103(1):56–115.
- Rastle, K. and Coltheart, M. (1999). Serial and strategic effects in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, 25(2):482–503.
- Seidenberg, M. and McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96:523–568.
- Taraban, R. and McClelland, J. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and Language*, 26:608–631.
- Treiman, R., Mullennix, J., Bijeljac-Babic, R., and Richmond-Welty, E. D. (1995). The special role of rimes in the description, use and acquisition of English orthography. *Journal of Experimental Psychology: General*, 124(2):107–136.
- Venezky, R. L. (1970). *The Structure of English Orthography*. Mouton, The Hague.

# Individual Differences in Exemplar-Based Interference During Instructed Category Learning

David C. Noelle

(NOELLE@CNBC.CMU.EDU)

Center for the Neural Basis of Cognition  
Carnegie Mellon University  
Pittsburgh, PA 15213 USA

Garrison W. Cottrell

(GARY@CS.UCSD.EDU)

Department of Computer Science & Engineering  
University of California, San Diego  
La Jolla, CA 92093 USA

## Abstract

Instructed category learning studies have shown that categorization practice on a fixed set of labeled training exemplars can cause learners to violate explicitly provided categorization instructions. We have previously proposed a connectionist account of this exemplar-based interference effect — an account which predicts that individuals who display initial difficulty in the application of a categorization rule will exhibit greater exemplar-based interference than good rule-followers. In this paper, we report on a study of human instructed category learning performance intended to test this prediction of the model, and we provide the results of additional connectionist simulations which are fit to the human experimental data.

## Introduction

Instructed category learning studies have revealed that experience with labeled examples can sometimes cause learners to deviate from previously provided categorization instructions, even when the category labels on the training items are perfectly consistent with the given instructions (Allen and Brooks, 1991; Brooks et al., 1991). Such experiments typically begin with the presentation of an explicit rule for categorizing stimuli. These initial instructions are followed by a sequence of trials in which stimuli are presented to the learner, one at a time. The learner is asked to make a categorization judgment for each stimulus, and this judgment is immediately followed by performance feedback, providing the correct category label for the object. After a substantial period of such training, the learner is presented with novel stimuli and is asked to provide category labels for these novel items. Previous studies have discovered that learners may sometimes violate the instructions that they were given when faced with a novel stimulus, assigning the category label of a similar training set exemplar to the novel item in lieu of accurately applying the given explicit rule.

We have previously presented a connectionist model of instructed category learning which explains this exemplar-based interference effect as emerging from the use of an error-correcting learning rule when learning from examples (Noelle and Cottrell, 1996). This model posits the existence of a working memory network which actively maintains a distributed pattern of activity encoding the explicitly provided categorization rule. The model also includes a categorization network — a system which assigns category labels based on stimulus features. The behavior of the categorization network is modulated by activity in the working memory network, allowing explicit instructions to shape categorization performance. Exemplar-based interference appears when connec-

tion weights in the categorization network are modified, by an error-correcting learning rule, as the result of performance feedback on the training exemplars. This mechanism makes a general prediction concerning exemplar-based interference: difficulty in rule-following should result in larger amounts of interference. If the network exhibits substantial residual error when applying a given explicit categorization rule, this error will produce large weight changes during exemplar-based training, and significant interference will arise. This prediction has a number of corollaries. First, the complexity of the categorization rule should impact interference, with more complex rules producing more interference. Second, a corresponding trend should be seen across individual learners, with individuals who are error prone at rule application exhibiting more interference than those who find rule application easy.

In this paper, we investigate these predictions. We report on a human instructed category learning study which investigates individual differences in exemplar-based interference, and we provide the results of detailed connectionist simulations which model the observed human learning performance.

## Individual Differences in Interference

### Method

Undergraduate students were asked to learn to categorize a set of simple geometric line drawings into two categories. Each geometric stimulus involved a circle with a radial line, drawn in green on a black computer screen. The circle stimuli varied along two continuous dimensions: size and orientation. Four different sized circles were used, with radii of approximately 5.0 mm, 7.0 mm, 10.0 mm, and 14.0 mm. The number of distinct orientations was also four, with the radial line of the circle rotated counterclockwise from the right-pointing vector by 30°, 60°, 120°, or 150°. Each of the four angles of rotation could be paired with each of the four sizes, producing a set of 16 different stimulus items. These stimuli may be graphically depicted as points in a two dimensional feature space, as shown on the right side of Figure 1. Of the 16 possible circle stimuli, seven were distinguished as training set items. These items are marked with boxes in Figure 1. Four training stimuli were to be placed in one category, called the “black” category here for convenience, and the remaining three were to be placed in the other, called the “white” category.

A typical experimental trial involved the presentation of a circle in the middle of a blank computer screen. Participants were expected to identify the appropriate category for each stimulus, communicating their judgment by depressing the appropriate key on the computer keyboard. No time pressure

was placed on the learners, and accuracy was stressed in initially provided task instructions. Once a category judgment was made for a given stimulus object, a message appeared above the circle indicating if the given classification was correct or not. The correct category label was also explicitly provided at this time. This feedback remained on the display for 2 seconds, after which time the next trial began.

Participants were randomly assigned to one of two experimental conditions: the simple rule condition or the complex rule condition. In each of these two conditions, the experimental session began with the presentation of an explicit rule for categorizing the circle stimuli. The learners in the simple rule condition were told that circles in the “black” category were those of the smallest size, or of the largest size, or rotated 150°. All other circles were to be placed in the “white” category.<sup>1</sup> Participants in the complex rule condition were given similar categorization instructions, only their rule included an “exception” clause. All circles of the smallest size, the largest size, or rotated 150° were to be placed in the “black” category *unless* they were rotated by 60°. All circles with radial lines at 60° were to be placed in the “white” category, even if they were of the largest or smallest sizes. The instructions were designed to ensure that the rules were clearly understood. The four stimuli sizes and the four angles of rotation were graphically displayed on the same screen with the textually presented categorization rule. The rule was described in plain English, making reference to the graphical examples. Furthermore, participants were not allowed to advance to the next stage of the experiment until they demonstrated an accurate memory of the rule by correctly identifying a reworded version of it in a list of three alternatives. Participants also demonstrated retention of the rule by describing it during an informal debriefing following the experiment.

Once categorization instructions were given, learners were presented with 36 blocks of training trials, each block consisting of one presentation of each of the seven training set items, appearing in a random order. Each trial involved the display of one of the stimuli, a categorization judgment on the part of the learner, and a period of performance feedback which provided the correct category label for the object. At the end of this *training phase*, participants were given a short break, during which time they were told that performance feedback would be suspended for the remainder of the experimental session. They were then presented with 8 blocks of trials incorporating all 16 of the possible stimulus objects. The stimuli were presented in a random order, with the learner providing a category label for each object but receiving no feedback concerning the accuracy of such judgments. Following this *testing phase*, the session was paused once more, and learners were given a new categorization rule to apply. They were told that they would soon be asked to classify circles according to the new rule without the benefit of performance feedback. The new rules involved rotating the structure of the original rules in feature space, keeping the complexity of the rules constant. In the simple rule condition, the new rule placed items in the “black” category if they were rotated to 30°, or to 150°, or if they were of the smallest size. The new complex rule was the same, ex-

<sup>1</sup>The stimuli used here and the simple categorization rule are derived from the experiments of Nosofsky, Clark, & Shin (1989).

cept all circles of the penultimate size (i.e., 10.0 mm radius) were to be placed in the “white” category, regardless of orientation. Following these new classification instructions, 8 more blocks of trials were given, each involving all 16 stimuli, randomly sequenced. The goal of this final collection of trials was to assess the rule-following ability of each learner when no exemplar-based feedback was made available. This final *rule-following phase* was conducted with new categorization instructions to avoid transfer from the earlier training phase.<sup>2</sup> The total number of categorization trials experienced by each participant was 508, and these were typically completed within a period of 45 to 55 minutes.

The performance of participants from a third experimental condition is also reported here. The data for these learners were collected during a previous experiment (Noelle et al., 2000). In this third condition, no explicit categorization instructions were given. Participants were asked to learn to categorize the circle stimuli from feedback on the training items alone. These learners experienced 252 training trials with the seven training stimuli, as the instructed learners did, and they were tested, without feedback, on all 16 objects for 128 trials, as before. Since no explicit instructions were given to these participants, no final rule-following test was conducted.

All of the participants in this experiment were undergraduate students enrolled in psychology or cognitive science courses at the University of California, San Diego during the 1996–1997 academic year. They received course credit in exchange for their participation. Data were collected for 36 uninstructed participants, 28 learners in the simple rule condition, and 27 in the complex rule condition. Some students did not appear to be engaged by the task, exhibiting chance level rule-following performance or chance level performance on the training set items even after the 252 training trials. The data for these participants were discarded, leaving valid data for 34 uninstructed learners, 27 learners in the simple rule condition, and 19 learners in the complex rule condition.

## Results

The mean frequency of classification responses, averaged over the uninstructed learners, are displayed in Figure 1. The mean results for the instructed participants are shown in Figure 2. The chart on the right side of Figure 1 presents the letter labels which will be used to refer to individual stimulus objects in the discussion, below. The other feature space graphs in these two figures show the frequency with which learners identified objects as being in the “black” category during various phases of the experiment.

Exemplar-based interference, if present, should be found in the testing phase categorization frequencies. Such interference involves a change in categorization performance away from that dictated by the rule instructions and towards that suggested by the distribution of training set items alone. The responses of the uninstructed learners may be taken as a characterization of the category structure suggested solely

<sup>2</sup>Previous experiments attempted to assess rule-following ability by testing each participant on the original categorization rule, over all 16 stimuli, without feedback, *prior* to the training phase. It was found, however, that such an initial rule-following test, even without performance feedback, impacted performance during the later training and testing phases in a manner which masked interference.

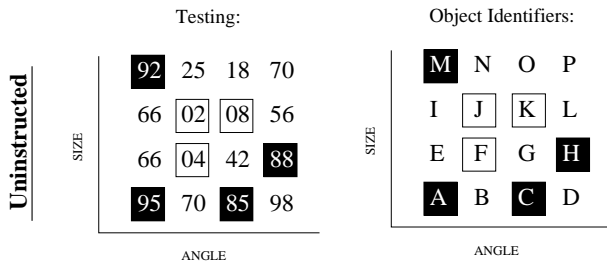


Figure 1: Uninstructed Condition Mean Responses: Categorization results are shown as the percentage frequency with which items were placed in the “black” category. Training set items are marked with boxes, colored according to the assigned category label. Also shown are letter labels for the stimuli, used to reference items in this report.

by the training exemplars. Performance during the final rule-following phase may be used to approximate the accuracy, over all 16 items, with which a learner might have applied the original rule prior to the training phase. Thus, exemplar-based interference may be said to have been present in a given condition to the degree that the pattern of testing phase responding deviated from rule application behavior in the direction of that exhibited by the uninstructed learners.

In order to assess if exemplar-based interference was present in a given experimental condition, careful attention must be given to the amount of error displayed by the participants when they apply an explicit rule without the benefit of exposure to training items. Exemplar-based interference may be said to exist only if deviation from perfect rule application *increases* as a result of performance feedback on the training set. This characterization suggests a quantitative measure of interference involving assessing the deviation from the explicit rule in both the rule-following and testing phases and taking the difference between these two values. Rule-following phase accuracy may be used as an approximation of how well learners might have applied the original explicit categorization rule prior to exemplar-based training. This estimate is compared to categorization accuracy during the testing phase, after training is complete. Using this measure reveals that deviation from the rule did not reliably change in the simple rule condition ( $t(26) = -0.664$ ) but did increase in the complex rule condition, with marginal reliability ( $t(18) = 2.02$ ;  $p = 0.06$ ).

Comparing deviation from the rule across the rule-following and testing phases is not a very powerful test of interference, however. The presence of exemplar-based interference does not entail increased deviation from the rule for *all* stimulus items. Indeed, it is reasonable to expect that the classification performance for some of the stimulus objects will become *more* consistent with the explicit rule as a result of exemplar-based performance feedback, since, for some objects, the category structure suggested by the training items is consistent with that specified by the rule. A more sensitive test of interference would restrict its consideration to those stimulus items for which interference is reasonably expected. A simple operational definition of this expectation can be based upon the uninstructed participant data, expecting

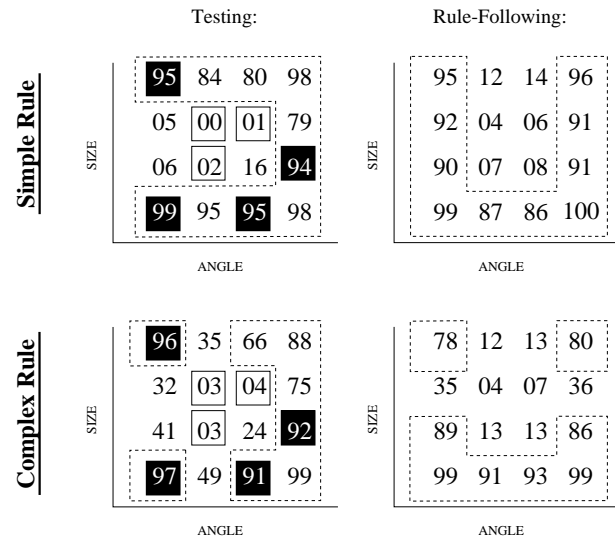


Figure 2: Instructed Conditions Mean Responses: Categorization results are shown as the percentage frequency with which items were placed in the “black” category. Explicitly provided rules are displayed as dashed boundary lines.

interference for those objects which the uninstructed learners, on average, placed in the opposite category as that specified by the explicit rule. The simple categorization rule conflicts with mean uninstructed learner performance at items “E”, “I”, “N”, and “O”. For the complex rule, interference is expected for items “B”, “E”, “I”, and “O”. Restricting our attention to these items, our measure of interference becomes: the increase in mean error, defined as deviation from the rule, from the rule-following phase to the testing phase, averaged only over those items for which interference was expected.<sup>3</sup>

Making use of this more sensitive measure reveals no reliable interference in the simple rule condition ( $M = 0.006$ ;  $SD = 0.244$ ;  $MSE = 0.059$ ;  $F(1, 26) < 1$ ) but substantial interference in the complex rule condition ( $M = 0.220$ ;  $SD = 0.244$ ;  $MSE = 0.060$ ;  $F(1, 18) = 15.478$ ;  $p < 0.001$ ).

Our previous connectionist model of instructed category learning explained exemplar-based interference as the result of connection weight modifications made during the training phase, driven by an error-correcting learning rule (Noelle and Cottrell, 1996). Under this view, large residual errors will produce large weight changes in the network, producing large amounts of interference. Thus, this model gave rise to the prediction that increased error during rule application (estimated by rule-following phase error) should be accompanied by increased exemplar-based interference. Support for this

<sup>3</sup>There may be concern that this measure of interference is inappropriate since different rules were used in the rule-following and testing phases. Indeed, it may seem odd to compare categorization error on a specific stimulus (e.g., item “O”) across the two phases when the relationship of that stimulus to the category boundaries changes between the phases. These concerns may be partially alleviated, however, by noting that none of the significance results reported here change if deviation from the rule is averaged over *all* stimuli in the rule-following phase, and this average deviation is then compared to the average testing phase error on stimuli for which interference is expected.

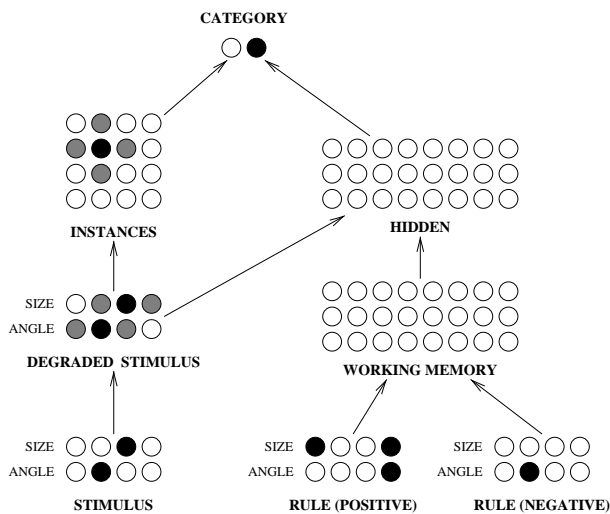


Figure 3: Network Architecture: Each circle represents a standard sigmoidal processing unit, with the units grouped into layers. Arrows represent connections between the units in these layers. Example activations for stimulus “J”, being classified according to the complex rule, are also shown.

prediction may be seen in the difference between the simple rule and the complex rule conditions. The complex rule condition, which elicited a greater degree of rule-following phase error than the simple rule condition (13.5% average error per stimulus object versus 7.9%), elicited a greater degree of interference during the testing phase. This prediction may also be investigated at the level of individual differences. According to the model, poor rule-followers should display more interference than good rule-followers. In order to test this prediction, a correlation was computed over participants in each condition between the mean classification error during the rule-following phase (averaged over all 16 stimuli) and the sensitive measure of interference. A statistically significant positive correlation would verify the prediction. When tested, a marginally reliable *negative* correlation was observed in the simple rule condition ( $r = -0.378$ ;  $t(25) = -2.041$ ;  $p = 0.052$ ), but a robust *positive* correlation was found in the complex rule condition ( $r = 0.629$ ;  $t(17) = 3.340$ ;  $p < 0.01$ ). In brief, the condition which displayed significant interference over all (the complex rule condition) also supported the individual differences prediction of the connectionist model, while the condition which showed no reliable interference (the simple rule condition) revealed a tendency for poor rule-followers to become more consistent with the explicit rule as a result of training.

## A New Connectionist Model

### Simulation Method

Our connectionist model of instructed learning (Noelle and Cottrell, 1996) has been augmented to provide a detailed account of the experimental results reported here, and new model simulations have been conducted. The network architecture used in these simulations is diagrammed in Figure 3. Standard connectionist processing elements were used,

grouped into layers, the activation level of each processing element being the result of applying a logistic sigmoid to the weighted sum of input activity levels. The activity of each unit was, thus, bounded between zero and one.

The network took a representation of categorization instructions and a pair of stimulus features as input and produced a category judgment at the output layer. The input rule representation included eight units corresponding to the four levels of size and the four levels of angle of rotation used in the human learning experiments. Activating one of these units indicated that all stimuli of the given size or of the given orientation were to be placed in the “black” category. These rule input units are shown in Figure 3 as the “RULE (POSITIVE)” layer. Alongside this layer is a collection of inputs, called “RULE (NEGATIVE)”, which encoded “exceptions” to the positive rule terms. Activating one of these eight “negative” units indicated that all stimuli incorporating the given feature level were to be placed in the “white” category, regardless of the category suggested by the positive terms. As an example, the input representation for the complex rule used in our human learning study is shown in Figure 3.

These explicit rule inputs fed activity to a 24 unit working memory layer. When modeling uninstructed learners, the activity levels of these working memory units were set to zero. The weights on the connections from the rule inputs were bound to be non-negative, forcing the working memory layer to encode explicit rule terms by an increase in the activation levels of its processing elements. These weights were initially set to small random values sampled uniformly from the range  $[0.0, 0.4]$ . Complete connectivity extended from the working memory layer to a pool of 24 instruction-sensitive hidden units, and these, in turn, provided activity to the category output units. There were no bounds on these weights. The hidden layer also received complete connections from the degraded stimulus layer, which is described below. All of these unrestricted weights were initialized to small random values, sampled uniformly from the range  $[-0.5, 0.5]$ . Bias values on the working memory and hidden units were initialized to  $-3.0$  in order to encourage sparse internal representations.

Each stimulus was encoded by activating exactly one of the size units and one of the angle units. In order to incorporate perceptual similarity information into the network, this “place coded” stimulus representation was mapped, through connections with fixed weights, to a degraded stimulus representation. In this modified stimulus representation, each unit responded preferentially to a particular stimulus size or stimulus orientation, with partial activity appearing for stimuli of similar sizes or orientations. Levels of partial activation were set to decay exponentially with the number of feature levels separating the given unit from the stimulus. For example, the activity of the “size 2” unit when the stimulus was of “size 4” was set to  $e^{-\beta(4-2)}$ , where  $\beta$  was a gain parameter which could be modified to fit the model to data. Each stimulus dimension, size and orientation, had its own independent gain parameter, making them analogous to the dimensional attention weights used in models like ALCOVE (Kruschke, 1992).

The “instances” layer contained 16 processing elements, with each unit corresponding to one of the possible stimulus objects. Each unit in this layer received input from exactly one size unit in the degraded stimulus layer and from exactly



one angle unit. Unlike other connections in this network, the activity from these two units was multiplied together, rather than summed, to get the resulting activity of the instances layer unit. Thus, each unit in this layer responded preferentially to a unique stimulus object, and activity declined exponentially with city-block distance in feature space. The weights which gave rise to this pattern of activation were fixed. This representational scheme was adopted because of its success in capturing perceived similarity between stimuli in models such as ALCOVE (Kruschke, 1992). The instances hidden layer provided complete connections to the two category output units, with these weights initialized to small random values uniformly sampled from the range  $[-0.5, 0.5]$ . The biases on the output units were initialized to  $-3.0$ .

In order to capture human performance, the network had to be able to apply categorization instructions from the very start of the experimental session. Connection weights which allowed the network to produce accurate categorization decisions immediately following explicit instruction were discovered through a training process conducted during a network initialization phase. During this phase, the network was iteratively presented with a randomly sampled categorization rule along with a randomly sampled stimulus object. It was trained to activate the rule-determined category unit for the given stimulus, modifying weights based on squared error at the output layer using the generalized delta rule (Rumelhart et al., 1986). A learning rate of 0.05 was used, with no momentum term. The gain terms used in the degraded stimulus layer were fixed at 0.8 during this initialization training. This phase continued for 5,000,000 training trials, after which the network consistently demonstrated essentially perfect rule application performance. The distribution of stimuli experienced by the network during this initial training was uniform over the 16 items, but the distribution of categorization rules was skewed towards simple category structures. This biased distribution of rules produced a network which exhibited slightly lower residual error when following a simple rule, as compared to a complex one, and it also encouraged the working memory layer to devote more representational resources to the encoding of simple category structures. The skewed rule distribution also reflected a belief that simple rule structures are much more common in the rule-driven categorization experience of most humans.

Once initialized, the network was presented with the same sequence of trials that was presented to the human learners. When uninstructed, the working memory units were turned off and the network was trained on the seven training items for 252 trials. To measure rule-following performance, the appropriate rule was presented at the input, and category outputs were recorded without performance feedback. To measure performance after both instruction and exemplar-based training, the network was given the appropriate rule at its input and trained on the seven exemplars for 252 trials. The network's performance on all 16 stimuli, without feedback, was then recorded. All exemplar-based training was conducted using the generalized delta rule, with a learning rate of 0.5 and no momentum. Only weights from the instances layer to the output category units, and the bias weights on the output units, were modified during this training process.

To simulate limitations in the cognitive resources applied

to the task, random noise was injected into the activation levels of the processing elements in the working memory layer. During each trial, a random deviant was sampled from a zero-mean normal distribution independently for each unit in the working memory layer and the absolute value of this deviant was subtracted from the activation level of the unit. This caused components of the distributed rule representation held in the working memory layer to become weakened. This use of random noise was intended as a simple and abstract way to capture the temporary failure of the working memory system to actively maintain complete representations of categorization instructions. Resampling the noise on every trial was meant to allow for the possibility of refreshing working memory contents from a longer term episodic memory store.

Network simulations were run 50 times for each experimental condition. Each network was initialized with the same set of connection weights, determined during the initialization phase, but both the injected noise and the order of stimulus presentation was randomized for each simulation. The results of each collection of 50 simulations were averaged to produce figures to be compared to mean human categorization behavior. Four free parameters were adjusted to fit the simulations to data. These included the two gain parameters on the exponential decay used in the degraded stimulus layer representation, the gain parameter used on a Luce choice ratio which converted output activity values to probabilities, and the variance of the noise injected into the working memory layer. A simple grid search was conducted over this parameter space to find values for these four parameters which minimized the squared difference in the probability of "black" category assignment between the human learners and the networks, over all experimental conditions.

## Simulation Results

The best fit of mean simulation results to the human data was had by sharpening the representation of stimulus orientation slightly (gain of 0.9) over the representation of stimulus size (gain of 0.8). This meant that the angle of rotation was slightly more discriminable by the networks than size. The best fit to the mean data required a Luce choice gain of 2.6 on the output activation levels and working memory injected noise with a variance of 0.3. The resulting probabilities of "black" category membership, as predicted by the model, are shown in Figure 4. That diagram also displays the variance accounted for by the model, over stimulus items, for each condition. Notice that, like the humans, the network simulations exhibit no interference in the simple rule case on average (a value of  $-0.017$  in the interference measure previously used with the human data), but they show substantial interference when given the complex rule (0.251).

These simulations involved only a single source of individual variation: the ability to actively maintain an accurate representation of the categorization instructions. Individual differences in exemplar-based interference, then, were to be explained in terms of the weight modifications which were driven by the rule application error introduced by a failure to maintain explicit rules in working memory. Variable working memory ability was reflected by the noise variance parameter in these networks. Thus, in order to examine individual differences in interference, networks with a range of values

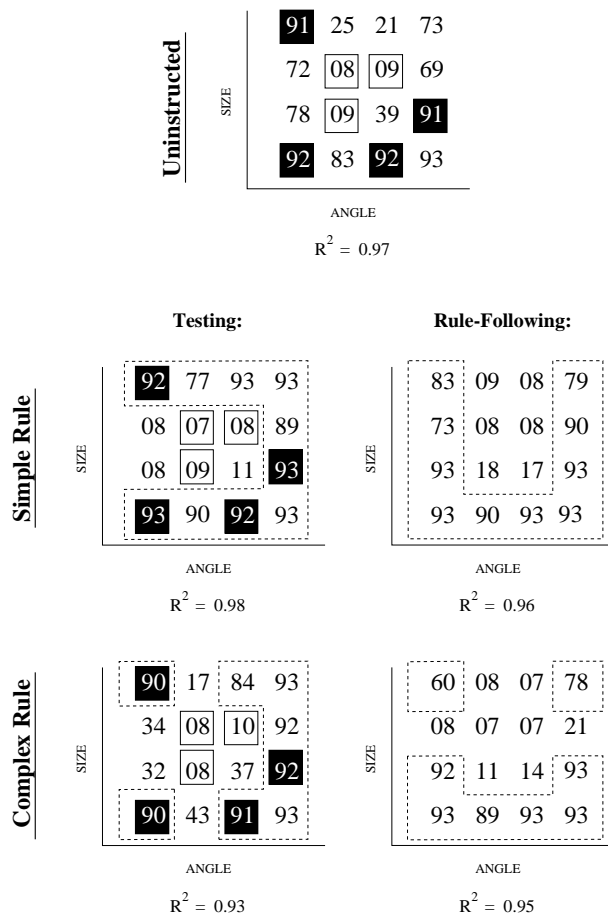


Figure 4: Simulation Results: Categorization results are shown as the mean probability with which items were assigned to the “black” category. Training set items are marked with boxes, colored according to the assigned category label. Variance accounted for is listed for each condition separately.

for this noise parameter had to be compared. Given a collection of networks with a variety of noise levels, the correlation between rule-following error and interference may be measured. Recall that human learners displayed a positive correlation when confronted with the complex rule, but showed a marginally negative correlation when given the simple rule.

When the variance on the noise injected into the working memory units was sampled uniformly from a bound range, these simulations matched the human findings. For example, if noise variance was sampled uniformly from the set  $\{0.0, 0.1, 0.2, 0.3\}$ , then the correlation between rule-following error and interference was reliably negative for the simple rule ( $r = -0.283$ ;  $t(198) = -4.15$ ;  $p < 0.0001$ ) and reliably positive for the complex rule ( $r = 0.412$ ;  $t(198) = 6.37$ ;  $p < 0.0001$ ). Similar results were found when the noise parameter was sampled in a manner sensitive to the observed distribution of human rule-following performance. This sampling was done by finding individual network simulations that matched, as closely as possible, the rule-following phase accuracies exhibited by individual human learners. These best match networks were found by varying the noise variance be-

tween 0.0 and 0.4. When correlations between rule-following error and interference were calculated for such participant-matched samples of network simulations, a positive correlation was found for the complex rule case ( $r = 0.545$ ;  $t(17) = 2.68$ ;  $p < 0.05$ ), and no correlation was found in the simple rule case ( $r = -0.166$ ;  $t(25) = -0.840$ ).

## Conclusions

The magnitude of exemplar-based interference was found to be sensitive to the complexity of the explicitly provided categorization instructions, with more complex categorization rules producing more interference. Also, in situations which elicit robust interference, a reliable correlation across individuals is observed: increased error at explicit rule application is paired with increased exemplar-based interference. A connectionist account of these effects, in which interference arises as the result of an error-correcting learning process, was found to fit the human performance data fairly closely.

## Acknowledgements

This work was supported, in part, by the NIH through a National Research Service Award (# 1 F32 MH11957-01) from the National Institute of Mental Health, awarded to the first author. We extend our thanks to Craig R. M. McKenzie, James L. McClelland, David Plaut, and the members of the UCSD-based *Gary's Unbelievable Research Unit* and the CMU based *PDP Research Group*, as well as to five anonymous reviewers, for their comments on this work.

## References

- Allen, S. W. and Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, 120(1):3–19.
- Brooks, L. R., Norman, G. R., and Allen, S. W. (1991). The role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*, 120(3):278–287.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44.
- Noelle, D. C. and Cottrell, G. W. (1996). Modeling interference effects in instructed category learning. In Cottrell, G. W., editor, *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, pages 475–480, La Jolla. Lawrence Erlbaum.
- Noelle, D. C., Cottrell, G. W., and McKenzie, C. R. M. (2000). Modeling individual differences in the specialization of an explicit rule. (in preparation).
- Nosofsky, R. M., Clark, S. E., and Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(2):282–304.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, chapter 8, pages 318–362. MIT Press, Cambridge, Massachusetts.

# Deep Learning in Virtual Reality: How to Teach Children That the Earth is Round

**Stellan Ohlsson (stellan@uic.edu)**

University of Illinois at Chicago  
Department of Psychology (MC 285)  
1007 West Harrison Street  
Chicago, IL 60607-7137

**Thomas G. Moher (moher@uic.edu)**

**Andrew Johnson (ajohnson@uic.edu)**

University of Illinois at Chicago  
Department of Electrical Engineering and Computer Science (MC 154)  
851 South Morgan Street  
Chicago, IL 60607-7053

## Abstract

To understand deep cognitive change, we have to understand how learners can go beyond their own prior knowledge. We propose a *displacement scenario* in which a learner acquires a target idea in a different context and then transfers that idea into a target context. We used virtual reality technology to implement a displacement scenario for teaching 2nd grade children that the Earth is round. The rather large pre- to posttest improvement was stable over four months.

## The Paradox of Deep Learning

Knowledge systems are organized along a center-periphery axis. One or more *central ideas* dominate more peripheral ones. The center-periphery structure is particularly obvious in scientific theories (Lakatos, 1980), but it also plays an important role in cognitive development (Chi, 1992; Vosniadou, 1994), social cognition (Eagly & Chaiken, in press; Rokeach, 1970) and elsewhere.

Changing the peripheral parts of a knowledge system by learning new facts or skills is easy enough, but revising its core concepts -- *deep learning* -- is a different matter (Ohlsson, 1995). Both direct experiences and communications are interpreted in terms of, and with the help of, prior ideas and hence tend to be understood as consistent with them. The result is that people assimilate information that is anomalous or inconsistent with current ideas or beliefs either by misunderstanding the former or by revising peripheral parts of the relevant knowledge system (Chinn & Brewer, 1993; Darden, 1992; Kuhn, Amsel & O'Loughlin, 1988; Strike & Posner, 1992). Consequently, neither direct experience nor communications have much power to change central ideas. Fodor (1976, Chap. 2) has argued that this is necessarily so: A less powerful

representational system cannot, in principle, replace itself with a more powerful one.

This conclusion leads to a paradox (Bereiter, 1985). It implies that central ideas never change, but of course they do. Scientists sometimes revise fundamental theoretical principles and non-scientists undergo radical changes in world view, particularly during childhood. Developmental psychologists have documented deep changes in children's understanding of a variety of domains (see, e.g., Hirschfeld & Gelman, 1994). Gopnic and Meltzoff (1997) argue that such developmental changes share many features with theory change in science.

How is deep cognitive change possible? How does the mind circumvent the learning paradox? One plausible hypothesis is that ideas that are new in one domain are brought into that domain from some other domain. According to this *cross-domain transfer hypothesis*, to acquire a new central idea in a target domain X, the learner must first acquire that idea in some source domain Y in which its acquisition is not hindered by prior knowledge, and then transfer the new idea to X and build a new understanding of X around it. The new understanding will gradually replace the old. This hypothetical three-step process might circumvent the distorting influence of the learner's prior ideas about X.

This hypothesis predicts that we can facilitate the acquisition of a deep idea if we displace the learner's attention from the target domain to some other domain, teach him or her the target idea in that domain, and then prompt him or her to transfer it into the target domain. We implemented this *displacement scenario* in a virtual reality environment for teaching children that the Earth is round. Empirical evaluation in a public school resulted in strong

and lasting improvement in the children's understanding of the shape of the Earth and related facts.

### **Mental Models of the Earth**

All direct experience supports the idea that the ground is a flat surface extending in all directions; hills and valleys are only local perturbations. The sky is parallel to the ground, the ground is always *down* and the sky is always *up*.

These ideas partition the universe into two unequal regions, above and below the Earth. They strongly imply that traveling in a straight line will bring the traveler further and further away from his or her starting point, until he or she reaches a boundary where the Earth stops. Furthermore, *down* and *up* do not vary with the observer's location; an arrow pointing upwards in one location is parallel to an arrow pointing upwards in any other location. Also, objects at a distance are hard to see either because they are occluded by another object or because the observer lacks visual acuity. Finally, the location of the sun and the moon when we cannot see them is problematic. Many children in Western (Nussbaum, 1985; Vosniadou & Brewer, 1992) as well as non-Western (Vosniadou, 1994) societies develop some version of this mental model.

The idea that the Earth is spherical has contrasting implications: It suggests that the surrounding space is uniform and it implies that a traveler who keeps going in a straight line will eventually return to his or her starting point. Furthermore, *down* and *up* varies with the observer's location; *up* in New York is not parallel to *up* in Hong Kong. Also, distant objects are invisible because they are occluded by the surface curvature. Finally, the sun and the moon are sometimes invisible because they are occluded by the planet itself.

The shift from a flat Earth to a round Earth view is an instance of deep learning. The two concepts, clearly stated, contradict each other and they influence many other aspects of one's understanding of Earth-related facts and events. Empirical research has shown that this shift takes considerable time, at least two years (Vosniadou & Brewer, 1992, Table 4) and possibly as long as six years (Nussbaum, 1985, Fig. 9.16) when it occurs spontaneously, and it requires one or more intermediate mental models. The question is whether this process can be speeded up with the displacement strategy.

### **A Virtual Asteroid**

Our approach to facilitating the shift from a flat to a round Earth is to teach the idea of a spherical planet in an unfamiliar context, unhindered by prior ideas, and then prompt the learner to apply this idea to his or her knowledge about the Earth. We accomplished the first step in this two-step procedure by using two linked virtual reality (VR) environments. The *Asteroid World* simulates the experience of walking on the surface of an asteroid with approximately 300 yards diameter. The virtual asteroid is roughly spherical

in shape and exhibits a desert-like landscape with a handful of geographical features (a bulge, a canyon, etc.), large rocks scattered here and there and fantasy structures that resemble trees made out of crystal, plus a shuttle-like space ship. The sky is black but features stars and a large, moon-like object. The *Asteroid World* was presented via a so-called *ImmersaDesk*, a VR projection device developed at the Electronic Visualization Laboratory at UIC. The *ImmersaDesk* is roughly 6 feet by 4 feet. The device supports full immersive VR with stereo vision, head tracking, hand tracking and audio; see Czernuszenko, Pape, Sandin, DeFanti, Dawe and Brown (1997) for a technical description.

When the *Asteroid World* user presses the forward-move button on the control stick, he or she has the visual perceptions that would be associated with a physical walk on a real asteroid with the same properties as the virtual one. When the diameter of the world is 300 yards, one can experience its sphericity directly. The horizon is very close, rocks and other large objects appear over the horizon very quickly, the stars in the sky are streaming past at a perceptible pace, objects are difficult to find because they are hidden by the curvature even when close by and circumnavigation is accomplished in a couple of minutes.

Our second environment, called the *Mission Control*, presents a satellite view of the virtual asteroid, projected in stereo on a computer monitor. When the user wears stereo glasses, he or she sees the virtual asteroid as a three-dimensional body floating in space against the background of stars. The various geographical features and the space ship are clearly visible. In addition, the *Mission Control* user sees the user of the *Asteroid World* as an avatar, a small space-suited figure. That is, the *Asteroid World* user and the *Mission Control* user access the same virtual reality at the same time but from different points of view. In particular, *Mission Control* can observe the movements of the astronaut on the virtual asteroid in real time. To remain in visual contact, *Mission Control* can rotate the asteroid (but not change his or her distance from it) by pressing a button on a control stick.

The *Asteroid* and *Mission Control* environments are described in more detail in Johnson, Moher, Ohlsson and Gillingham (1999). By alternating between them, the learner can experience or perceive the uniformity of the surrounding space, circumnavigation, the relativity of up and down, and occlusion by surface curvature. Furthermore, these experiences occur in a context in which the learner has no prior, conflicting ideas about the shape of the world. The second step in our learning scenario -- to transfer and apply this idea to the everyday experience of the Earth -- is described below.

## **Empirical Study**

### **Method**

**Materials** The equipment needed to project the two virtual environments was set up in a large room in a public school in a Chicago suburb. The user of one environment could not see the other environment or its user, but the two users were close enough so that they could talk to each other.

In addition, our instructional procedure required two physical models. One was a foam rubber model of the virtual asteroid, approximately eight inches in diameter, painted and equipped with a model space ship, rocks and other features to make it recognizable as a model of the virtual asteroid as seen in the Mission Control environment. The second physical model was a standard Earth globe purchased in a book store.

**Knowledge test** To assess children's understanding of the shape of the Earth, we developed a structured interview derived from those used by previous researchers (Nussbaum, 1985; Vosniadou & Brewer, 1992). The interviewer (a project team member) asked 18 questions about the shape of the Earth, the content of the region below the Earth, circumnavigation, the relativity of up and down and occlusion by curvature. The children's answers were classified at testing time by the interviewer, using a set of coding categories derived from a pilot study (Johnson, Moher, Ohlsson & Gillingham, 1999). The knowledge test interview took 10-20 minutes. The same test was used as pretest, posttest and delayed posttest.

**Subjects** All fifty second-grade children in the participating class rooms were pretested. The 28 children who answered 10 or fewer pretest questions correctly were included in the *treatment group*. Due to the small number of such students, we preferred to include all of them in a pretest-posttest design over dividing them into two groups in a treatment-control design. The 22 children who answered 11-13 questions correctly will be referred to as the *comparison group*, although it is not a control group in the statistical sense due to the non-random group assignments.

**Procedure** For the children in the treatment group, the procedure consisted of pretest, VR experience, bridging activity, posttest and delayed posttest. For the children in the comparison group, the procedure consisted of pretest and posttest.

(a) **VR experience.** The children were paired into teams of two. During the *familiarization phase*, the two experimenters who acted as guides helped the children put on the stereo glasses and guided them around their respective environments for five minutes. The two children then switched places and the familiarization process was repeated for another five minutes. During familiarization, the guides pointed out visual features related to sphericity (nearness of horizon, objects coming up over the horizon, the avatar seeming to be up side down, circumnavigation, etc.).

During the *game phase*, the children were told that they were stranded on the asteroid for lack of fuel and their task was to find extra fuel cells scattered over the asteroid so that their space ship could return to Earth. The child on the

asteroid collected the fuel cells, but the child in Mission Control assisted by locating fuel cells (the latter were clearly visible in the Mission Control view) and by giving directions to the other child. The children played this game for ten minutes, switched places and continued for an additional ten minutes. Each child thus had a total of 30 minutes (5+5+10+10) of interaction with the two VR environments.

(b) **Bridging dialogue.** Immediately after the VR experience, the two children were escorted to two different rooms for the *bridging dialogue*, a structured conversation with a member of the project team. The purpose of this dialogue was to prompt reflection on the VR experience and to help the child transfer the spherical planet idea to his or her mental model of the Earth. In each phase of the dialogue, the experimenter reminded the child of his or her VR experience with the help of the physical model of the asteroid, re-enacting some facet of that experience (e.g., circumnavigation) with toy figures. The experimenter then shifted the child's attention to the globe of the Earth and told him or her that what was the case on the asteroid is also the case on the Earth, enacting the relevant facet with toy figures vis-à-vis the Earth globe. The conversation then switched back to the asteroid model to cover another facet of sphericity, which was also illustrated with the Earth globe; and so on. The bridging dialogue took approximately 15 minutes.

(c) **Posttest.** The subjects were posttested 24 hours after the learning experience.

(d) **Delayed posttest.** The delayed posttest was administered four months after the learning experience.

## Results

Figure 1 shows the outcome. The performance of the treatment group increased from a mean of 7.3 correct answers on the pretest to a mean of 12.9 correct answers on the posttest. We tested the posttest mean with a single-sample t-test, using the pretest mean as the comparison value. The difference is statistically significant ( $t = 13.68$ ,  $p < .000$ ). Hence, the treatment group improved from pretest to posttest. The magnitude of the improvement is  $12.9 - 7.3 = 5.6$  scale units, which is 1.9 times the standard deviation on the pretest. The mean number of correct answers on the delayed posttest was 11.4. Almost the entire pre- to posttest improvement was retained four months later.

Because the posttest questions were identical to the pretest questions, there is a possibility that the improvement in the children's understanding of the Earth was caused by the test itself. We can use the comparison group to measure the effect of the test. The members of the comparison group were pre- and posttested but did not undergo the VR experience. The mean number of correct answers in this group was 12.2 on the pretest and 14.0 on the posttest. A single-sample t-test of the posttest mean, using the pretest mean as comparison value, showed that the pre- to posttest difference is statistically significant ( $t = 4.6$ ,  $p < .000$ ).

Hence, taking the test prompted some learning, even in the absence of the VR experience. The magnitude of the effect is  $14.0 - 12.2 = 1.8$ , which is .6 times the standard deviation on the pretest. This improvement is considerably smaller than the improvement in the treatment group. Due to the non-random assignment of subjects to groups, the evidence provided by this analysis is admittedly weaker evidence than that provided by a proper control group.

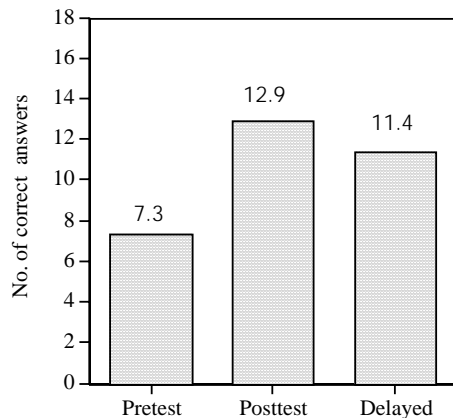


Figure 1. The mean number of correct answers on three test occasions.

A t-test for independent samples shows that the difference between the treatment and comparison groups on the pretest was statistically significant ( $t = 10.71, p < .000$ ). There was no significant difference between the two groups on the posttest ( $t = 1.90, p > .06$ ).

## Discussion

The children in the treatment group almost doubled their understanding of the shape of the Earth, as measured by our knowledge test. The treatment group initially performed considerably below the comparison group, but performed as well as the latter on the posttest. That is, our learning scenario allowed those children who had not spontaneously acquired an understanding of the shape of the Earth to catch up with those who had. Unlike the spontaneous acquisition process, which occurs over several years (Nussbaum, 1985; Vosniadou & Brewer, 1992), the displacement scenario enabled children to acquire the target idea in one day. They retained it four months later.

Why was the displacement scenario successful? An explanation for these results must deal with the paradox of deep learning: Central ideas are seldom transformed by novel input; they are too protected by the surrounding belt of auxiliary ideas and beliefs. So how does deep learning ever come about? The cross-domain transfer hypothesis claims that central ideas are not transformed but replaced by ideas transferred from other contexts, domains or situations (Chi, 1992). In the present study, both our virtual asteroid and the Earth can be said to belong to the domain of elementary astronomy, but the crucial point for learning is that our

subjects had no prior knowledge about the shape of the virtual asteroid but they did about the shape of the Earth.

This model of deep learning differs significantly from other models, e.g., attempts to view deep learning in children as analogous to scientific theory change (Gopnik & Meltzoff, 1997; Hewson & Hewson, 1984; Posner et. al, 1982). One difficulty with this theory theory, as it has come to be known, is that human beings are not conspicuously good at evaluating evidence, presumably the central process in theory change. The theory theory describes cognitive change in logical rather than naturalistic terms (Ohlsson, 2000). It does not explain our results, because we did not present our subjects with evidence of any kind: We familiarized them with a previously unfamiliar environment and then asserted that what was true in that environment is also true about the Earth. The cross-domain transfer hypothesis does better because, unlike the theory theory, it does not claim that dissatisfaction with prior ideas is a prerequisite for learning. Prior ideas are not necessarily falsified or rejected; instead, they fall into disuse when another, more useful idea becomes available.

Unlike the knowledge-in-fragments theory of DiSessa (1988, 1993) and Smith, DiSessa and Roschelle (1995), the present theory does not represent deep learning as a process of clarifying, organizing and systematizing so-called phenomenological primitives. Instead, it claims that a central idea that has been transferred from a different context can serve as a starting point for a new understanding of the target context. One difficulty with the knowledge-in-fragments view is that it is unclear how systematizing and organizing can engender a new idea that directly contradicts one of the ideas available at the outset. For example, it seems implausible that experience of the virtual asteroid would prompt our subjects to organize their no doubt fragmented knowledge of the Earth in such a way that they suddenly realized that it must be spherical.

Although our results are more consistent with the cross-domain transfer hypothesis than with these alternative hypotheses, the present study is limited in several respects. The number of children was small, we had no proper control group and the results do not allow us to separate the effects of the virtual reality experience from the effects of the bridging dialogue. We are currently completing a follow-up study that addresses these limitations.

In addition to its theoretical interest, the cross-domain transfer hypothesis might have practical importance. It is a commonplace in educational discourse that good instruction should connect to the students' prior knowledge and experience. However, this pedagogical tactic is unlikely to be productive in those situations in which the target subject matter conflicts with the students' prior knowledge (Ohlsson, 1999; Strike & Posner, 1992). The alternative is to teach the new idea in a different context and help the student transfer it to the target domain. Because many scientific ideas conflict with ideas derived from experience (e.g.,

inertia), the displacement scenario has the potential to be a useful tool in science education.

### Acknowledgments

The work reported here was supported, in part, by grant #EIA 9720352 from the Learning and Intelligent Systems program to Thomas DeFanti and, in part, by grant #BCS 9907839 from the Child Learning and Development program to the first author. Both funding programs are part of the National Science Foundation (NSF).

### References

- Bereiter, C. (1985). Toward a solution of the learning paradox. *Review of Educational Research*, 55, 201-226.
- Chi, M. T. H. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. N. Giere, (Ed.), *Cognitive models of science*. Minneapolis, Minnesota: University of Minnesota Press.
- Chinn, C., & Brewer, W. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, 63, 1-49.
- Czernuszenko, M., Pape, D., Sandin, D., DeFanti, T., Dawe, G. L., & Brown, M. D. (1997). The *ImmersaDesk* and *Infinity Wall* projection-based virtual reality displays. *Computer Graphics*, 31(2), 46-49.
- Darden, L. (1992). Strategies for anomaly resolution. In R. Giere, (Ed.), *Cognitive models of science*. Minneapolis, MN: University of Minnesota Press.
- DiSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. Pufall, (Eds.), *Constructivism in the computer age*. Hillsdale, NJ: Erlbaum.
- DiSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, 10, 105-225.
- Eagly, A. H., & Chaiken, S. (in press). Attitude strength, attitude structure, and resistance to change. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences*. Hillsdale, NJ: Erlbaum.
- Fodor, J. A. (1976). *The language of thought*. Sussex, UK: Harvester Press.
- Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.
- Hewson, P. W., & Hewson, M. G. A. (1984). The role of conceptual conflict in conceptual change and the design of science instruction. *Instructional Science*, 13, 1-13.
- Hirschfeld, L., & Gelman, S. (1994). *Mapping the mind*. Cambridge, UK: Cambridge University Press.
- Johnson, A., Moher, S., Ohlsson, S., & Gillingham, M. (1999, November/December). The Round Earth project: Collaborative VR for conceptual learning. *IEEE Computer Graphics and Applications*, pp. 60-69.
- Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills*. San Diego, CA: Academic Press.
- Lakatos, I. (1980). *The methodology of scientific research programmes*. Cambridge, UK: University of Cambridge Press.
- Larkin, J., & Chabay, R., (Eds.), (1992). *Computer-assisted instruction and intelligent tutoring systems*. Hillsdale, NJ: Erlbaum.
- Nussbaum, J. (1985). The Earth as a cosmic body. In R. Driver, E. Guesne, & A. Tiberghien (Eds.), *Children's ideas in science*. Milton Keynes, UK: Open University Press.
- Ohlsson, S. (1995). Learning to do and learning to understand: A lesson and a challenge for cognitive modeling. In P. Reimann and H. Spada, (Eds.), *Learning in humans and machines: Towards an interdisciplinary learning science*. Oxford, UK: Elsevier.
- Ohlsson, S. (1999). Theoretical commitment and implicit knowledge: Why anomalies do not trigger learning. *Science & Education*, 8, 559-574.
- Ohlsson, S. (2000). Falsification, anomalies and the naturalistic approach to cognitive change. *Science & Education*, 9, 173-186.
- Posner, G., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accomodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, 66, 211-227.
- Rokeach, M. (1970). *Beliefs, attitudes and values*. San Francisco, CA: Jossey-Bass.
- Smith, III, J. P., DiSessa, A. A., & Roschelle, J. (1995). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, 3, 115-163.
- Strike, K. A., & Posner, G. J. (1992). A revisionist theory of conceptual change. In R. A. Duschl & R. J. Hamilton, (Eds.), *Philosophy of science, cognitive psychology, and educational theory and practice*. New York: State University of New York Press.
- Vosniadou, S. (1994). Universal and culture-specific properties of children's mental models of the earth. In L. Hirschfeld and S. Gelman (Eds.), *Mapping the mind*. Cambridge, UK: Cambridge University Press.
- Vosniadou, S., & Brewer, W. F. (1992). Mental models of the Earth: A study of conceptual change in childhood. *Cognitive Psychology*, 24, 535-585.

# ANCHOR: A Memory-Based Model of Category Rating

Alexander A. Petrov (apetrov@andrew.cmu.edu)

Department of Psychology; Carnegie Mellon University  
Pittsburgh, PA 15213 USA

John R. Anderson (ja+@cmu.edu)

Department of Psychology; Carnegie Mellon University  
Pittsburgh, PA 15213 USA

## Abstract

This paper attempts to draw a bridge between psychophysics and memory research by proposing a memory-based model of category rating. The model is based on the cognitive architecture ACT-R and uses *anchors* stored in memory that serve as prototypes for the stimuli classified within a response category. The anchors are retrieved by a partial matching mechanism and updated dynamically by an incremental learning mechanism. Anchors also have base-level activations that reflect the frequency and recency of the responses. These mechanisms give rise to sequential effects and nonuniform response distributions. A psychological experiment involving category rating of physical length is reported and the predictions of the model are compared against the empirical data. The psychophysical implications of the model are discussed.

## Introduction

Category rating is a widely used method of data collection in experimental psychology. A category-rating situation arises whenever the participants are asked to assign each stimulus to one of several ordered categories such as *1, 2, ..., 9* or *very dissimilar, ..., very similar*. Procedures of this kind are common for many studies ranging from psychophysical scaling to similarity judgment to personality inventories. Therefore a detailed analysis of the cognitive mechanisms underlying this task is potentially relevant to a diverse set of situations.

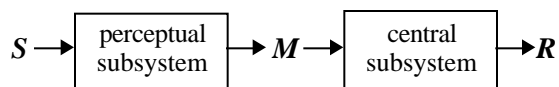


Figure 1: Simplified decomposition of the category-rating process. The external stimulus  $S$  maps to an internal magnitude  $M$  which in turn gives rise to the overt response  $R$ .

A rough decomposition of the process of category rating is presented in Figure 1. (This diagram is by no means complete or accurate; it is provided for expository purposes only.) The *perceptual subsystem* maps the external stimulus  $S$  onto an internal representation  $M$  on a psychological continuum. In this paper the internal representation is called *magnitude*. The magnitude  $M$  then serves as a basis for generating an overt response  $R$  on the category scale. The latter

transformation is the responsibility of the *central* (or *cognitive*) *subsystem*. Both subsystems are characterized with internal states that unfold in time and may differ from trial to trial. Thus each box in Figure 1 has underlying dynamics and the whole system is more complex than the open-loop pipeline suggested by the diagram.

The present paper focuses on the central subsystem and the computational mechanisms converting subjective magnitudes into external reports. While the perceptual aspects of the process are certainly important, they are not central to the research reported here. Therefore the research strategy has been to try to minimize the contribution of the perceptual subsystem so that the properties of the central one can show through. This dictated the choice of a modality for which the perceptual transformation is as simple as possible—physical length.

The empirical relation between stimulus intensities  $S$  and averaged category ratings  $R$  tends to follow a power function:  $R = k \cdot S^n$  (Stevens, 1957). The exponent  $n$  is characteristic of the perceptual modality. For physical length, this exponent is very close to 1.0 (Stevens, 1957). In other words, the scale is linear. Thus it seems reasonable to assume that the perceptual subsystem delivers veridical representations of physical length, with little if any systematic distortions (Krantz, 1972). Under this assumption, any patterns in the category-rating data for length are largely due to the central subsystem.

The psychophysical literature reports several phenomena related to category rating. The most basic finding is that the participants are able to perform this task without major difficulties and provide robust and regular data: the average rating values vary smoothly with stimulus intensity (Stevens, 1957). This is true whether or not feedback is provided (e.g. Ward & Lockhead, 1970). The second major finding is Stevens' power law stated above. In addition to these first-order results, there are several second-order effects as well.

The *sequential effects* are of special interest here because they shed light on the dynamics of the rating process. Numerous studies have indicated that the successive trials in a rating experiment are not independent (Ward & Lockhead, 1970; Jesteadt et al., 1977; Petzold, 1981; Schifferstein & Frijters, 1992). The responses, regarded as a time series, show autocorrelational structure. Typically the data are analyzed using multiple regression in which the stimulus  $S_{t-1}$  and the response  $R_{t-1}$  on the preceding trial enter as predictors after the contribution of the current stimulus  $S_t$  has



been partialled out. A robust finding is that current responses tend to be contrasted (i.e. negatively correlated) with previous stimuli and assimilated (positively correlated) toward previous responses. Moreover, there is an interaction between the two time-lagged variables  $S_{t-1}$  and  $R_{t-1}$ . The assimilation towards the previous response seems to be modulated by the difference between the two consecutive stimuli  $S_{t-1}$  and  $S_t$  (Jesteadt et al., 1977; Petzold, 1981). The closer the stimuli, the stronger the assimilation.

Theoretical analysis of the task also invites the hypothesis that some form of memory is involved in the rating process. Consider a trial in a category-rating experiment. The presentation of the stimulus evokes some subjective percept in the participant. The participant is then faced with the problem of communicating this subjective percept using the particular response scale chosen by the experimenter. There is no a priori correspondence between the subjective magnitudes and the response categories. Such correspondence must be established at the beginning of the experiment and then applied consistently until the end. This is a role for memory.

This hypothesis is supported by a study of Ward and Lockhead (1970). The experiment involved 8 sessions on 8 consecutive days. Feedback was provided at the end of each trial. Unbeknown to the participants the feedback was manipulated so that the response categories were associated with different stimuli on different days. This caused systematic shifts in participants' responses.

The thesis of the present paper is that memory plays an important role in category rating and in particular in the transition from internal magnitudes to overt responses. Memory maintains the consistency of responses over periods of hours and even days. Moreover, the hypothesis is that failures to achieve perfect consistency—manifested as response drifts, sequential effects, and context effects—are due to the plasticity of the memory system and reflect the dynamics of its operation.

This paper reports the initial steps towards a memory-based theory of category rating. The theory is instantiated in a computational model called ANCHOR and the predictions of the model are compared with empirical data.

## Psychological Experiment

The ANCHOR model makes detailed predictions on a trial-by-trial basis. To estimate the parameters of the model and evaluate its adequacy as a psychological theory one needs empirical data at the same level of granularity. The psychophysical literature cited in the introduction reports aggregate data only and hence falls short of this standard. Therefore, a psychological experiment was carried out. In addition to providing the necessary data, it replicates the sequential effects from the literature and tests the assumption of linearity of the scale of physical length.

## Method

**Stimulus Material.** The stimuli were pairs of white dots presented against black background on a 17-inch Apple-Vision monitor. The only independent variable in the experiment was the distance between the two dots measured in pixels. The distance used on each trial was drawn independ-

ently from a uniform distribution ranging from 250 pixels (80 mm) to 700 pixels (224 mm). The viewing distance was approximately 500 mm. The imaginary segment formed by the dots was always horizontal and was randomized with respect to its absolute horizontal and vertical position on the screen. The stimulus set for each participant was generated and randomized separately. The maximal distance representable on the monitor was 1000 pixels (320 mm). Each dot was roughly circular in shape with a diameter of 16 pixels (5 mm).

**Participants:** 24 students participated in the experiment to satisfy a course requirement.

**Procedure.** The participants were asked to rate the “distance between the dots” on a scale ranging from 1 to 9. The participants entered their responses on the numeric keypad of the computer keyboard. Each trial began with a 500 ms beep followed by 3300 ms stimulus presentation followed by 200 ms inter-trial interval. There were 17 demonstration and 450 experimental trials divided into 10 blocks with short rest periods between the blocks. The demonstration presented stimuli of length 275, 325, 375, ..., 625, 675, 625, ..., 275 pixels and the participants were encouraged to practice pressing the keys 1, 2, ..., 8, 9, 8, ..., 1. No feedback was given during the experimental trials. The whole procedure lasted about 40 minutes.

## Results and Discussion

The data are analyzed at the level of individual participants.

**Linearity of the Scale.** To estimate the exponent of Stevens' power law, a function of the form  $R = a + k \cdot S^n$  is fitted to the data of each individual participant. The exponents  $n$  range from 1.01 to 1.12 in the sample of 24 participants, with mean 1.06. Thus the exponent is empirically indistinguishable from unity for all participants. (The correlations between the functions  $S^{0.95}$ ,  $S^{1.00}$ , and  $S^{1.10}$  are greater than 0.99 in the domain [250;700].) This suggests that the assumption of linearity of the scale is correct, at least within the precision of measurement.

**Overall Accuracy.** The linearity of the scale allows the data to be analyzed by simple linear regression of  $R$  on  $S$ . The squared correlation coefficient  $R^2$  is a measure of the accuracy of the respective participant. It ranges from 0.65 to 0.91 for the 24 participants, with mean 0.80 and std.dev. 0.070. In other words, the immediate stimulus accounts for full three quarters of the response variance, sometimes up to 90%.

**Response Distributions.** Even though the stimuli are uniformly distributed, the responses are not. Figure 2 shows the response distributions for two representative participants. A marked feature of these distributions is the predominance of responses in the middle of the scale at the expense of extreme ones. The response standard deviation ranges from 1.20 to 2.44, with mean 1.96 and s.d. 0.28. For comparison, if the 450 responses were evenly distributed in 9 categories, the standard deviation would be 2.58.

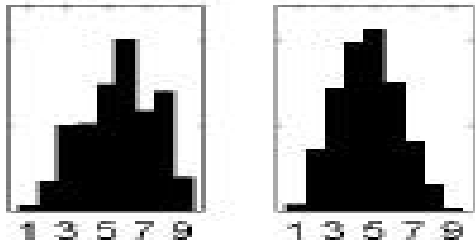


Figure 2: Response distributions for two representative participants.

It seems unlikely that the perceptual subsystem maps the uniform stimulus distribution onto a highly non-uniform distribution of internal magnitudes. Therefore the shape of the response distribution appears to be largely due to the cognitive subsystem. It is possible that the participants reserve the extreme responses for distances that are very short (close to zero) or very long (filling the width of the screen). Such extreme stimuli are not presented during the experiment and this may be one of the reasons for the non-uniformity of responses. However, this explanation does not address the peak in the middle of the scale. The memory-based theory of category rating offers an alternative explanation in terms of self-reinforcing buildup of strength for the frequent responses and corresponding loss of strength for the infrequent ones.

**Sequential Effects.** A multiple linear regression is performed with the following variables entering as predictors: the current stimulus  $S_t$ , the previous stimulus  $S_{t-1}$ , and the previous response  $R_{t-1}$ . The signs of the regression coefficients of the time-lagged variables are of special interest. For the previous stimulus  $S_{t-1}$ , the standardized coefficient  $\beta_s$  ranged from  $-0.53$  to  $-0.08$ , with mean  $-0.25$  and s.d.  $0.10$ . Conversely, the standardized coefficient  $\beta_r$  for the previous response  $R_{t-1}$  ranged from  $+0.15$  to  $+0.55$ , with mean  $+0.30$  and s.d.  $0.10$ . Thus all 24 participants without exception show evidence of stimulus-driven contrast and response-driven assimilation.

Additional regression analyses involving interaction terms replicate the finding of Jesteadt et al. (1977) that the assimilation towards  $R_{t-1}$  is modulated by the difference between the two consecutive stimuli  $S_{t-1}$  and  $S_t$ . These analyses are not reported here because of lack of space.

## Memory Based Model of Category Rating

As argued in the introduction, memory seems to play an important role in the category-rating process. The remainder of this paper outlines one particular proposal about the computational mechanisms that may carry out this process. The ANCHOR model proposed here is based on a general theory of memory incorporated in the ACT-R cognitive architecture (Anderson & Lebière, 1998). The ACT-R theory is consistent with a broad range of memory phenomena. Thus ANCHOR draws a bridge between psychophysics and memory research. The following two subsections describe the model first in general terms and then with details and equations.

## Main Principles of the Model

The centerpiece of the ANCHOR model is the construct of an *anchor*. An anchor is an association between an internal magnitude and a category on the response scale. There is one anchor per category and it can be construed as an internal representation of the prototypical member of this category.

The collection of all anchors defines a mapping from the continuum of magnitudes to the discrete categories of the response scale. This mapping is partly constrained and partly arbitrary. The constraints come from the demand for homomorphism implied by the category-rating task. There is intrinsic ordering of the intensity of the physical stimuli and hence of the magnitudes on the subjective continuum. Also, there is ordering of the response categories. When reporting their subjective magnitudes, the participants try to align the ordering of the two domains.

Another constraint implied by the task is to maintain consistency over time. If, for whatever reason, a stimulus is labeled with a particular response on a given trial, there is pressure to label this stimulus with the same response on subsequent trials. This extends not only to the stimulus that happened to be presented but to other stimuli that evoke similar subjective magnitudes.

These constraints motivate the following mechanisms of the ANCHOR model. When a stimulus is presented and encoded as an internal magnitude, a *partial matching mechanism* activates an anchor whose magnitude is similar to the magnitude of the target stimulus. In so far as anchor magnitudes are relatively stable, categorization of the stimuli is consistent over time.

The partial matching is stochastic and depends on other factors besides similarity (viz. recency and frequency, discussed below). Therefore it is not guaranteed to retrieve on each trial the anchor that best matches the target magnitude. In the cases when there is large discrepancy between the target magnitude evoked by the stimulus and the anchor magnitude retrieved from memory, a *correction mechanism* may increment or decrement the response suggested by the anchor. The correction mechanism is stochastic and error-prone too but it does tend to enforce homomorphism between magnitudes and responses.

Phenomenologically, an introspective report of a category-rating trial might run like this, “I see the dots... The distance looks like a 7... No, it’s too short for a 7. I’ll give it a 6.”

So, the stimulus has been encoded, matched against anchors, and a response has been produced. Is this the end of the trial? According to the ANCHOR model and the broader ACT-R theory (Anderson & Lebière, 1998), the answer is no. The cognitive system is plastic (within limits) and each experience seems to leave a mark on it. It is impossible to step into the same river twice. The model postulates an obligatory *learning mechanism* that pulls the magnitude of the relevant anchor in the direction of the magnitude of the stimulus that has just been presented. Thus each trial results in a slight change of the magnitude of one of the anchors—namely the one that corresponds to the response given on that particular trial. The notion of obligatory learning is similar to the ideas of Logan (1988), although ANCHOR learns prototypes rather than individual instances.

The implications of this incremental learning mechanism are worth considering in detail. After a long sequence of trials, each anchor magnitude ends up being a weighted average of the magnitudes of all stimuli classified in the corresponding response category. Thus the anchors are true prototypes. However, recent stimuli weigh more heavily than earlier ones, introducing bias. The influence of the initial instructions and demonstrations gradually wash away.

More importantly, the performance of the system on each trial depends on the history of its performance on previous trials. This makes it a dynamic system capable and even forced to exhibit gradual shifts, sequential effects, and self-reinforcing preferences. Each run of the model becomes idiosyncratic in systematic ways apart from the random noise even when tested on the exact same sequence of stimuli.

One final aspect of the model remains to be introduced. There is abundant evidence that the human memory system is sensitive to the frequency and recency of the encoded material. These two factors enter the ACT-R theory and the ANCHOR model through a construct called *base-level activation* (BLA). Each memory element, anchors included, has some base-level activation that goes up and down with time. The partial matching mechanism is sensitive not only to the similarity between the target magnitude and the anchor magnitudes but also to the activation levels of the anchors. Overall, anchors with high BLA are more likely to win in the matching process than anchors with low BLA, the target stimulus notwithstanding.

The form of the base-level learning equation (Eq. 6 below) entails that when a response is produced on a trial the BLA of the corresponding anchor receives a sharp transient boost followed by small residual increase. On the other hand, when some response is not used for a long time the activation of the corresponding anchor gradually decays away. In terms of observable behavior, the rapid transient manifests itself as sequential response assimilation and the long-term overall strength leads to rich-gets-richer differentiation of the response frequencies.

## Details and Equations

Figure 3 shows a schematic diagram of the various quantities used in the model and the dependencies among them.

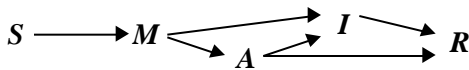


Figure 3: Schematic diagram of the quantities used in the model: physical intensity of the stimulus  $S$ , target magnitude  $M$ , anchor magnitude  $A$ , increment  $I$ , and overt response  $R$ .

The perceptual subsystem (cf. Figure 1) is modeled by a single equation [1]. It transforms the physical intensity of the stimulus  $S$  into an internal magnitude  $M$ . The transformation is linear, with some multiplicative noise. The magnitudes are arbitrarily scaled between 0.25 and 0.70, given that  $S$  varies between 250 and 700 pixels. The random variable  $\epsilon$  is normally distributed with zero mean. Thus the term  $(1+\epsilon)$  is centered around 1.0. The standard deviation of

the noise is a free parameter of the model. In the simulation experiments reported in the next section this parameter was set to 0.050. The multiplicative relationship between the scale value (i.e. the mean of the magnitude distribution induced by a given stimulus  $S$ ) and the noise term implements Ekman's law (Ekman, 1959).

$$M = S \cdot (1 + \epsilon) / 1000 \quad [1]$$

There are 9 anchors with magnitudes  $A_1 \dots A_9$ , respectively. The partial matching mechanism has to select one of them according to their similarity to the target magnitude  $M$  and their base-level activations  $B_1 \dots B_9$ . This process is governed by two equations. First, a *score* is produced for each anchor according to Eq. 2. Second, one anchor is chosen according to the *softmax* Equation 3.

$$Score_i = B_i - MP \cdot |M - A_i| \quad [2]$$

The *mismatch* (or dissimilarity) between two magnitudes is simply the absolute difference between them. The mismatch is multiplied by a *mismatch penalty factor*  $MP$  and subtracted from the base-level activation of the anchor to produce the combined score for this anchor.  $MP$  is a free parameter of the model that scales the mismatches relative to the activation values. It was set to 7.0 in the simulations.

$$P_i = \exp(Score_i / t) / \sum_j \exp(Score_j / t) \quad [3]$$

Equation 3 converts scores into retrieval probabilities.  $P_i$  is the probability of retrieval of anchor  $i$  and  $\exp(\cdot)$  denotes the exponentiation function. The *temperature*  $t$  is a free parameter of the model controlling the degree of non-determinism of the partial-matching process. It was set to 0.40 in the simulations.

Having retrieved an anchor, the model has to determine the correction  $I$  to produce the final response. Under the current settings of the model, the correction can be 0,  $\pm 1$ , and occasionally  $\pm 2$ . The correction depends, stochastically, on the discrepancy between the target magnitude  $M$  and the anchor magnitude  $A$ . One free parameter of the model— $d$ —defines a set of five *discrepancy reference points*  $\{-2d, -d, 0, d, 2d\}$ . They are compared with the algebraic difference  $(M - A)$  to produce *correction scores*:

$$CorrScore_k = |d_k - (M - A)|, \quad k = -2 \dots +2 \quad [4]$$

The correction scores are converted to choice probabilities by an equation analogous to Eq. 3. The only differences are that the correction scores enter with negative signs, thus transforming the softmax rule into softmin, and that a separate temperature parameter is used. In the simulations this parameter was set to 0.040. The discrepancy reference parameter was  $d=0.090$ . To illustrate these settings, suppose the anchor magnitude  $A$  is 0.050 below the target magnitude  $M$ , which is roughly the width of one response category. Then there is 51% chance that the model will increment the anchor response by +1, 39% chance to leave it unchanged, and marginal chance to increment it by +2 or decrement it.

The final response  $R$  is the algebraic sum of the anchor label and the increment, clipped between 1 and 9 if needed.

At the end of the trial the learning mechanism updates the magnitude of the anchor corresponding to the response  $R$ . (Note that this does not necessarily coincide with the anchor retrieved from memory.) The anchor magnitude  $A$  is updated according to Eq. 5, which is a form of competitive learning.

The learning rate  $\alpha$  weighs the most recent trial relative to earlier ones. The simulation experiments used  $\alpha=0.50$ .

$$new\_A = \alpha \cdot M + (1-\alpha) \cdot old\_A \quad [5]$$

The base-level learning equation is somewhat less transparent. The ACT-R theory postulates Equation 6a which contains an explicit term for each instant the anchor is updated (Anderson & Lebière, 1998, p.124). Suppose a particular response has been given at time lags  $t_1, \dots, t_n$  from the present trial. Then the base-level activation  $B$  of the corresponding anchor is the logarithm of a sum of powers [6a], where  $d$  is a decay parameter.

$$B = \ln \left( \sum_i t_i^{-d} \right) \quad [6a]$$

Because Equation 6a is computationally expensive, the model uses Eq. 6b which closely approximates the theoretical formula. The approximation disregards the detailed update history and retains only the time lag since the last usage  $t$ , the lag  $T$  since the beginning of the experiment, and the total number of times the corresponding response has been given up to the current trial. In the simulation experiments the decay parameter was set to  $d=0.5$ , which is a default value used in many ACT-R models. The duration of each trial was 4 sec, as in the psychological experiment.

$$B = \ln \left[ t^{-d} + n \cdot (T^{1-d} - t^{1-d}) / [(1-d)(T-t)] \right] \quad [6b]$$

Equations 2, 3, 4, and 6a are taken verbatim from the ACT-R architecture (Anderson & Lebière, 1998) and thus establish continuity between the ANCHOR model and a broad spectrum of memory-related models. Equation 1 is ANCHOR's connection to Stevens' and Ekman's psychophysical laws.

## Evaluation of the Model

### Simulation Experiment

In order to test the model, its computer implementation was run on the 24 random sequences of stimuli used in the psychological experiment. To mimic the effect of the introductory demonstration, the magnitudes of the anchors were initialized as follows. Anchor 9 was set to 0.800—a compromise value between the longest stimulus presented on the demonstration (675 pixels) and the total width of the screen (1000 pixels). Anchor 1 was initialized to 0.150 and the remaining anchors were evenly spaced in between. The other parameters were set as reported in the previous section. The model generated 24 sequences of responses which were then analyzed in the same way as the psychological data.

Table 1: Comparison of the performance of the model and the psychological data. See text for details.

Statistic	Human data				Model			
	min	mean	max	s.d.	min	mean	max	s.d.
Accuracy ( $R^2$ )	.65	<b>.80</b>	.91	.07	.65	<b>.76</b>	.84	.05
Resp. std.dev.	1.20	<b>1.96</b>	2.44	.28	1.58	<b>1.81</b>	2.57	.21
Multiple $R^2$	.67	<b>.83</b>	.93	.07	.73	<b>.78</b>	.84	.03
Increase in $R^2$	.00	<b>.02</b>	.06	.01	.00	<b>.02</b>	.10	.02
$\beta$ for $S_t$	.80	<b>.90</b>	.93	.04	.80	<b>.87</b>	.92	.03
$\beta$ for $S_{t-1}$	-.53	<b>-.25</b>	-.08	.10	-.47	<b>-.23</b>	-.10	.09
$\beta$ for $R_{t-1}$	+.15	<b>+.30</b>	+.55	.10	+.13	<b>+.25</b>	+.53	.10

Table 1 summarizes the outcome of these various analyses and compares the performance of the model with the human data. The overall accuracy of the model, operationalized as the squared correlation between stimuli and responses, ranges from 0.65 to 0.84 in the sample of 24 runs, with mean 0.76 and standard deviation 0.046. The mean  $R^2$  for the psychological data is 0.80. The degree of non-uniformity of the response distribution is reflected in the standard deviations reported in the second row of Table 1.

The remainder of Table 1 summarizes the multiple regression analysis of the response  $R_t$  on the current stimulus  $S_t$ , previous stimulus  $S_{t-1}$ , and previous response  $R_{t-1}$ . The model shows the same pattern of sequential effects as the psychological data.

Overall, the results of the simulation experiment suggest that the ANCHOR model closely matches human category-rating behavior. The biggest discrepancy between the two data sets is that the model responses are less variable. The human data, however, includes both within-subject and between-subject variability whereas the parameter settings of the model were fixed for all 24 runs. Individual differences can be modeled by using different parameter settings for the different runs.

### Explanation of the Empirical Phenomena

The fact that a model fits the data indicates that its computational mechanisms hang together and can be brought in line with the empirical observations. A much more acid test for the utility of the model, however, is the degree to which it contributes to the theoretical understanding of the psychological phenomena. This closing section discusses the empirical effects in light of the ANCHOR model.

**Nonuniformity of the Response Distribution.** The model shifts the level of theorizing from aggregate scale values to individual responses. At that level of granularity the entire response distribution becomes important. Two salient features of this distribution appear to be the predominance of responses in the middle of the scale and the relative infrequency of extreme responses (Figure 2). Several factors conspire to produce such distributions in the model. The base-level learning mechanism (Eq. 6a/b) tends to differentiate the response frequencies—more frequent anchors build up strength which in turn makes them more likely to be retrieved in the future. This makes flat distributions unstable—small differences tend to grow. This self-reinforcing dynamics cannot go out of hand, however, because of three stabilizing factors. First and foremost, the immediate stimulus controls about 75% of the response variance and hence the responses cannot stray too much from the stimuli. Second, the correction mechanism redistributes the strength among neighboring anchors. This inhibits the formation of isolated spikes or gaps in the distribution, making the smooth unimodal shape the most stable configuration. The third stabilizing factor is related to the context effects discussed below.

**Context Effect.** If the stimuli control 75% of the response variance and the base-level learning tends to amplify inequalities, what happens when the stimuli are unevenly

distributed themselves? It may appear that the model would produce responses that are even more skewed. This would directly contradict the finding of several studies (Parducci, 1965; Parducci & Wedell, 1986; Schifferstein & Frijters, 1992). Empirically, the responses tend to be less skewed than the stimuli, not more so. However, simulation experiments with the ANCHOR model that are too long to be detailed here indicate that it produces context effects consistent with the empirical data. In a nutshell, this is due to the anchor adjustment Equation 5. Because the anchors are prototypes, they tend to cluster in those regions of the magnitude continuum that are densely populated with stimuli. In turn, this reduces the skewness of the response distribution.

**Sequential Effects.** The positive autocorrelation between responses on successive trials is a direct consequence of the recency component of base-level activations (Eq. 6a/b). When a particular response is given, the BLA of its corresponding anchor goes up, which in turn improves the probability of retrieving the same anchor on the next trial. This produces assimilation towards the previous response. However, the increase of the activation level matters only when the two successive stimuli are similar enough (cf. Eq. 2). If they are too far apart, the response on the first trial primes an anchor that is too remote from the target on the second trial to have any influence on the final outcome. The closer the two consecutive stimuli, the stronger the assimilation.

Another sequential effect is the negative correlation between the response  $R_i$  on a given trial and the stimulus  $S_{i-1}$  on the previous trial. Part of this effect is probably due to the perceptual subsystem and its tendency to enhance contrasts. The ANCHOR model, however, has a deliberately simplified front end that precludes any interaction between the stimuli at the perceptual level. Still, the model exhibits contrast effects due to the plasticity of anchor magnitudes (Eq. 5) and the discrepancy penalizing aspect of the partial matching mechanism (Eq. 2). The magnitude of the past stimulus  $S_{i-1}$  is averaged into the magnitude of one of the anchors, which then serves as a proxy of that stimulus on subsequent trials. The anchor magnitudes  $A_i$  are subtracted from the new target magnitude  $M$  during the partial matching process. In other words, one of the  $A_i$  terms in Eq. 2 is positively correlated with  $S_{i-1}$ ,  $M$  is positively correlated with  $R_i$ , and  $A_i$  and  $M$  are subtracted from each other. This creates negative relationship between the response  $R_i$  and the previous stimulus  $S_{i-1}$ .

**Memory-Related Effects.** The anchors are stored in memory and decay only slowly with time. Therefore, the mapping from stimuli to responses implicit in these anchors can influence the performance hours and even days later.

This paper argues in favor of the hypothesis that category ratings are produced in a memory-based manner. A range of category-rating phenomena seem to arise naturally from a set of principles that are also consistent with a large body of memory research. In so far as the ANCHOR model is successful, it illustrates the advantages of its integrative methodology and the utility of general architectures for cognitive modeling.

## Acknowledgments

This research is supported in part by grant AFOSR F49620-99-10086 awarded to John Anderson. The authors thank Stefan Mateeff, Stephen Gotts, and two anonymous reviewers for their valuable comments on the paper. The contribution of Stefan Mateeff is especially gratefully acknowledged.

## References

- Anderson, J. R. & Lebière, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Ekman, G. (1959). Weber's law and related functions. *Journal of Psychology*, 47, 343-352.
- Jesteadt, W., Luce, D., & Green, D. M. (1977). Sequential effects in judgments of loudness. *Journal of Experimental Psychology: Human Perception and Performance*, 3 (1), 92-104.
- Krantz, D. H. A. (1972). A theory of magnitude estimation and cross-modality matching. *Journal of Mathematical Psychology*, 9, 168-199.
- Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, 95 (4), 492-527.
- Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, 72 (6), 407-418.
- Parducci, A. & Wedell, D. H. (1986). The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance*, 12 (4), 496-516.
- Petzold, P. (1981). Distance effects on sequential dependencies in categorical judgments. *Journal of Experimental Psychology: Human Perception and Performance*, 7 (6), 1371-1385.
- Schifferstein, H. N. J. & Frijters, J. E. R. (1992). Contextual and sequential effects on judgments of sweetness intensity. *Perception & Psychophysics*, 52 (3), 243-255.
- Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64 (3), 153-181.
- Ward, L. M. & Lockhead, G. R. (1970). Sequential effects and memory in category judgments. *Journal of Experimental Psychology*, 84 (1), 27-34.

# The Area Activation Model of Saccadic Selectivity in Visual Search

**Marc Pomplun (marc@psych.utoronto.ca)**  
**Eyal M. Reingold (reingold@psych.utoronto.ca)**  
**Jiye Shen (jiye@psych.utoronto.ca)**  
**Diane E. Williams (diane@psych.utoronto.ca)**  
University of Toronto, Department of Psychology  
100 St. George Street, Toronto, Ontario, Canada M5S 3G3

## Abstract

We present an approach towards a simple, explicit model of saccadic selectivity in visual search tasks. The model in its present state includes weights for target-distractor similarities and fixation field size as its only adjustable parameters. Based on these, the model predicts the statistical distribution of saccadic endpoints for any given visual search display. Besides providing an explicit and complete mathematical specification of the model, we demonstrate the performance of its computer simulation in a triple-conjunctive search task. The model successfully simulates empirical data reported by Williams and Reingold (in press).

## Modeling Visual Search

How do we detect a prespecified target item among a set of distractors? Numerous studies employing the paradigm of *visual search* have attempted to answer this question (see Treisman, 1988 and Wolfe, 1998, for reviews). In a typical visual search task, subjects have to decide whether a search display contains a designated target item, indicating their decision by pressing either a “yes” or a “no” button. In most studies, reaction times (RTs) and error rates were analyzed as a function of the number of items in the display (display size). The majority of current models of visual search were based on data obtained within this paradigm.

An early attempt to model visual search is the Feature Integration Theory (Treisman & Gelade, 1980; Treisman, 1988). This theory proposes the existence of preattentive *feature maps*, one for each stimulus dimension such as color or shape. These maps are created in parallel after stimulus onset and allow immediate target detection if the target is defined by a unique feature in any single dimension (feature search). If the target is defined by a specific combination of features (conjunctive search), attention is necessary to locally combine the information of the corresponding feature maps. As a result, subjects have to inspect the display in an item-by-item fashion until target detection or exhaustive search. The Feature Integration Theory thus explains the finding that reaction time tends to increase with display size in conjunctive search tasks, while it is almost constant in feature search tasks.

A more recent approach is the Guided Search Model (Cave & Wolfe, 1990; Wolfe, Cave & Franzel, 1989; Wolfe,

1994), which proposes a two-stage model of visual search. In the first, parallel stage, an *activation map* containing likely target locations is created on the basis of both top-down and bottom-up sources of activation. The second, serial stage uses the activation map to guide visual attention from item to item, starting with the item with the highest activation, then proceeding to the second highest, and so on, until the target is found or the current activation falls below a certain threshold (see Chun & Wolfe, 1996).

Besides many variations of these two models, there are also more complex approaches like the one by Grossberg, Mingolla and Ross (1994). Their model uses artificial neural networks to achieve perceptual grouping of search displays into subregions. Visual search is assumed to proceed serially between these subregions and in parallel within them.

Recently, several researchers have analyzed participants' eye movements during visual search to supplement traditional RT and accuracy measures (e.g. Findlay, 1997; Hooge & Erkelens, 1999; Jacobs, 1987; Luria & Strauss, 1975; Motter & Belky, 1998; Rayner & Fisher, 1987; Scialfa & Joffe, 1998; Shen, Reingold, & Pomplun, in press; Viviani & Swenson, 1982; Williams, Reingold, Moscovitch, & Behrmann, 1997; Williams & Reingold, in press; Zelinsky, 1996; see Rayner, 1998, for a review). Some of these studies have further examined *saccadic selectivity*, i.e. the proportion of saccades directed to each distractor type, by assigning saccadic endpoints to the closest display item. Such studies have found a strong selectivity towards distractors sharing a particular feature with the target item (e.g. Findlay, 1997; Hooge & Erkelens, 1999; Luria & Strauss, 1975; Motter & Belky, 1998; Scialfa & Joffe, 1998; Shen, Reingold & Pomplun, in press; Williams & Reingold, in press; but see Zelinsky, 1996). Given that eye movements are usually accompanied by shifts of attention (see Hoffman, 1998, for a review), it seems that subjects can selectively attend to a critical subset of items in the display rather than perform an item-by-item search as suggested by the original Feature Integration Theory.

To date, no explicit model has been proposed which allows for simulating saccadic selectivity in visual search. In the present article, we propose such an approach, referred to as the Area Activation Model. Following the description of

the model, we examine its performance by simulating the saccadic selectivity findings reported by Williams and Reingold (in press).

### The Area Activation Model

The Area Activation Model is based on assumptions concerning three aspects of visual search performance: (1) the extent of available resources for processing, (2) the choice of fixation positions, and (3) the scan-path structure.

**Processing resources** - The extent of available resources for processing is determined by a two-dimensional Gaussian function with its peak centered at the current gaze position (e.g. Pomplun, Ritter & Velichkovsky, 1996). The standard deviation  $\sigma_f$  of the Gaussian function would be affected by a variety of factors such as task difficulty, item density, and item heterogeneity, but in essence should be a function of the area from which information is extracted during a fixation (henceforth “fixation field”). For example, if the target and distractors are easily discriminable and the density and heterogeneity of items are low, we would expect the fixation field to be larger than when discriminability is low and density and heterogeneity are high. This theoretical measure is likely to be correlated with the number or density of fixations in a given area. If the fixation field is smaller, we would expect more fixations per display area. In fact, in the current simulation we are using the empirically observed number of fixations per trial to adjust  $\sigma_f$ .

**Fixation positions** - Fixation positions are chosen to optimize the amount of information acquired. However, the execution of saccades entails a certain amount of error, which causes fixations to deviate from these optimal positions. Another source of error in empirical data is related to inaccurate measurement of eye movements. It is important for a valid comparison between empirical and simulated data to consider both saccadic error and measurement error.

For every point in the display it is possible to calculate its informativeness or relevance to the search task, creating an activation map. In the present simulation, we use weights corresponding to features along several dimensions to determine activation for individual items. A variety of models may suggest different activation maps (e.g. Cave & Wolfe, 1990; Wolfe, 1994).

In order to make the method transparent and applicable to a wide variety of tasks, we provide a general, explicit specification of the model. A search display consists of  $N$  items with positions  $(x_n, y_n)$  and features  $f_n^{(d)}$  along  $D$  dimensions,  $n \in \{1, \dots, N\}$ ,  $d \in \{1, \dots, D\}$ . The search target has the features  $t^{(d)}$ . Each dimension  $d$  is assigned a weight  $w^{(d)}$ , which currently has to be estimated on the basis of the results from a pilot-study. If, for example, subjects rely entirely on color, the color weight should be set to 1 and all other weights set to 0.

If an item  $n$  is identical to the target in dimension  $d$ , the item's feature activation  $a_n^{(d)}$  is set to the weight of that dimension:

$$a_n^{(d)} = \begin{cases} w^{(d)}, & \text{if } f_n^{(d)} = t^{(d)} \\ 0, & \text{otherwise} \end{cases}, \quad n \in \{1, \dots, N\}, \quad d \in \{1, \dots, D\}$$

The total activation of item  $n$  is then calculated as the sum of its feature activations, implying the possibility of simultaneous guidance of attention by two or more dimensions:

$$a_n = \sum_{d=1}^D a_n^{(d)}, \quad n \in \{1, \dots, N\}$$

In a triple-conjunction search task, for instance, with color, shape, and orientation weighted 1, 0.5, and 0 respectively, a distractor item of the same color and shape as the target would receive a total activation of 1.5, surpassing those distractors with single-feature correspondence. Results from empirical studies support the hypothesis of combined activation across dimensions (see Williams & Reingold, in press).

As argued above, the activation map function  $m(x, y)$  should reflect the amount of information that could be processed during a fixation at any position  $(x, y)$  in the display, given a Gaussian distribution of resources for processing. In the current model,  $m(x, y)$  is calculated as the sum of total activations of all the items, with each item weighted by the amount of resources it receives, as a function of its distance from  $(x, y)$ :

$$m(x, y) = \sum_{n=1}^N a_n \cdot \exp \left[ -\frac{(x-x_n)^2 + (y-y_n)^2}{2\sigma_f^2} \right]$$

The fixation targets are chosen as those maxima (peaks) of  $m(x, y)$  that are greater than or equal to the activation of a single target item, i.e. those coordinates  $(x_p, y_p)$  meeting the following two requirements:

$$\exists \varepsilon > 0: |x - x_p| + |y - y_p| < \varepsilon \Rightarrow m(x_p, y_p) > m(x, y) \forall x, y$$

$$m(x_p, y_p) \geq \sum_{d=1}^D t^{(d)}$$

While the first requirement achieves a plausible selection of fixation points for most efficient acquisition of information, the second requirement simulates a subject's ability to give a negative response even before attending to every item in the display. According to this equation, subjects can decide whether a peak in the activation map is high enough to possibly contain a target item. They can thus stop the search after inspecting all relevant peaks, without directing their attention to the irrelevant ones.

After calculating the fixation targets, the actual fixation points are determined by simulating normally distributed saccadic error and measurement error. Saccadic error is assumed to increase with a larger fixation field, which corresponds to faster search, longer saccades, and a more diffused activation map. Accordingly, in the present

simulation, we set the saccadic error parameter to equal the fixation field parameter  $\sigma_f$ . Measurement error is set to a constant standard deviation  $\sigma_m$  corresponding to the precision of the eye tracker used in the empirical study. The actual fixation point for an activation peak  $(x_p, y_p)$  is thus determined on the basis of the following probability distribution  $p(x, y)$ :

$$p(x, y) = \frac{1}{2\pi(\sigma_f^2 + \sigma_m^2)} \cdot \exp\left[-\frac{(x-x_p)^2 + (y-y_p)^2}{2(\sigma_f^2 + \sigma_m^2)}\right]$$

**Scan paths** - The structure of scan paths is governed by the principle that every fixation target, i.e. every relevant peak in the activation map, is visited exactly once. The order in which these fixation targets are inspected is chosen in terms of spatial optimization, as suggested by empirical results (e.g. Zelinsky, 1996). Among the unvisited peaks, the current implementation of the model always chooses the one that is nearest to the current gaze position. This type of local scan-path minimization - also termed the "Greedy Heuristic" - has been shown to resemble human scanning strategies without assuming extensive planning processes (see Pomplun, Carbone, Koesling, Sichelschmidt & Ritter, submitted).

Turning back to the distinction between feature and conjunctive search, the current model makes the following predictions: If the distractors' activations are too weak to form peaks that exceed the target activation - for example, if the target has a unique feature in one dimension (feature search) - the target item produces the only relevant peak in the display, yielding a highly efficient "pop out" search. In contrast, increasing target-distractor similarity (e.g. conjunctive search) leads to more fixations and a stronger influence of display size on search performance. These predictions of the model are consistent with empirical results.

## Empirical Validation of the Model

The Area Activation model is illustrated by simulating saccadic selectivity findings reported by Williams and Reingold (in press). The authors reported two visual search experiments with 32 participants in each experiment. Participants were presented with displays of 6, 12, and 24 items, half of them containing a target item defined by a unique combination of three dimensions - color, shape, and orientation. Each experiment consisted of a single-feature (SF) and a two-feature (TF) condition, in which the distractor items shared one or two dimensions respectively with the target item. While both experiments used the same colors (red and blue) and orientations (upright and rotated clockwise by 90 degrees), the stimuli differed in the discriminability of the shape dimension. Experiment 1 employed the similar letters E and F (low discriminability), whereas Experiment 2 used the distinct letters T and C (high discriminability). Figure 1 (upper row) presents a sample stimulus for each of the two experiments. Eye movements

were measured with the SR Research Ltd. EyeLink system. The measurement error in this study was determined as  $\sigma_m = 0.6$  deg.

In our comparison of empirical and simulated data, only target-absent trials were analyzed in order to avoid the disruptive influence of target items (see Zelinsky, 1996). In the present article, only the results for display size 24 were simulated.

Since we had no a-priori knowledge about the subjects' fixation field in each of the four conditions (SF and TF conditions in Experiments 1 and 2), we used an iterative algorithm to adjust the model's fixation field parameter  $\sigma_f$  in such a way that the simulated number of fixations per trial matched the empirical one.

Another problem was to determine the weights  $w^{(d)}$  for the color, shape, and orientation dimensions. We used the SF conditions in both experiments to adjust these weights and we tested their generality by applying them to the TF conditions. In the SF condition of Experiment 1, subjects showed strong saccadic selectivity towards color and equally low selectivity towards shape and orientation (see Figure 2, top row). This suggested that only the color dimension induced feature guidance, while shape and orientation were irrelevant to the search process. Consequently, for both the SF and TF conditions in Experiment 1, the weights were set to 1, 0, and 0 for color, shape, and orientation respectively. Experiment 2 differed from Experiment 1 only in the shape discriminability. Therefore, a larger shape weight was required in Experiment 2, but the other two weights had to be the same. We adjusted the shape weight to 0.6 in order to match the empirical saccadic selectivity towards the shape dimension in the SF condition of Experiment 2.

With these adjustments, the computer simulation of the Area Activation Model attempted to address several important questions: Is the model able to quantitatively reproduce the empirical saccadic selectivity? Does the implemented concept of simultaneous guidance by multiple dimensions match the human data, i.e. do the parameters for the SF conditions predict selectivity values in the TF conditions? Do the simulated gaze trajectories correspond to the empirical ones, as indicated by the distribution of saccade amplitudes?

Figure 1 (lower left) shows the activation map calculated by the computer simulation for the sample stimulus of Experiment 1. It reveals four peaks induced by groups of distractors sharing the target color blue, since in this condition only color features contribute to the activation map. As shown in Figure 1 (lower right), the simulation fixates once in the vicinity of each peak while always choosing the nearest unvisited peak as the next saccade target.

Figure 2 allows a comparison between simulated and empirical results, with each row referring to one of the four conditions. The first row shows a remarkable correspondence in the SF condition of Experiment 1, for both the amplitude and the feature selectivity of saccades.



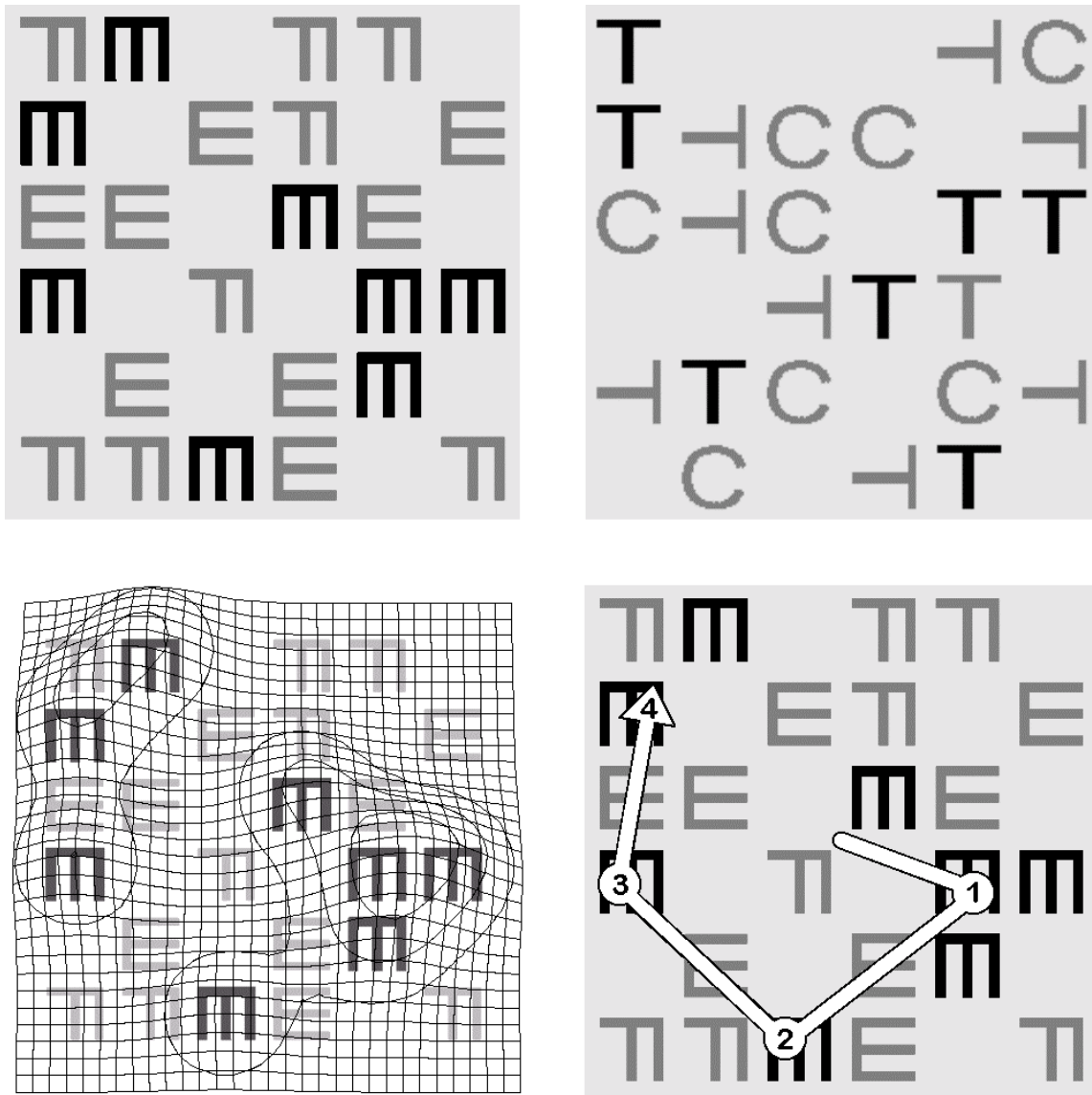


Figure 1: Sample stimuli and illustration of the Area Activation Model. Blue and red items are displayed in black and gray respectively. Upper left: Experiment 1, SF condition, target is a blue, upright “F” (absent). Upper right: Experiment 2, TF condition, target is a red, upright “T” (present). Lower left: Activation map (“activation landscape”) calculated for the sample stimulus of Experiment 1. Lower right: Scan path generated by the model for the same stimulus. The four fixations correspond to the four peaks in the activation map.

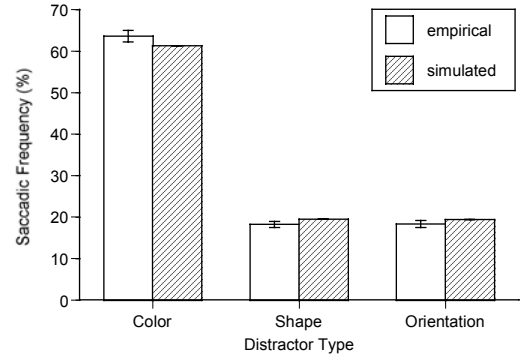
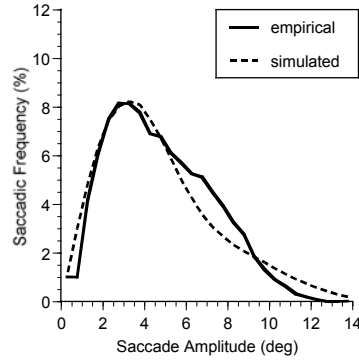
The same is true for the TF condition, as shown in the second row. Despite a profound difference in search efficiency between these two conditions (3.77 versus 10.41 fixations per trial), the distribution of saccades and their selectivity is well predicted with the same set of parameters used in the SF condition.

With regard to the SF condition of Experiment 2, the model's saccadic selectivity once again closely resembles the empirical one, whereas the saccade histogram indicates a significant mismatch. The empirical data revealed a peak at an amplitude of approximately 3 degrees, but the model produced a smoother distribution extending further towards

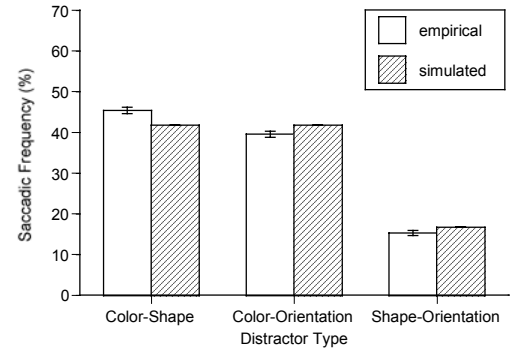
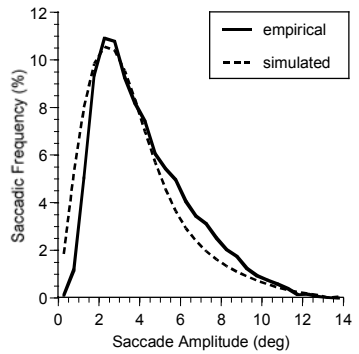
higher amplitudes. This discrepancy might be related to the high search efficiency in this condition (only 2.59 fixations per trial).

Finally, the TF condition, which is substantially less efficient (6.31 fixations per trial), showed an excellent correspondence between simulated and empirical data. The same parameters that failed to replicate the distribution of saccade amplitude in the SF condition almost perfectly reproduced the empirical amplitude histogram in the TF condition. Again, the model precisely predicted the effect of simultaneous guidance by two dimensions.

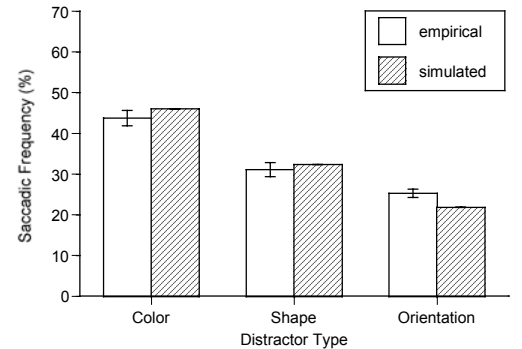
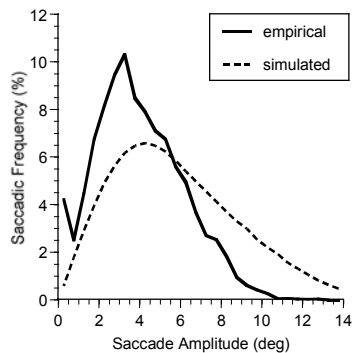
Experiment	1
Condition	SF
Fixations per Trial	3.77
Fixation Field Size	1.05 deg



Experiment	1
Condition	TF
Fixations per Trial	10.41
Fixation Field Size	0.96 deg



Experiment	2
Condition	SF
Fixations per Trial	2.59
Fixation Field Size	1.82 deg



Experiment	2
Condition	TF
Fixations per Trial	6.31
Fixation Field Size	1.06 deg

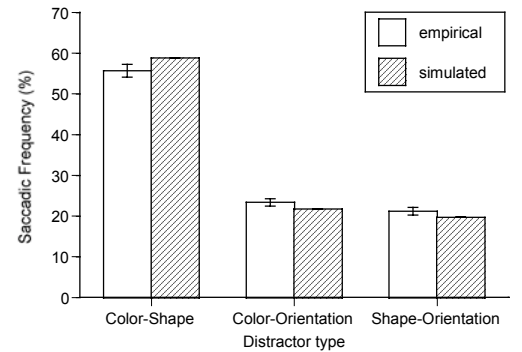
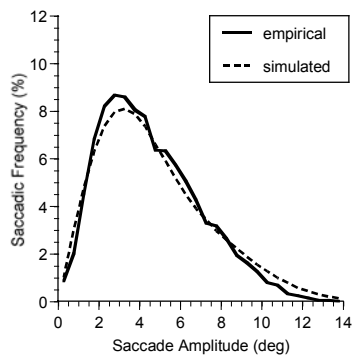


Figure 2: Comparison between empirical and simulated data with each row corresponding to one of the four experimental conditions. Left column: Empirical number of fixations per trial and simulated visual span size required to match the number of fixations. Middle column: Comparative histograms of saccade amplitude. Right column: Comparative diagrams of saccadic selectivity towards different distractor types.

## Conclusions

In all four conditions, empirical saccadic selectivity was precisely replicated, supporting the concept of simultaneous guidance by multiple dimensions. Moreover, saccade amplitude produced by the model was remarkably accurate. One exception found was the SF condition in Experiment 2. This is perhaps due to the fact that search in this condition was highly efficient. It may be the case that highly efficient searches induce a qualitatively different saccadic scanning behavior. For example, if it is always possible to detect the target from the central gaze position, an efficient strategy could be to avoid any eye movements to the periphery. Another factor could be an increased amount of corrective saccades due to faster scanning of the display. Further research is necessary to investigate this issue.

As indicated by the model's accurate saccadic selectivity, not only the area-based activation map, but also the implementation of saccadic error - as identical to the fixation field size  $\sigma_f$  - have passed their first test. The generally successful replication of saccade amplitude supports the hypothesis of spatial scan-path optimization within the relevant display areas.

All in all, the current version of the Area Activation Model can be considered a promising approach towards an explicit, quantitative model of saccadic selectivity in visual search.

## Acknowledgments

The preparation of this paper was supported by a grant to Marc Pomplun from the Deutsche Forschungsgemeinschaft (DFG) and a grant to Eyal Reingold from the Natural Science and Engineering Research Council of Canada (NSERC).

## References

- Cave, K.R., & Wolfe, J.M. (1990). Modeling the role of parallel processing in visual search. *Cognitive Psychology*, 22, 225-271.
- Chun, M.M., & Wolfe, J.M. (1996). Just say no: How are visual searches terminated when there is no target present? *Cognitive Psychology*, 30, 39-78.
- Findlay, J.M. (1997). Saccade target selection during visual search. *Vision Research*, 37, 617-631.
- Grossberg, S., Mingolla, E., & Ross, W.D. (1994). A neural theory of attentive visual search: Interactions of boundary, surface, spatial, and object representations. *Psychological Review*, 101, 470-489.
- Hoffman, J.E. (1998). Visual attention and eye movements. In H. Pashler (Ed.), *Attention*. England UK: Hove.
- Hooge, I. T., & Erkelens, C. J. (1999). Peripheral vision and oculomotor control during visual search. *Vision Research*, 39, 1567-1575.
- Jacobs, A.M. (1987). Toward a model of eye movement control in visual search. In J.K. O'Regan & A. Levy-Schoen (Eds.), *Eye movements: From physiology to cognition*. North-Holland: Elsevier Science Publishers.
- Luria, S. M., & Strauss, M. S. (1975). Eye movements during search for coded and uncoded targets. *Perception and Psychophysics*, 17, 303-308.
- Motter, B. C., & Belky, E. J. (1998b). The guidance of eye movements during active visual search. *Vision Research*, 38, 1805-1815.
- Pomplun, M., Velichkovsky, B.M., & Ritter, H. (1996). Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception*, 25 (8), 931-948.
- Pomplun, M., Carbone, E., Koesling, H., Sichelschmidt, L., & Ritter, H. (submitted). Modeling visual scanning strategies in two-dimensional object distributions. *Cognitive Science*.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372-422.
- Rayner, K. & Fisher, D.L. (1987). Eye movements and the perceptual span during visual search. In J.K. O'Regan & A. Levy-Schoen (Eds.), *Eye movements: From Physiology to Cognition*. North Holland: Elsevier.
- Scialfa, C. T., & Joffe, K. (1998). Response times and eye movements in feature and conjunction search as a function of eccentricity. *Perception & Psychophysics*, 60, 1067-1082.
- Shen, J., Reingold, E. M., & Pomplun, M. (in press). Distractor ratio influences patterns of eye movements during visual search. *Perception*.
- Treisman, A. (1988). Features and objects: The fourteenth Bartlett Memorial Lecture. *The Quarterly Journal of Experimental Psychology*, 40A, 201-237.
- Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, 12, 97-136.
- Viviani, P., & Swenson, R.G. (1982). Saccadic eye movements to peripherally discriminated visual targets. *Journal of Experimental Psychology: Human Perception and Performance*, 8 (1), 113-126.
- Williams, D.E., Reingold, E.M. (in press). Attentive guidance of eye movements during triple conjunction search tasks. *Psychonomic Bulletin and Review*.
- Williams, D.E., Reingold, E.M., Moscovitch, M., & Behrmann, M. (1997). Patterns of eye movements during parallel and serial visual search tasks. *Canadian Journal of Experimental Psychology*, 51, 151-164.
- Wolfe, J.M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1, 202-238.
- Wolfe, J.M. (1998). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13-71). Hove, England UK.
- Wolfe, J.M., Cave, K.R., & Franzel, S.L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 419-433.
- Zelinsky, G.J. (1996). Using eye saccades to assess the selectivity of search movements. *Vision Research*, 36, 2177-2187.

# The use of a high-dimensional, “environmental” context space to model retrieval in analogy and similarity-based transfer

Michael Ramscar and Daniel Yarlett

{michael,dany}@cogsci.ed.ac.uk

Division of Informatics

University of Edinburgh

Edinburgh, Scotland EH8 9LW

## Abstract

Current models of the retrieval of analogies from a long-term memory store assume mental representations that are generally either underspecified or implausible. In this paper we conduct two experiments which demonstrate that an ‘environmental’ approach to retrieval can produce appropriate retrieval patterns on cognitively plausible styles of representation, utilising information that can be easily learned from a linguistic environment.

## Introduction: Similarity-Based Transfer

Analogy (and similarity-based transfer) is a central cognitive process that represents a versatile problem-solving and reasoning strategy, allowing agents to bring previous experience to bear on novel problems. Its operation embodies two distinct processes: (i) reminding, or *retrieval*, of appropriate analogs from a long-term memory store; after which (ii) candidate analogs are *mapped* onto the representation of the current problem (the target) to determine deeper relational matches, and to allow inferences to be made (Gentner, Ratterman & Forbus, 1993; Forbus, Gentner & Law, 1995; Holyoak & Thagard, 1995).

The latter mapping process has been shown to rely largely on structural commonalities (Gentner, 1983; Holyoak & Thagard, 1995; Hummel & Holyoak, 1997), and computational models of the mapping processes that determine structural commonalities have been subject to much critical scrutiny (Falkenhainer, Forbus & Gentner, 1989; Holyoak & Thagard, 1989; Keane, Ledgeway & Duff, 1994, Hummel & Holyoak, 1997). In contrast *retrieval* has been subject to less investigation. Here, we subject the relatively more neglected issue of modelling analog retrieval to a more focussed theoretical examination.

## Four Constraints on Retrieval

Empirical studies by Gentner, Ratterman and Forbus (1993) established four primary constraints on the patterns that an appropriate theory of retrieval should produce given a specific context or probe:

1. *Primacy of the mundane*: The majority of retrievals evoked should be literally similar to the context, sharing both surface and structural characteristics (e.g. a

bicycle should call to mind memories of other bicycles).

2. *Surface superiority*: Retrievals based on surface similarity alone (without structural similarity) should also be frequent (e.g. a fairy story about a frog might call to mind other stories about frogs, although the structure of the stories might differ greatly).

3. *Rare insights*: Memories that are structurally similar to the target context should be retrieved only occasionally (e.g. the orbits of the solar system reminding one of electrons orbiting an atom).

4. *Scalability*: The model must plausibly extend to realistically sized memory pools because people typically have vast numbers of memories, and are able to access them in a matter of seconds.

Gentner, Ratterman and Forbus’ (1993) investigation demonstrated that retrieval is sensitive to surface (or ‘semantic’, Hummel and Holyoak, 1997) similarities between a target representation and a base analogy that needs to be retrieved. (As opposed to the shared relational structure that determines an analogical match.) The retrieval process, being relatively computationally cheap, acts as an efficient prefilter to the more expensive process of structural alignment (albeit at the expense of potentially passing over useful analogies that share structural commonalities with the target domain).

## Meeting the Constraints

MAC/FAC (Forbus, Gentner & Law, 1994) and LISA (Hummel & Holyoak, 1997) are the two foremost models of similarity-based transfer. Below we review the approach taken by both models with regards to retrieval, and examine the theoretical basis for each.

## MAC/FAC: Content Vectors

MAC/FAC models retrieval by generating a *content vector* for each representation that is stored in its memory-pool. A content vector summarises the surface features of a representation by recording the frequency with which each *lexically distinct* predicate occurs in it. Thus, the following proposition:

```
(CAUSE (STRIKES-WITH JOHN CUE CUE-BALL)
      (AND (POTS CUE-BALL) (POTS BLACK)))
```

would be assigned the following content vector:

```
((CAUSE . 1) (STRIKES-WITH . 1)
 (AND . 1) (POTS . 2))
```

A measure of the degree that two representations share the same surface features can then be derived by calculating the dot-product of their content vectors (if a particular predicate does not appear in a representation then it is implicit, adopting a sparse-encoding approach, that it has a frequency of zero). It is important to note that only predicates that are *identical* from one another can contribute to the magnitude of a dot-product between two content vectors: there is no potential for multiplying the frequencies of distinct predicates in the dot-product calculation.

Forbus, Gentner and Law (1994) argue that the dot-product between two content vectors provides an empirically adequate measure of the *retrievability* of one representation, given another as a context, because it satisfies the four constraints on retrieval performance.

### A Critique of Content Vectors

In order to model the way that *lexically distinct* items in stimuli prime one another for retrieval, the content vector theory makes a commitment to a theory of mental representation we shall call canonical representation (CR) theory. This presupposes a translation procedure that allows tokens that are lexically distinct but share similar semantic “meanings” to be re-encoded using identical tokens. This translation procedure accounts for cross-lexeme priming effects by identically encoding distinct lexemes that should prime for one another, thus ensuring that they can contribute to the dot-product score between the two representations in which they feature. CR theory assumes that during the process of comprehension (representation building):

“Two concepts that are similar but not identical (such as ‘bestow’ and ‘bequeath’) are decomposed into a canonical representation language so that their similarity is expressed as a partial identity (here, roughly, give).” Forbus, Gentner and Law, 1994, pp. 153

### ‘Canonical Form’?

According to CR theory, complex semantic elements can be recursively decomposed -- or *re-represented* -- until a canonical measure of their semantic significance is reached. Hence CR theory assumes that the mental encoding of semantically complex concepts can ultimately be analysed in terms of a stock of canonical forms. Clearly the correctness or otherwise of this assumption is an empirical matter. However, it does seem worth noting that research into the mental representation of concepts suggests that human conceptual representations are anything but canonical. The proposals for generalised theories of representation that exist in the concepts literature fall well short of providing the kind of “neat” account of concepts that canonical conceptual representation assumes (see Komatsu, 1992; Ramsar & Hahn, 1998 for reviews). Lacking as it does an account of what a canonical conceptual form is, in its current form CR theory is under-specified, and thus fails to operationalise the notion of semantic similarity in a sufficiently tight manner. This prevents specific predictions being made from the theory (e.g. how strongly do ‘cat’ and ‘dog’ prime for one another based on an analysis of the overlap in their shared semantic features?).

### LISA: Semantic Features

The other leading model of analogy in the literature, LISA (Hummel & Holyoak, 1997) also relies upon the notion of semantic units (or links) – and re-representations into ‘semantic primitives’ – in its structured representations to model retrieval. These semantic elements are largely constrained by the representation strategy adopted in LISA (e.g. `\verb+likes1+` or `\verb+likes2+`). Hummel and Holyoak’s claim is that these allow appropriate patterns of retrieval to be produced by their model. However, they offer no empirical support for the selection of their particular set of primitive semantic features. At present, the semantic information in LISA’s representations is *hand-coded*, and ultimately reliant upon humanistic intuitions about similarities of meaning.

### Summary of Current Approaches

Both MAC/FAC and LISA present models of retrieval that are theoretically under-specified. Both accounts rely on the problematic (i.e. currently undefined) notion of re-representation, either into ‘canonical conceptual representations’ (MAC/FAC) or ‘semantic primitives’ (LISA). Ultimately, this means that both models rely on hand-coded information to drive their retrievals. Neither LISA nor MAC/FAC actually *models* the representation of lexical information. They rely instead on imported information (primarily intuition) to underpin their behaviour, thus neither can be said – at present – to offer any real explanation of the role of lexico-conceptual knowledge in retrieval.

None of this means, of course, that the shortcomings that we describe in each of the two theories could not ultimately be addressed. We do, however, feel that in the light of these shortcomings there is room for an investigation of whether another approach to the representation of lexico-conceptual knowledge might be used to ground an alternative theory of retrieval.

### Co-occurrence Models of Semantics

One approach to lexico-conceptual knowledge that seems promising in this respect is the high dimensional modelling of context spaces. This is a data-intensive technique that analyses a set of corpora, and from this derives a summary of the variety of different contexts that different words can be *used* in. There is a growing body of evidence that the frequency with which different lexemes co-occur with one another (that is, are used together within a particular context, such as a paragraph or moving-window) can provide useful information about the semantic properties of those lexemes.

In co-occurrence analyses, a contextual distribution is calculated for each lexeme encountered in a corpus analysis by counting the frequency with which it *co-occurs* with every other lexeme in the corpora being analysed. The contextual distribution of a lexeme can then be summarised by a vector showing the frequency with which it is associated with the other lexemes in a common linguistic environment. One can think of this information as defining a model containing a network of links between the lexemes in a language, each with varying strengths,

and representing the varying contextual co-occurrences of lexemes in that language. Two such co-occurrence models are the *Latent Semantic Analysis* (LSA) model (Landauer and Dumais, 1997; Landauer, Foltz & Laham, 1998), and the *Hyperspace Analog to Language* (HAL) model (Burgess & Lund, 1997).

There is good evidence that co-occurrence analysis extracts information from corpora that can be used to model certain linguistic behaviour. For example, Landauer and Dumais (1997) report that the LSA model can pass a multiple-choice TOEFL synonym test. Lund, Burgess and Atchley (1995) present evidence that co-occurrence data can act as a good predictor of various priming effects. Burgess and Lund (1997) demonstrate that the HAL model can produce clustering in its high-dimensional space of lexemes from differing grammatical categories.

Whilst the exact parameters of LSA and HAL are different, they both adopt the general approach outlined above to generate co-occurrence vectors. We feel that there are a number of attractive benefits to be gained from modelling the semantic information used in analogical and similarity based retrieval in this way:

1. The proposed semantic metric is clearly specified. By proposing that the semantic information used in retrieval is learned from observing the varying contextual co-occurrences of lexemes in a language, we avoid having to postulate entities – such as ‘semantic primitives’ whose theoretical and psychological nature is massively under-specified.
2. The semantic information used could be easily learned from the environment,<sup>1</sup> thus avoiding the problems inherent in positing entities whose learnability is somewhat controversial, and whose innateness might otherwise have to be treated as axiomatic (as canonical concepts seem to be; see Laurence & Margiolis, 1999; Fodor, 1981).
3. An environmental context model contains representationally cheap, summarised information, the usage of which makes only limited processing demands. Thus it allows one to avoid the theoretical problems inherent in theories of re-representation which explain cheap surface matches in terms of semantic decomposition and expensive structural alignment (c.f. Holyoak & Hummel, 1997; Forbus *et al*, 1997).
4. Environmental context models are relatively objective: they do not require that a particular set of ‘semantic features’ are defined before textual analysis begins. Instead the co-occurrence technique takes the lexemes themselves as features, and uses frequency relations between them to define their associativity. This is an advantage given the difficulty we have already highlighted of empirically grounding claims as to the identity of semantic features. Furthermore, the use of

<sup>1</sup> Indeed, despite some of the stronger claims made for co-occurrence models of language (c.f. Landauer & Dumais, 1997) we feel that they are best characterised as being essentially models of the associativity of lexemes in a common linguistic environment, such that we prefer to call them “environmental context models”. It is also worth noting that co-occurrence techniques are also compatible with a neural implementation. Lowe (1997) demonstrates that a co-occurrence model can easily be implemented as a self-organising Kohonen map, and this offers some support for the idea that some form of co-occurrence counting could occur in the brain.

dimensional reduction techniques on the vectors associated with each lexeme (Landauer & Dumais, 1997) offers evidence that, in fact, there may not be a *unique* set of semantic features used in the encoding of semantic relations, but rather that multiple encodings can provide sufficient information to meet empirical constraints

5. Because co-occurrence techniques do not rely on a predefined set of semantic features (such as gender, plurality, animacy and so on), this eliminates subjectivity from the decisions that are made during the process of hand-coding representations during the modelling process.

The success of co-occurrence techniques in accounting for priming effects (c.f. Lund, Burgess and Atchley, 1995), has shown them to be useful models of lexical retrieval. Here, we seek to establish whether these models can be used to account for the retrieval of structured composite representations, and not just individual lexemes, from a memory-pool.

### The ‘Karla the Hawk’ Stories

The experiments detailed below use the ‘Karla the Hawk’ materials as originally used by Gentner, Ratterman and Forbus (1993). The Karla materials consist of twenty sets of stories written in natural language. Each set consists of a base story, and four systematic variations of that story. Two factors are crossed over the four variant stories, as shown in Table 1.

	+ST	-ST
+SF	Literal Similarity	Surface similarity
-SF	Analogy	1st Order Relations

Table 1: The Karla materials

The four story categories systematically vary the commonalities that are shared with the base-story from which they are derived. Each variant can either share or not share surface ( $\pm$ SF) and structural ( $\pm$ ST) commonalities with the corresponding base-story. This 2 x 2 materials design allows for the controlled examination of the sensitivity of various putative measures of retrieval. Gentner, Ratterman and Forbus (1993) found that the prime determinant of retrievability was shared surface commonalities, whilst shared structural commonalities had a nonsignificant effect. This is the pattern of results that we will look for in our experiments. The empirical results reported in Gentner, Ratterman and Forbus (1993) are summarised in Table 2.

	LS	SS	AN	FOR
Retrieval Scores	1.92	1.64	0.44	0.27
Inferential Soundness	4.41	2.70	4.16	2.58

Table 2: The results of the experiments conducted by Gentner, Ratterman and Forbus (1993).

Below, we report two experiments that compare the performance of the content vector (CV) theory of retrieval, as implemented in MAC/FAC, against the measure provided by the LSA model.

## Experiment 1: Stripped Natural Language.

Experiment 1 was designed to determine whether there is sufficient informational content in a reduced representation of the Karla the Hawk stories to produce retrieval patterns conformable to the empirical data.

It is clear from experimental studies that in addition to the accretion of structural information during comprehension, there is a concomitant loss of superficial verbatim information as propositional representations are built up (Sachs, 1967; Gernsbacher, 1985). Since we wanted to simulate retrieval of what subjects in Gentner *et al*'s studies actually stored (and there is good evidence that people do not store texts verbatim), we decided to initially test retrieval on versions of Gentner *et al*'s stimuli that had all of the *closed-class*<sup>2</sup> lexemes removed from them.

Applying this principle resulted in a set of words for each story which constituted the words which are, in some sense, maximally informative about the context that the representation defines. For example, some words (generally the closed-class words) may occur in almost any (and every) possible context (e.g. 'the' can co-occur plausibly with an extremely diverse set of lexemes). Thus encountering such a word in a probe representation has little informational utility with respect to retrieval because it fails to narrow the set of candidate retrievals at all. Such lexemes are unlikely to influence the kind of retrieval studied by Gentner, Ratterman and Forbus (1993).

The original Karla the Hawk base-story after it had been pruned of all closed-class lexemes is given below, as an example of the characteristic 'bag of words' that remained once the natural language representations had been stripped:

Karla old hawk lived top tall oak tree  
afternoon saw hunter ground bow crude arrows  
feathers hunter aim shot hawk missed Karla  
knew hunter wanted feathers glided down hunter  
offered give hunter grateful pledged shoot  
hawk shot deer

### Method

The base story for each story-set of the reduced representations was compared with each of its four variants in turn, using the LSA and CV (MAC/FAC content vector) models. This was done in order to reproduce the experimental format embodied in Gentner's original retrieval experiments. The LSA model was set to compare items in *document-to-document* mode, using the 300 most significant factors extracted by the model from a corpus that approximates the general reading a first year college student is exposed to (which seemed appropriate given the participants in Gentner *et al*'s studies). Because of the 2 x 2 design of the experiment, a repeated-measure ANOVA analysis is the appropriate test to determine which of the factors,  $\pm$ SF or  $\pm$ ST, the two metrics are sensitive to.

### Results

The results of the inter-story comparisons conducted with the LSA and CV models of retrieval are recorded in Table

<sup>2</sup> Closed-class words belong to the set of words which are closed under the grammatical rules of a language.

3. As noted above, each variant story either exhibits  $\pm$ SF and  $\pm$ ST, depending on whether it shares or does not share object-attributes and higher-order relations (structure) with the base story it is derived from. The ANOVA analysis revealed that the CV metric was sensitive to both  $\pm$ SF ( $F(1,19) = 11.965$ ,  $p < 0.01$ ) and  $\pm$ ST ( $F(1,19) = 10.027$ ,  $p < 0.01$ ), with no significant interaction effect ( $F(1,19) = 3.717$ ,  $p > 0.05$ ). For the LSA metric there was a main effect of  $\pm$ SF ( $F(1,19) = 68.985$ ,  $p < 0.01$ ); no effect of  $\pm$ ST ( $F(1,19) = 2.611$ ,  $p > 0.05$ ), and no significant interaction between the factors ( $F(1,19) = 2.428$ ,  $p > 0.05$ ).

	LS	SS	AN	FOR
CV Metric	0.116	0.084	0.057	0.053
LSA Metric	0.442	0.412	0.151	0.152

**Table 3:** Experiment 1 -- The category means for the CV and LSA scores derived from comparing each base-story with its four variants on the stripped ('bag-of-words') representations. All twenty story-sets had closed-class lexemes removed from them, and were used in the comparison.

### Discussion

The clustering in the mean LSA scores for each category of variant (LS-SS and AN-FOR) mirrors the subject data in Gentner, Ratterman and Forbus's (1993) study closely. The same pattern is not observable in the CV metric. Furthermore, the only significant factor in Gentner's original retrieval experiments was  $\pm$ SF and only the LSA scores conform to this pattern. The CV metric was also sensitive to the  $\pm$ ST factor, which indicates that it is sensitive to a factor which has been shown to have little significant impact on retrieval performance. It appears that there is sufficient information remaining in the reduced representation to allow different contexts for retrieval to be discriminated from one another in a way that simulates the empirical findings discussed.<sup>3</sup> Moreover, it seems clear from these results that LSA models the original empirical data more accurately than CV.

## Experiment 2: Faithful Dgroups

Experiment 2 investigated the performance of the CV and LSA measures on a style of representation that explicitly encodes the structural features implicit in the original stories. This structural information is required to be able to complete the mapping phase of similarity-based transfer, and so these experiments were conducted to determine whether a single style of representation would be sufficient to underpin both the retrieval and mapping processes of similarity-based transfer. The style of representation that we chose shares the substantial core of its form with that used in SME and MAC/FAC, but we developed a series of constraints for translating text into these structured representations whilst avoiding any commitment to the CR theory (we call these representations Faithful Dgroups, 'Dgroup' being the usual term used to describe individual – "chunked" – structured representations in the SME literature.).

<sup>3</sup> It should be noted here that the LSA retrieval scores remain more or less unchanged from pilot testing on the full NL versions. The CV scores, however, are significantly reduced from the original NL materials. This seems to indicate that the LSA model is more robust across representations.

### Producing The Faithful Dgroups

Humans are capable of extracting more meaning from language than the basic information that is encoded in the surface structure of texts and dialogues might suggest. To take the following as an example:

John hit Mary; Mary cried. The Headmaster expelled John.

In interpreting this passage, a reader has to infer firstly that John's hitting Mary caused her to cry, and secondly that the relationship between John's hitting Mary, and her crying, caused the Headmaster to expel John. We might express this information in terms of the following nested propositional structure:

```
cause( cause( hit(john,mary), cry(mary) ),
        expel(headmaster, john) )
```

None of this causal information appears explicitly in the original utterance, so it is clear that it must in some way be inferred from a prior source. (The need for inference here is uncontroversial: all theories of comprehension agree that language comprehension requires a great deal of active involvement on the part of the comprehender when it comes to inferring information that is not explicitly encoded in language (e.g. McKoon & Ratcliff, 1992); where they disagree is on what, and how much, inference actually happens.)

Whilst we haven't attempted to make a commitment to a particular theory of comprehension in specifying the procedure for translating texts into Faithful Dgroups, what we have tried to do is to provide the beginnings of a method that requires a minimal amount of inference, and is broadly compatible with the bulk of the available data in this area (again, see McKoon & Ratcliff, 1992).

The basic outline of a procedure for forming the Faithful Dgroups from natural language samples is described below.

#### Algorithm for Construction of Faithful Dgroups

Seeking to maximally preserve closed-class lexical information:

1. Identify the objects that are referred to in the text, and list them using (`sme:defEntity ...`) commands.
2. Identify all the lexeme structures used to express attributes of the objects in the text, and express these as unary expressions.
3. Identify the lexeme structures used to express relations between the identified objects, and express these in the Dgroup form as expressions with two or more arguments, taking only objects as arguments.
4. Now deal with higher-order information (i.e. temporal and causal information that is frequently *implicit* in NL representations). Express this information as expressions taking other expressions as arguments. Note that because this information is often implicit in the NL forms of the stories, a standard (or canonical) lexical identity for each expression must be adopted (this has the effect of minimising the influence of inferred structures on retrieval, which is in accordance with Gentner's empirical findings). The set of inferred

relations should be the minimum set required to articulate the narrative structure of the story.<sup>4</sup>

Thus we sought to minimise unwarranted inferences, and the addition of features not warranted by their inclusion in the original materials. In contrast to the original Dgroups, the Faithful Dgroups incorporate much of the lexical information that is present in the original natural language representations.

#### Method

Faithful Dgroups representing nine of the original story-sets were created.<sup>5</sup> The faithful Dgroup representing the base story for each story-set was then compared with each of its four variants in turn, again using both the CV and LSA models. The LSA model was again set to compare items in *document-to-document* mode, using the 300 most significant factors extracted by the model from the "first year college student, general reading" corpus.

#### Results

The result of the CV and LSA comparisons on the Faithful Dgroups are presented in Table 4 below.

For the CV method there was no significant effect of  $\pm$ SF ( $F(1,8) = 3.647$ ,  $p > 0.05$ ), no significant effect of  $\pm$ ST ( $F(1,8) = 3.383$ ,  $p > 0.05$ ), and no interaction effect ( $F(1,8) < 1$ ). For the LSA method there was an effect of  $\pm$ SF ( $F(1,8) = 66.091$ ,  $p < 0.01$ ); no significant effect of  $\pm$ ST ( $F(1,8) = 2.190$ ,  $p > 0.05$ ); and no significant interaction between the factors ( $F(1,8) = 1.094$ ,  $p > 0.05$ ).

	LS	SS	AN	FOR
CV Metric	0.751	0.718	0.735	0.688
LSA Metric	0.670	0.633	0.466	0.456

**Table 4:** Experiment 2 -- The category means for the CV and LSA scores derived from comparing each base-story with its four variants in the Faithful Dgroups. Nine of the NL story-sets were encoded in this format

#### Discussion

As expected, on representations make no commitment to CR theory – using instead the lexico-semantic information derived from the external representations to drive retrievals – these results demonstrate that the CV method is insensitive to the surface-features of the stories, and thus fails to produce empirically adequate retrieval patterns. This is because the CV method only permits priming between lexically identical items. The LSA method, however, performs much better: its retrievals are only sensitive to the  $\pm$ SF factor, which is what is required to model the empirical evidence.

It is particularly noteworthy that the LSA method assigned high retrieval scores to the LS and SS categories in this experiment, when their representations need not share any *identical* lexemes with their corresponding base representation. It follows that the LSA model is not simply relying on identical lexemes in distinct

<sup>4</sup> Thus, as with other models of similarity-based transfer, some hand coding of representations does occur (though the freedom to make unprincipled coding decisions is greatly reduced in comparison with other models). This procedure was designed to minimise the influence of such hand coding, although our ultimate goal is the automation of this process.

<sup>5</sup> For comparison purposes, we encoded the same set of stories that Forbus, Gentner and Law (1994) coded for MAC/FAC.



representations to facilitate retrievals, but is modelling instead a more complex kind of relationship between the ways that individual lexemes are used in differing linguistic contexts.

## Conclusion

The performance of the LSA measure on both styles of representation offers concrete evidence that it can act as a good predictor of retrieval. That it can do so even when operating on a style of representation that remains faithful to the natural language source of information, and relies on only a psychologically plausible range of inferences for its structure (i.e. a structured, propositional representation that handles lexeme-encoding realistically) is encouraging. As is the fact that we were able to model the empirical data *without* hand tailoring a model of semantics, instead using an objectively, and independently, derived model of lexico-semantic information.

We alluded above to a potential problem in employing the idea of re-representation in retrieval: that studies have shown retrieval to act as a cheap pre-filter for the more computationally expensive – and conceptually rich – process of analogical mapping. Yet the use of re-representation in this process will result in multiple structural mappings being carried out at the conceptual decomposition stage (as many as there are lexically distinct but "semantically" similar items in representations to be mapped). It doesn't take much reflection to realise that will lead to a situation where more structural mapping is required in reconciling semantic differences than in mapping an analogy itself.

At some point mappings between richly represented structure will have to stop, if only because cognitive processing capacity is limited. Our contention is that re-representation – in retrieval at least – is expensive and unnecessary. Structure mappings can be retrieved – and conceptualised – using a far cheaper source of information. Not only does the use of high-dimensional, "environmental" context space to model retrieval in analogy and similarity-based transfer appear to be a plausible approach, it also seems to satisfy Gentner, Ratterman and Forbus' *scalability* constraint better than other models as well.

Given the role structure appears to play in concepts, any *conceptual* solution to matching semantics may suffer from to re-representation problem as well. It may be that *all* conceptualisation – analogical and literal – is about retrieving and mapping the right information in context. Gentner, Ratterman and Forbus (1993) showed that an inexpensive source of information was all that was needed to contextualise retrieval: our results indicate that a of high-dimensional, "environmental" model can provide that context in analogy and similarity-based transfer. Our suspicion is that it might also serve to contextualise broader conceptual processing as well.

## Acknowledgements

We are grateful to Andrew Wishart for his help in coding the Faithful Dgroups, and also for insightful comments on an initial draft of this paper. We would like to thank Lera Boroditsky, Ken Forbus and Dedre Gentner for

helpful discussion of these issues. This work was supported in part by EPSRC Grant GR/M59846

## References

- Burgess, C., and Lund, K. (1997). Modelling Parsing Constraints with High-Dimensional Context Space. *Language and Cognitive Processes*, 12, 177-210.
- Falkenhainer, B., Forbus, K.D., and Gentner, D. (1989). The Structure-Mapping Engine: Algorithm and Examples. *Artificial Intelligence*, 41, 1-63.
- Fodor, JA (1981) The present status of the innateness controversy. In J Fodor (ed) *RePresentations*, MIT Press
- Forbus, K., Gentner, D., and Law, K. (1994). MAC/FAC: A Model of Similarity-based Retrieval. *Cognitive Science*, 19, 141-205.
- Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7, 155-70.
- Gentner, D., Ratterman, M., and Forbus, K. (1993). The Roles of Similarity in Transfer: Separating Retrievability from Inferential Soundness. *Cognitive Psychology*, 25, 524-575.
- Gernsbacher, M. A. (1985). Surface Information Loss in Comprehension. *Cognitive Psychology*, 17:324-363.
- Holyoak, KJ & Thagard, P (1995) *Mental Leaps*. MIT Press, Cambridge, Ma.
- Hummel, J.E., and Holyoak, K.J. (1997). Distributed Representations of Structure: A Theory of Analogical Access and Mapping. *Psychological Review*, 104, 427-66.
- Keane, M., Ledgeway, T., and Duff, S. (1994). Constraints on Analogical Mapping: A Comparison of Three Models. *Cognitive Science*, 18, 387-438.
- Komatsu, L K (1992) Recent views of conceptual structure. *Psychological Bulletin*, 112(3), 500-526
- Landauer, T.K., and Dumais, S.T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104, 211-40.
- Landauer, T.K., Foltz, P.W., and Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-84.
- Laurence, S. & Margiolis, E. (1999) Concepts and cognitive science. In Laurence & Margiolis (eds.) *Concepts*. MIT Press, Cambridge, Ma.
- Lowe, W. (1997). Semantic Representation and Priming in a Self-Organizing Lexicon. *Proceedings of the 4th Neural Computation and Psychology Workshop* pp. 227-39. Springer-Verlag.
- Lund, K., Burgess, C., and Atchley, R.A. (1995). Semantic and Associative Priming in High-Dimensional Semantic Space. In *Proc. 17th Annual Conference of the Cognitive Science Society* LEA pp. 660-65. LEA.
- McKoon, G., and Ratcliff, R. (1992). Minimal Inference: A Framework for Discourse Processes. *Psychological Review*, 99:440-66.
- Ramscar, M.J.A. and Hahn, U. (1998) What family resemblances are not In *Proc., 20th Annual Conference of the Cognitive Science Society*, LEA, pp 865-870
- Sachs, R.S. (1967) Recognition Memory for Syntactic and Semantic Aspects of Connected Discourse. *Perception and Psychophysics*, 2, 422-437.

# Organising Principles in Lexical Representation: Evidence from Polish

Agnieszka Reid (agnieszka.reid@mrc-cbu.cam.ac.uk)

William Marslen-Wilson (william.marslen-wilson@mrc-cbu.cam.ac.uk)  
MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 2EF, UK.

## Abstract

Cross-linguistic research into the structure of the mental lexicon potentially allows us to deconfound factors which are language specific from factors which are cross-linguistically universal. In a series of three experiments we provide preliminary evidence for the structure of the Polish lexicon, which belongs to the Slavonic language family. As in English, semantic compositionality plays a crucial role, so that semantically compositional, morphologically complex words are stored in a combinatorial fashion, and semantically opaque words seem to be represented as full forms. At the same time, clear evidence is found for priming between derivational and inflectional affixes, and for interference effects between suffixed words competing for the same underlying stem. Overall the data support a combinatorial and decompositional approach to lexical representation.

## Introduction

To begin to discover the organising principles underlying the representation and processing of lexical knowledge, it is necessary to conduct comparable research programmes across a variety of different languages. In the studies reported here, we take as a starting point a body of research on English (Marslen-Wilson, Tyler, Waksler & Older, 1994), and ask whether the general properties that seem to emerge for English can be found to operate for Polish, a language with a much more complex and developed morphological system.

Two kinds of claim are made for English. The first is that underlying representations of morphologically complex forms, both derivational and inflectional, are fundamentally decompositional and combinatorial in nature. Evidence for this came from three main sources. Marslen-Wilson et al. (1994) report extensive priming, in an immediate cross-modal repetition priming task, between suffixed and prefixed words sharing the same stem. At the same time, they also report the phenomenon of suffix-suffix interference, where semantically transparent pairs such as *government - governor* do not prime, despite sharing the same stem. Marslen-Wilson et al. (1994) interpreted this as evidence for competition between different affixes for attachment to the same underlying stem. Thirdly, and perhaps most compellingly, Marslen-Wilson, Ford, Older & Zhou (1996) demonstrate strong priming between derivational affixes, as in pairs like *toughness/darkness* and *rearrange/rethink*. Affixes like *'-ness'* or *'re-'* appear to be isolable and independent structures in the mental lexicon, participating in a

dynamic and combinatorial manner in the representation of many different words.

The second important claim is that, cutting across this evidence for decompositional morphemically based representation, the further factor of semantic transparency plays a crucial role in determining the representation of morphologically complex words. Marslen-Wilson et al. (1994) found that semantically transparent morphologically complex words, such as *darkness-dark* prime each other, but that morphologically related, semantically opaque pairs, such as *department - depart*, do not, indicating that words such as *department* are stored as full forms. Marslen-Wilson et al. (1994) made the argument that this reflects choices made during the language acquisition process, where the language learner rejects a decompositional analysis of *department* (as *depart + ment*) on the grounds that this delivers the incorrect semantics.

## Cross-linguistic research

The broader status of these claims about the structure of lexical representation – as fundamentally decompositional but conditioned by semantic factors – remains hard to interpret unless comparable bodies of research, using parallel techniques, are conducted across a typologically contrasting sample of the world's languages. Research of this type is only now starting to emerge, and is already suggesting illuminating contrasts with the patterns proposed for English.

A salient example is the contrasting importance of semantic factors in Semitic languages, such as Hebrew and Arabic, as opposed to English. Hebrew and Arabic are characterised by non-linear morphological processes which operate on roots and word patterns. The most striking feature of this morphological system is that morphemes are not combined linearly, but a root, which usually consists of three consonants, is interleaved in a discontinuous manner with a word pattern, to create the phonetic surface form. Deutsch & Frost (1998) demonstrated that in Hebrew, words which are morphologically but not semantically related, prime each other strongly, in contrast to the findings on English. More recently Boudelaa & Marslen-Wilson (2000) demonstrate comparable findings for Arabic, using both cross-modal and masked priming tasks, and finding equally strong priming between prime target pairs sharing the same roots, irrespective of semantic transparency.

The finding that semantic transparency is a crucial factor in the structure of the English mental lexicon, but seems to play no role in the morphological decomposition of Semitic words, is hard to interpret on its own, because of the many

ways in which languages like Hebrew and Arabic contrast with a language like English. One of the goals of the research reported below is to add another typologically distinct data point to these contrasts, asking for Polish not only whether there is comparable evidence here for decompositional representation, but also whether semantic factors play a critical role in determining whether or not complex forms are represented in decompositional format.

## Research on Polish

A striking characteristic of Polish, a member of the Slavonic language family, is the richness of its morphological systems. Almost every word in Polish exists within a very rich paradigm, declensional for nouns, adjectives, numerals and pronouns or conjugational for verbs. The derivational morphology is comparable to English, being based on concatenative processes of prefixation and suffixation, but includes a number of qualitatively very different affixes, for instance verbal aspectual prefixes, aspectual-derivational prefixes and diminutival suffixes. Also, as far as derivational suffixes are concerned, they are considerably more numerous. Polish permits the formation of morphologically very complex words, such as secondary imperfectives described below, which allows a challenging test of claims about combinatorial representation and access.

## Experiment 1

The main goals of the first experiment we report here were to investigate morphological phenomena that are absent in English, as well as to investigate parallel phenomena in the two languages. To do this we used the cross-modal immediate lexical decision task (Marslen-Wilson et al. 1994). In this task subjects hear an auditory prime, at the offset of which, they immediately see a visual target (for 500 ms) and have to decide, by pressing an appropriate button, whether a target word is a real word or a non-word.

Taking advantage of the range of qualitatively different affixes in Polish, we probed their representation in pairs of semantically transparent words, which share the same affixes. The stimuli included (a) 24 pairs of verbs which share the same aspectual prefix, e.g. *skorzystać* 'to benefit, Perfective' - *straciły* 'they lost, Perfective'; (b) 22 verbs which share the same aspectual-derivational prefix, for instance: *nagrzać* 'to heat up, Perfective' - *nakroiła* 'she cut, Perfective'; (c) 18 nouns which share the same diminutive suffix, e.g. *kotek* 'a little cat' - *ogródek* 'a little garden' and (d) 24 nouns which have the same derivational suffix, e.g. *kucharz* 'a cook' - *piłkarz* 'a footballer'. Also, having in mind the difference in findings on English and Hebrew/Arabic regarding words which are morphologically related, but semantically opaque, we included (e) 22 pairs, such as *jalowiec* 'juniper' - *jalowy* 'futile', as a test of whether semantically opaque words prime each other. It seemed plausible that these pairs would prime in a language such as Polish, where the dynamics of morphological processing are much stronger than in English. We will refer to them as [+Morph, - Sem]. We also included a condition (f) 20 pairs which share the same stem, e.g., *szycie* 'sewing' - *szyć* 'to sew'. Because many studies document a robust effect of

stem priming, this condition served as a test of the procedure in our experiment. Finally, we included (g) 20 semantically related pairs, e.g., *kokos* 'coconut' - *banan* 'banana', which also served as a test of the experimental procedure. Many experiments on English found semantic priming in the cross-modal priming. We will refer to them as [-Morph, +Sem]. In addition, to investigate whether any observed priming in affix conditions was due to pure phonological overlap, we included two control conditions where the stimuli were phonologically but not morphologically or semantically related: (h) 18 words with phonological overlap at the onset, e.g. *numer* 'a number' - *nuda* 'boredom' and (i) 18 with overlap at the word offset, e.g. *hałas* 'noise' - *szałas* 'a shelter'.

## Results

6 subjects from version 1 and 4 subjects from version 2 were discarded from the analysis, because of high error percentage on real words (equal to or above 15%) or/and slow mean reaction times to real words (equal to or above 1000 ms). A total of 20 subjects per version was entered into the analysis. All subjects were in their twenties, and were native Polish speakers living and studying in Poland. 7 items were removed from the analysis: 3 because of high error percentage (equal to or above 30% on both versions or equal to or above 40% on one version) and 4 because of homophony. Every reaction time was inversely transformed in order to reduce the influence of outliers. The inversely transformed data were analysed in a Repeated Measures ANOVA separately for items (F2) and for subjects (F1). See Figure 1 for details of the descriptive statistics.

First the overall repeated measures ANOVA with Prime (related, unrelated) and Condition (1-9) was run. There was a main effect of Prime, indicating that RTs were faster for targets when preceded by a related prime than an unrelated prime,  $F(2,163)=22.62$ ,  $p<0.001$ ;  $F(1, 38)=37.32$ ,  $p<0.001$ . The main effect of Condition was significant,  $F(8, 163)=17.82$ ,  $p<0.001$ ;  $F(8, 304) = 162.89$ ,  $p<0.001$ . However, there was also a significant two-way interaction of Condition x Prime  $F(8,163) = 4.49$ ,  $p<0.001$ ;  $F(8, 304) = 7.45$ ,  $p<0.001$ .

The finding that there was 18 ms of priming on average in all the affix conditions treated as a group was explored further in an ANOVA. The results showed that there was a main effect of Prime,  $F(2, 75)=12.06$ ,  $p<0.001$ ,  $F(1, 38)=11.77$ ,  $p<0.001$ . The main effect of Condition was also significant,  $F(3, 75)=19.54$ ,  $p<0.001$ ,  $F(3, 114)=209.88$ ,  $p<0.001$ , with no interaction between Condition x Prime,  $F(3,75)=0.14$ ,  $p>0.05$ ,  $F(3,114)=0.37$ ,  $p>0.05$ . This result indicates that there was a facilitatory effect of Prime in all affix conditions treated as a group.

We then conducted an analysis of simple effects of Prime on each level of Condition in the remaining Conditions. The results show no facilitatory priming for [+Morph,-Sem] pairs,  $F(2,20)=0.56$ ,  $p>0.05$ ;  $F(1,38)=0.72$ ,  $p>0.05$ . There was no priming for either of the Phonological Overlap conditions: Phonological Overlap at the Word Onset,  $F(2,16)=0.05$ ,  $p>0.05$ ,  $F(1,38)=0.62$ ,  $p>0.05$  and Phonological Overlap at the Word Offset  $F(2,16)=2.15$ ,  $p>0.05$ ,  $F(1,38)=3.98$ ,  $p>0.05$ . On the other hand, there

was clear priming in the Stem Condition  $F(2,18)=25.0$ ,  $p<0.001$ ,  $F(1,38)=53.43$ ,  $p<0.001$  and in Semantically, but not Morphologically Related Pairs  $F(2,18)=13.36$ ,  $p<0.01$ ,  $F(1,38)=22.10$ ,  $p<0.001$ .

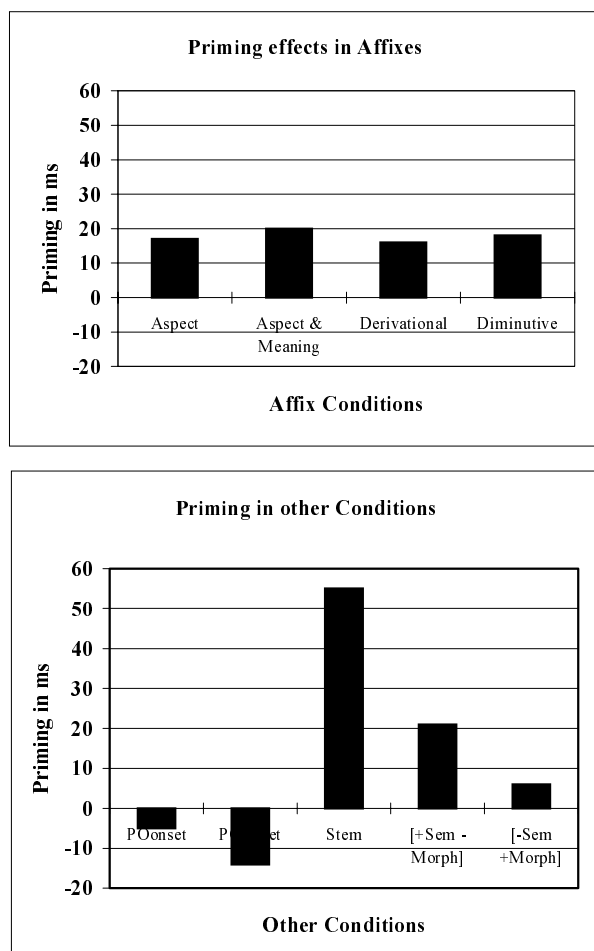


Figure 1. Priming effects for Experiment 1.

## Discussion

The results of Experiment 1 show clear priming in all the Affix conditions treated as a group as well as in the Stem condition. The absence of priming in the two phonological overlap conditions indicates that the priming obtained in the affix conditions cannot be attributed to simple phonological overlap. The results show that affixes and stems are isolable and independent structures in the Polish mental lexicon. Polish affixes, although qualitatively different from English affixes, seem to be stored in a combinatorial manner. On the other hand, the evidence shows that morphologically related, semantically opaque words do not prime each other, indicating that they are stored as full forms. This indicates that the factor of semantic compositionality determines the representation of morphologically complex words in Polish. This is in line with the findings on English, but is in contrast with Hebrew and Arabic, where semantic compositionality

does not determine the representation of morphologically complex words.

The combinatorial storage of affixes in the Polish mental lexicon is also supported by evidence on a Polish Wernicke's aphasic patient (Ulatowska and Sadowska, 1988). In tests of production, the patient occasionally made mistakes involving derivational morphology. When asked to produce a word denoting a little plate, she produced *tależek*, using an existing, but incorrect diminutival suffix, instead of saying *tależyk*. When verbal aspectual morphology is considered, the patient produced an incorrect form, such as *z-siwiał*, instead of *o-siwiał* 'he got grey', substituting a correct aspectual prefix 'o-' with an incorrect one 'z-' for this verb. Although the origin of these errors may be partially conditioned by phonological deficits in the patient's language output system, it seems to be hard to account for these errors only in these terms. The errors include an incorrect combination of existing morphemes, rather than a combination of non-existing units. Hence, we take this as a further evidence in support for the combinatorial storage of words in the Polish mental lexicon.

## Experiment 2

The findings on affix priming in Polish reported in experiment 1 left us with two further questions. Firstly, does the combinatorial representation of affixed words also hold for much more complex forms? Secondly, will we get more reliable priming, in comparison to the relatively weak priming in the four individual affix conditions in Experiment 1, if two affixes are shared by the prime and target?

Highly polymorphemic, semantically compositional words, such as secondary imperfectives, which occur in Polish, are a particular challenge for the combinatorial view of the mental lexicon. On one hand their complex structure would make them potentially more difficult to parse in comprehension and assemble in production if they are represented as a combination of morphemes, rather than as full forms. On the other hand, the intuition of native speakers of Polish is that they can process highly polymorphemic forms with the same efficiency as the less complex forms. More generally, for productive complex morphological forms, it is generally accepted that simple learning of each complex form is not a plausible language acquisition procedure (e.g., Hankamer, 1989).

We used (a) 30 pairs of secondary imperfectives, which shared the same prefix and suffix, e.g. *roz-pakow-ywa-l-em* (prime) 'to unwrap, 1<sup>st</sup> person sing., masculine, past tense, secondary imperf.' and *roz-walkow-ywa-ć* (target) 'to flatten something using, a rolling-pin, secondary imperf.'. These words consisted of a derivational prefix, e.g. 'roz-', a secondary imperfective suffix '-ywa-', past tense morpheme 'ł' (prime only) and a morpheme '-em' (prime only), which denotes the 3<sup>rd</sup> person singular, masculine. To ensure an appropriate paradigm-check we also included (b) 24 standard stem priming pairs, *myśl-ę* 'I think' - *myśl-e-ć* 'to think'; and (c) 24 semantically related, but morphologically unrelated pairs, e.g. *dom* 'a house' - *garaż* 'a garage', to dissociate the morphological and semantic effects.

Because we wished to avoid possible confounds with semantic priming, we used here a different task. This was an

auditory-auditory priming experiment with 12 items intervening between prime word and target. At these long lags, it is generally found that semantic priming drops away whereas morphological priming does not (Marslen-Wilson & Tyler, 1998).

## Results

10 subjects were discarded from the analysis according to the same criteria as in experiment 1. Data from 23 (version 1) and 24 (version 2) participants were entered into the analysis. One item had to be discarded from the analysis, because of high error percentage on one version. See Figure 2 for the details on the descriptive statistics.

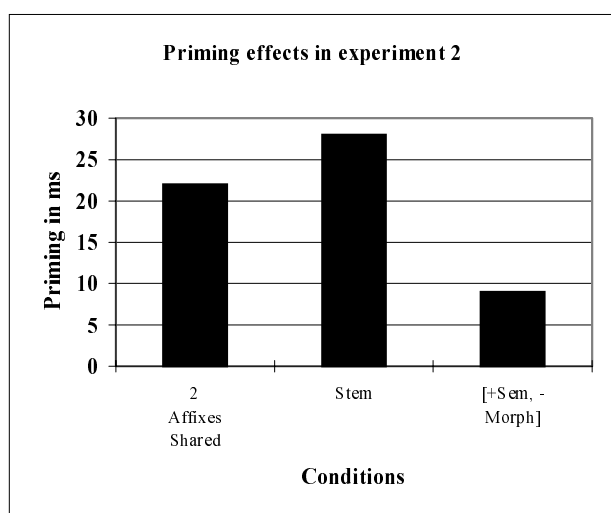


Figure 2. Priming effects for Experiment 2

The reaction time data were prepared for the analysis as described in experiment 1. The overall ANOVA revealed that the main effect of Prime was significant,  $F(2,71)=16.006$ ,  $p<0.001$ ,  $F(1,45)=507.888$ ,  $p<0.001$ . The main effect of Condition was also significant,  $F(2,71)=20.426$ ,  $p<0.001$ ,  $F(1,45)=81.308$ ,  $p<0.001$ . The two-way interaction of Prime and Condition was not significant in the items analysis,  $F(2,71)=2.004$ ,  $p>0.05$ , but it was significant in the subject analysis,  $F(2,90)=72.625$ ,  $p<0.001$ .

We then carried out an analysis of simple effects of Prime on every level of Condition. There was clear priming in the Shared Affixes Condition;  $F(1,28)=4.8$ ,  $p<0.05$ ,  $F(1,45)=7.58$ ,  $p<0.01$ . The results for the paradigm-check conditions were straight-forward: as predicted there was significant priming for the Stem Priming Condition  $F(1,22)=15.48$ ,  $p<0.01$ ;  $F(1,45)=10.98$ ,  $p<0.01$  and there was no priming in the Semantically Related, but Morphologically Unrelated Condition:  $F(1,21)=1.01$ ,  $p>0.05$ ;  $F(1,45)=0.249$ ,  $p>0.05$ .

## Discussion

Firstly, the results show reliable priming for secondary imperfectives, e.g. *roz-pakow.ywa-l-em*, which indicates that they are in fact represented in a combinatorial fashion, de-

spite their morphological complexity. Secondly, it appears that, when two affixes are repeated in prime and target, we obtain a more robust priming effect, of the magnitude of stem priming, in comparison with the relatively weak priming in the affix conditions in experiment 1. This is consistent with claims for combinatorial underlying processing mechanisms, comparable to those claimed for English, and matching the claims for somewhat different forms of underlying combinatorial systems in the non-concatenative morphologies of Hebrew and Arabic.

## Experiment 3

One of the main pieces of evidence in support of the combinatorial approach to the English mental lexicon comes from the finding that semantically transparent pairs which share the same stem and have different derivational suffixes do not prime each other. This finding has been replicated many times in English, since the original report in Marslen-Wilson et al. (1994). For instance Marslen-Wilson & Zhou (1999) show that pairs which exhibit allomorphy, e.g., *sincere-ly* & *sincer-ity* as well as non-allomorphic pairs, e.g., *excit-able* & *excite-ment* do not prime each other either in a cross-modal priming task or in an auditory-auditory lexical decision task with 0 or 8 intervening lags. The results at 8 intervening lags established that the suffix interference effect is robust and can be elicited under conditions where morphological but not semantic factors are likely to be responsible.

Because we found evidence for the combinatorial storage of morphologically complex, semantically compositional words in Polish, we wanted to test whether we would find convergent evidence from suffix interference, tested in a language system where suffixation is one of the main derivational processes.

The stimuli included (a) 32 derived - derived words which shared the same stem, but had different derivational suffixes. Half of the stimuli were deverbal derivatives, e.g. *pis-anie* 'writing' - *pis-arz* 'a writer'. The other half were denominal derivatives, e.g. *balon-owy* 'balloon like, adj.' - *balon-ik* 'a little balloon',  $SR^1 = 8.1$ ,  $SD = 0.5$ ; (b) 32 inflected - derived pairs which shared the same stem. Half of the stimuli had an inflected verb as a prime and a deverbal derivative as a target e.g. *pisa-la* 'to write, 3<sup>rd</sup> person, sing., feminine, past tense' - *pis-arz* 'a writer'. The other half had an inflected noun as a prime and a denominal derivative as a target, e.g. *balon-em* 'balloon, instrumental' - *balon-ik* 'a little balloon',  $SR = 8.4$ ,  $SD = 0.4$ ; (c) 24 stem priming pairs, as before e.g., *mysl-e* 'I think', *mysl-e-c* 'to think' were included as a paradigm check,  $SR = 8.2$ ,  $SD = 0.2$ .

<sup>1</sup> SR denotes a mean score (across 10 participants) on a Semantic Relatedness pre-test, where native speakers of Polish judged on a 9- point scale (where 9 is the highest possible score), to what degree a given pair of words is semantically related. We use these scores as a measure of Semantic Transparency between a prime and target, which is highly correlated with semantic compositionality.

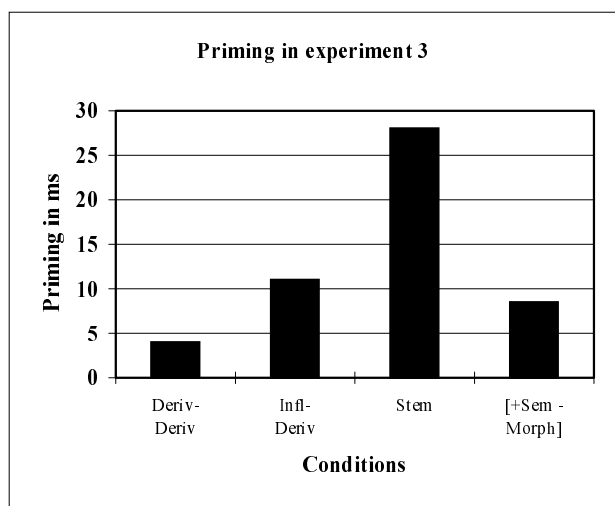


Figure 3. Priming effects for Experiment 3

Both words in the stem condition shared the same stem, the prime had an inflectional ending, denoting person, tense and number, whereas the target was an infinitive and had an infinitival marker 'ć' and (d) 24 [- Morph, +Sem] pairs, included to dissociate the morphological and semantic effects,  $SR = 8.0$ ,  $SD = 1$ . In order to ensure the most rigorous comparison between conditions 1 and 2, the same target was used in both conditions. Experimental items were assigned to 4 versions, so that the same target was preceded by one of the two related primes or one of the two unrelated primes, with one combination of a prime and a target per version. Because we were mainly interested in the morphological effects, we again used an auditory-auditory priming, lexical decision experiment with 12 intervening items to dissociate the morphological from the semantic effects.

## Results

10 participants were rejected on the same criteria as in experiment 1. A total of 89 participants: 22 (version 1), 24 (version 2), 21 (version 3) and 22 (version 4) were entered into an analysis. No experimental items were removed. See Figure 3 for the details of the descriptive statistics.

The overall repeated measures ANOVA analysis with Prime (related, unrelated) and Condition (1-4) revealed that the main effect of Prime was significant  $F(2, 96)=22.2$ ,  $p<0.001$ ,  $F(1, 85)=32.48$ ,  $p<0.001$ . The main effect of Condition was not significant in the item analysis  $F(3, 96)=0.537$ ,  $p>0.05$ , but it was significant in the subject analysis  $F(3, 255)=6.776$ ,  $p<0.001$ . A two-way interaction of Prime and Condition was significant,  $F(2, 1, 3)=2.9$ ,  $p<0.05$ ,  $F(3, 255)=2.71$ ,  $p<0.05$ .

The simple effects analysis was conducted to investigate effect of Prime at each level of Condition. The analysis showed that there was no significant priming in Derived-Derived Condition,  $F(2, 1, 28)=1.35$ ,  $p>0.05$ ,  $F(1, 85)=0.548$ ,  $p>0.05$ , nor in Inflected-Derived Condition,  $F(2, 1, 28)=3.615$ ,  $p>0.05$ ,  $F(1, 85)=2.702$ ,  $p>0.05$ . The remaining analyses on the individual conditions revealed that

there was a strong priming in Stem Condition,  $F(2, 1, 20)=15.24$ ,  $p<0.01$ ,  $F(1, 85)=19.874$ ,  $p<0.001$  and that there was no significant priming for [+Sem, -Morph] Condition,  $F(2, 1, 20)=3.9$ ,  $p>0.05$ ,  $F(1, 85)=1.821$ ,  $p>0.05$ .

## Discussion

The results are clear. There is no priming in Derived-Derived Condition nor in Inflected-Derived Condition while there is robust priming for the Stem Condition and no priming for [-Morph, +Sem] pairs. The findings for the latter two conditions have been replicated in many of our experiments (e.g. experiment 2) and are in line with the predictions.

Not finding priming for semantically transparent suffixed words which share the same stem, but have different derivational suffixes indicates that there must be an inhibitory process between the suffixes. Hearing *pis-anie* 'writing', as a prime inhibits the combination of the root *pis-* with another suffix, e.g. *-arz*, hence the recognition of the target *pis-arz* 'a writer' is slowed, even though the root *pis-* is active. This finding parallels the findings reported for English (Marslen-Wilson et al. 1994).

Interestingly, the results also suggest that there is suffix interference in Inflected-Derived Condition between inflectional and derivational suffixes which are attached to the same stem in semantically compositional pairs. It is hard to see how this could be the case if both types of suffix were not represented in the lexicon.

Most influential linguistic models of word formation in generative grammar assume that inflections are not represented lexically, but they are added by syntactic rules which are outside the mental lexicon. This is supported by data from the lexical decision experiments on English, but not by the findings on Polish. One plausible source of the difference between the findings for Polish and English comes from the characteristics of the Polish inflectional system, which in contrast to English is extremely rich and carries a lot of very complex information. The findings (although from a different paradigm) on Italian (Miceli & Caramazza 1988) which is a morphologically rich language, similarly to findings on Polish support the claim that the inflectional suffixes are stored in the mental lexicon.

The suffix-suffix interference leaves, at the current stage, at least one unresolved issue. All the inflectional suffixes of the primes for deverbal targets in Inflected-Derived Condition were 3<sup>rd</sup> person singular, masculine or feminine, past tense. In our previous experiments, we found priming for pairs which shared the same stem and where prime had a derivational suffix whereas the target had an infinitival ending - 'ć'. The question which arises is: what is special about the infinitival suffix that it does not cause interference with a derivational suffix? One possible explanation is that the infinitival ending does not have the same linguistic status as the inflectional suffixes, which carry a lot of information, e.g. person, number, gender, tense, etc. An issue which has to be resolved in our future research is whether suffix interference occurs for two inflectional suffixes. This will provide a more stringent test of the representation of the Polish inflectional morphemes in the lexicon.

## General Discussion

We have reported the findings on the Polish mental lexicon in a series of three experiments, in an attempt to examine the organising principles affecting the structure of the mental lexicon of a morphologically complex language from the Slavonic family. In the first experiment we concentrated on probing the representation of morphologically complex words, which included affixes which are qualitatively different from English. The findings indicated that affixes are represented in a combinatorial fashion. Secondly, the results show that semantic compositionality is an important factor in determining the lexical representation in the Polish mental lexicon. In the second experiment we confirmed that a combinatorial representation also holds for words with a much more morphologically complex structure, such as the secondary imperfectives, at the same time confirming the existence of strong priming between derivational and inflectional affixes. Finally, in the third experiment we addressed the issue of suffix interference in Polish, finding clear evidence for interference in derived-derived pairs as well as in inflected-derived pairs. This is further evidence for underlying combinatorial representations and processes.

In summary, the overall picture which has emerged as a result of our investigation of the Polish mental lexicon is that, Polish, similar to English and Hebrew is characterised by a combinatorial mental lexicon. However, different factors which condition the structure of the mental lexicon have different 'weightings' in Polish as in comparison with Hebrew and English. The factor of semantic compositionality is crucial in determining the structure of the representation of words in Polish, similarly to English, but in contrast with Hebrew (and Arabic). On the other hand, the factor of the type of inflectional morpheme is important in the structure of the Polish lexicon, in that both types of inflectional morphemes verbal and nominal seem to be represented in the Polish lexicon. This contrasts with English, where neither verbal nor nominal inflections seem to be represented as lexical processing structures.

## Acknowledgements

The research reported here is supported in part by an ESRC research studentship to A. Reid, and in part by the UK Medical Research Council. We would like to thank K. Kowalik, M. Smoczyńska and B. Szymanek for very helpful comments made during the selection of the stimuli. We also would like to thank P. Brzuski who kindly provided the facilities in Poland to run our experiments.

## References

- Boudelaa, S., & Marslen-Wilson, W. D. (2000). On The Use of Word Pattern Morphemes in Modern Standard Arabic. *The Fourteenth Annual Symposium on Arabic Linguistics*. The University of California at Berkeley.
- Deutsch, A., & Frost, R. (1998). Verbs and nouns are organized and accessed differently in the mental lexicon: evidence from Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24, 1238-1255.

- Hankamer, J. (1989). Morphological parsing and the lexicon. In Marslen-Wilson, W. D. (Ed.), *Lexical representation and process*. Cambridge, MA: MIT Press.
- Miceli, G., & Caramazza, A. (1988). Dissociation of inflectional and derivational morphology. *Brain and language* 35, 24-65.
- Marslen-Wilson, W. D., Ford, M., Older, L., & Zhou, X. (1996). The combinatorial lexicon: priming derivational affixes. *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (pp. 223-227). La Jolla, California, Lawrence Erlbaum Associates.
- Marslen-Wilson, W. D., & Tyler, L. K. (1998). Rules, representations, and the English past tense. *Trends in Cognitive Science* 2, 419-463.
- Marslen-Wilson, W. D., Tyler, L., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review* 101, 3-33.
- Marslen-Wilson, W. D., & Zhou, X. (1999) Abstractness, allomorphy, and lexical architecture. *Language and Cognitive Processes*, 14, 321-352.
- Ulatowska, H., K., & Sadowska, M. (1988). Features of agrammatism in Polish - a case study. *Journal of Neurolinguistics* 3, 77-88.

# From Studying Examples to Solving Problems: Fading Worked-Out Solution Steps Helps Learning

**Alexander Renkl** (renkl@psychologie.uni-freiburg.de)

University of Freiburg; Department of Educational Psychology; Belfortstr. 16  
D-79085 Freiburg; Germany

**Robert K. Atkinson** (atkinson@ra.msstate.edu)

Department Counselor Education and Educational Psychology; Box 9727  
Mississippi State, MS 39762; USA

**Uwe H. Maier** (uwe.harald.maier@t-online.de)

Educational University of Schwäbisch Gmünd; Oberbettringer Str. 200  
D-73525 Schwäbisch Gmünd; Germany

## Abstract

Research has shown that it is effective to combine example study and problem solving in the initial acquisition of cognitive skills. Present methods for combining these learning modes are, however, static and do not support a transition from example study in early stages of skill acquisition to later problem solving. Against this background, we propose a successive integration of problem-solving elements into example study until the learners solve problems on their own (i.e., complete example → increasingly more incomplete examples → problem to-be-solved). We tested the effectiveness of such a fading procedure against the traditional method of employing example-problem pairs. In a field experiment and in a more controlled lab experiment, we found that the fading procedure fosters learning, at least when near transfer performance is considered. Moreover, this effect is mediated by a lower number of errors under the fading condition as compared to the example-problem condition.

## Introduction

Worked-out examples consist of a problem formulation, solution steps, and the final solution itself. Research has shown that learning from such examples is of major importance for the initial acquisition of cognitive skills in well-structured domains such as mathematics, physics, and programming (for an overview see VanLehn, 1996). In addition, novices prefer this learning mode, and they are right: It is quite an effective way of learning. Studies performed by Sweller and his colleagues (e.g., Sweller & Cooper, 1985; for an overview see Sweller, van Merriënboer, & Paas, 1998) showed that learning from worked-out examples can be more effective than learning by problem solving.

Although worked-out examples have significant advantages, their employment as a learning methodology does not, of course, guarantee effective learning. First, the extent to which learners profit from the study of examples depends on how well they explain the solutions of the examples to themselves (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Renkl, 1997). Second, it is important how the learning materials (examples and problems) are structured (cf. Atkinson,

Derry, Renkl, & Wortham, in press). The second aspect is the focus of this study. More specifically, this study investigates one possible approach to integrating elements of problem solving into example study. We propose that these learning modes can be combined by successively introducing more and more elements of problem solving in example study until learners are solving the problems on their own. This rationale can also be used as a way to structure the transition from studying examples in initial skill acquisition to problem solving in later phases of the learning process.

In the next section, the literature with respect to the issue of combining example study and problem solving is discussed. Then we outline open questions and give preliminary answers that were tested in two studies, first in a field experiment and then in a more controlled lab experiment.

## How to Combine Example Study and Problem Solving? – State of the Art

Empirical evidence has shown that pure example study (i.e., examples alone) is not as effective as learning from examples in which elements of problem solving are integrated. There are two traditional ways to combine example study and problem solving: (1) Making the solutions of examples incomplete and (2) employing example-problem pairs.

### Incomplete Examples

Some researchers argue that incomplete examples, which the learners have to complete, effectively support the acquisition of cognitive skills (van Merriënboer, 1990; van Merriënboer & de Crook, 1992; Paas, 1992; Stark, 1999). Stark (1999) conducted a controlled experiment designed to examine the extent to which the insertion of “blanks” into the solution of examples—which, in a certain sense, forced the learners to determine the next solution step on their own—fostered learning. In his study, half of the participants studied incomplete examples (experimental group), while the other half learned from complete examples (control group). In the experimental group, portions of the example solutions presented to the participants were replaced by “question marks.” The learners were then asked to identify what solution step



was missing. After doing that, or at least making the attempt, the complete solution step was presented so that learners received feedback on the correctness of their anticipation. When compared to studying complete examples, Stark found that incomplete examples fostered the quality of self-explanations and, as a consequence, the transfer of learned solution methods. The results of Stark's study contrast with observations by Paas (1992), who did not find any difference in performance among participants presented with either incomplete or complete examples. However, the main purpose of Paas' study was not to investigate the effects of complete versus incomplete examples. Taken together, the results of Stark (1999) show that making examples incomplete (at least) can support learning.

### Example-Problem Pairs

Sweller and his colleagues (e.g., Sweller & Cooper, 1985) have conducted several classic studies documenting the effectiveness of learning from worked-out examples. However, in these studies the authors did not compare *pure* learning from examples (worked-out solutions only) with *pure* learning by problem solving (problems to-be-solved only). That is, these empirical examinations did not examine the impact of studying examples exclusively with solving practice problems only. Instead, the example condition usually consisted of examples followed by isomorphic problems to-be-solved (example-problem pairs). Thus, the studies of Sweller and colleagues mainly showed that combined learning from examples and problems is more effective than learning by solving problems.

Studies on learning from worked-out examples performed by other researchers have focussed on pure learning from examples (e.g., Renkl, 1997). Explicit comparisons between pure example learning and learning from example-problem pairs are, however, rare. One such study was performed by Trafton and Reiser (1993), in which the authors designed two treatments, alternating and blocked: Participants in the alternating condition were exposed to six example-problem pairs, where each example was followed directly by a isomorphic problem, while participants in the blocked condition were exposed to the entire set of six examples, followed by the entire set of six practice problems. The authors found that, as predicted, participants in the alternating-example condition took less time and produced more accurate solutions on the transfer posttest than their counterparts in the blocked-example condition. Based on these findings, the authors asserted that "the most efficient way to present material to acquire a skill is to present an example, then a similar problem to solve immediately following" (Trafton & Reiser, 1993, p. 1022).

In a recent study, Stark, Gruber, Renkl, and Mandl (in press) examined whether there might be another effective variation of the traditional method of pairing examples with practice problems. Based on a study of learning diagnostic strategies in medicine in which it was found that it is more effective to learn from a "cognitive model" (which can also be regarded as a kind of worked-out example) after an initial problem solving experience (Gräsel & Mandl, 1993), the authors argued that presenting practice problems first followed by isomorphic examples (problem-example pairs)

should be an effective mode of instruction. Specifically, the authors proposed that initial problem solving difficulties should motivate the learners to process the examples that followed more deeply and, in particular, more focussed with respect to the specific difficulties the individual learners have in solving such problems. In a comparison between pure example learning and learning from problem-example pairs (domain: calculation of compound and real interest), it was found that the combined learning method (i.e., problem-example pairs) substantially fostered active example processing and, as a result, learning outcomes.

Taken together, combining practice problems and examples is obviously more effective than exposing learners to either of the two pure learning conditions, that is, either to sets of practice problems or sets of examples.

### Open Questions and Answers to be Validated

Although there can be little doubt on the effectiveness of a combined learning method, two questions still remain open: (1) Are there more effective ways of combining example study and problem solving than presenting incomplete examples or pairs of examples and problems? (2) What is a sensible rationale for designing the transition from learning from examples in initial stages of cognitive skill acquisition to problem solving in later stages?

Instructional models such as *Cognitive Apprenticeship* (Collins, Brown, & Newman, 1989) propose a smooth transition from modeling to scaffolded problem solving to independent problem solving in which instructional support fades during the transition. The use of incomplete examples, at least as realized in previous studies, has not incorporated such a dynamic fading component. To date, studies incorporating the "pairs arrangement" have also not used a fading component. In fact, these studies typically contain abrupt transitions from examples, as a type of model, to independent problem solving. Against this background, it is sensible to combine problem solving and example study in the following way. First, a complete example is presented (model). Second, an example is given in which one single solution step is omitted (scaffolded problem solving). Then, the number of blanks is increased step-by step until just the problem formulation is left, that is, a problem to-be-solved (independent problem solving). In this way, a smooth transition from modeling (complete example) over scaffolded problem solving (incomplete example) to independent problem solving is implemented. This rationale provides possible answers to the open questions outlined above.

### Experiment 1: Field Experiment

As a first test of our assumptions we conducted a small-scale field experiment in which we tested whether a smooth transition from example study to problem solving (gradual insertion of blanks into the solutions of examples) is more effective than learning by example-problem pairs as they are used in many studies on learning from examples. As a method of fading out the solution steps, we choose to first omit the last solution step, then the last two steps, and finally all three steps ("backward rationale").

## Methods

**Sample and Design.** Two ninth-grade classrooms from a German *Hauptschule* (lowest track of the German three track system) participated in this quasi-experiment. In both classrooms, the same teacher (third author) conducted a physics lesson on electricity based on four examples/problems. In one classroom ( $n = 20$ ) a fading procedure was used and in the other classroom ( $n = 15$ ) traditional example-problem pairs were employed. Each example/problem involved three solution steps. Across both conditions half of the steps were worked-out whereas the other half was to be generated. Thus, learners in both conditions were required to solve the same number of solution steps.

**Learning Environment.** In the experimental phase, the third author (a professional teacher) conducted a 45 minute lesson in each classroom. Both groups worked on four examples/problems in which the cost for running a variety of electric devices for a certain time had to be determined (e.g., "A aluminum factory has a big melting furnace which is run with 1000 V. A power of 20 A has to flow through the furnace in order to melt aluminum. What does the factory have to pay per month when the furnace always runs and the kWh costs DM 0.22?"). Although the examples/problems were printed on work sheets, the problem formulation of each example/problem was read aloud by one of the students from the class. Following the reading of the problem formulation, the students were permitted to ask clarifying questions (of course, no questions on the solution) before working individually on the example or problem. At the end of each incomplete example or problem, the complete solution was presented on an overhead transparency and, if necessary, the students corrected or supplemented their solutions. Then the teacher proceeded to the next example/problem.

In the fading classroom, the teacher presented the instruction in the following order: (1) a complete example, (2) an example with the last solution step left out, (3) an example with the last two steps omitted, and (4) a problem where all three steps were missing. In the example-problem group in contrast, a complete example was presented twice, each time it was followed by a corresponding problem.

**Procedure.** The overall procedure was identical in both classrooms. Basic knowledge of the concepts and rules of electricity was introduced in the context of regular instruction followed by a pretest that tapped into prior knowledge with respect to the ability to apply the abstract rules to domain problems. Two days later, the school lessons in which the experimental variation took place were conducted. Finally, after additional two days, the students worked on a posttest.

**Instruments.** The pretest consisted of four problems from the physics domain of electricity that were structurally equivalent to the problems in the posttest (e.g., "The electronic motor of an electronic locomotive is supplied by a voltage of 0.6 kV. In the average, a current of 18 A flows through the motor. What does an eight-hour trip from Stuttgart to Hamburg cost when you assume that the *German Railway* pays DM 0.12 per kWh?"). For the correct solution

of an item, a maximum of three points was assigned. For partly correct solutions partial credit was dispensed. The score was divided by the theoretical maximum score (12) so that it represent the percentage of points in relation to perfect performance. The pretest had a sufficient reliability (Cronbach's Alpha: .87).

The posttests consisted of six problems. The four near-transfer problems had the same underlying structure (solution rationale) as the examples and problems employed in the learning phase but different surface features (cover story, numbers). Two problems were classified as far transfer because both the underlying structure and the surface features differed (e.g., "Tanja pays for her frig DM 40 per year. One kWh costs DM 0.22. What power does the frig have if you assume that it runs all the time?"). For the correct solution on a posttest problem, which always included three solution steps, three points were dispensed. Partial credit was given for partly correct solutions (1 or 2 points). The scores for both scales were finally divided by the theoretical maximum score (12 or 6 respectively) so that they represented the percentage of points in relation to perfect performance. We obtained sufficient reliabilities (Cronbach's Alphas) for both posttest scales: .85 for near transfer and .60 for far transfer.

## Results

Table 1 shows the means and standard deviations of the two experimental groups on the pretest and the posttest scores. Both groups showed almost identical pretest performance ( $t(33) = 0.01$ ;  $p > .10$ ). Hence, there was no *a priori* difference between groups with respect to prior knowledge.

Table 1: Group means (standard deviations in brackets) of pretest and posttest scores.

	Fading	Example-problem pairs
Pretest	24.06 (28.12)	23.96 (29.02)
Posttest: near transfer	79.38 (27.42)	62.22 (24.82)
Posttest: far transfer	36.25 (37.29)	21.11 (27.61)

With respect to treatment effects we descriptively obtained higher means in the fading group for both near and far transfer. Comparisons between the experimental conditions by means of an ANCOVA (controlling for prior knowledge) yielded a significant difference for near transfer performance ( $F(1,32) = 4.44$ ;  $p < .05$ ). The group difference in far transfer performance failed to reach the level of significance ( $F(1,32) = 2.28$ ;  $p > .10$ ). Thus, the fading procedure clearly fostered near transfer performance. We can not, however, claim that this is also true for far transfer performance.

## Discussion

We obtained a positive effect of our fading procedure with respect to near transfer performance. The far transfer performance was also superior in the fading group, but not at the level of statistical significance. Before theorizing on

possible reasons for potential differential effects of the fading procedure on near and far transfer, we should wait and see whether the respective finding can be replicated.

A replication is necessary because a field study such as the present one always has some factors that might diminish the internal validity of the findings. For example, the teacher that conducted the instruction in both classrooms was not "blind" with respect to the experimental expectations. Furthermore, the present investigation was "merely" a quasi-experiment (no random assignment of participants to the experimental conditions). Hence, the conditions in both classrooms might not have been totally identical except for the independent variable (fading vs. example-problem pairs). Finally, no data on possible processes that mediate the effects of the fading procedure on the learning outcomes were recorded. These issues were addressed in Experiment 2.

## Experiment 2: Lab Experiment

In order to conceptually replicate the results of the preceding field experiment under more controlled conditions, we ran a lab experiment. We also tested for one possible mediating mechanism that may explain the effect found in Experiment 1.

As outlined above, there are quite abrupt changes with respect to the demands placed on the learners in the example-problem conditions. After a first example, the learners have to solve a whole problem totally on their own. In the fading procedure, the first problem solving demand is to generate just a single step, and the demands are only gradually increased. Against this background, we expect that the learners will make fewer errors during learning in the fading condition. In the example-problem condition, in contrast, we expect a relatively high number of errors during learning that may prevent rapid learning progress. This assumption was tested in Experiment 2.

In order to see whether the effects of the fading procedure are robust against variations in its concrete implementation, we did not use a "backward", but a "forward rationale" in this study. This means that firstly the first step was omitted, then the first two steps, and finally all three steps.

## Methods

**Sample and Design.** The participants of this study were 54 students of psychology (Mississippi State University). They were randomly assigned to the fading or to the example-problem condition, respectively ( $n = 27$  in each group). As with our field experiment, the number of unsolved solution steps was held constant across both conditions.

**Learning Environment.** A computer-based learning program was employed that had been originally developed by Renkl (1997), modified by Stark (1999), and finally adapted to the present needs by the second author. It presented worked-out examples and problems from the domain of probability (e.g., "Jonathan has recently bought a new camera. Independently of each other he frequently makes two errors when he takes a picture. He manages to blur the image in 40% of his photos ( $p=2/5$ ) and he forgets to activate the flash in 10% of the photos ( $p=1/10$ ) so that the pictures end up too dark. If you randomly choose one of Jonathan's

developed pictures, what is the probability that it will be flawless?"). The examples/problems were displayed in a step-by-step procedure. On the first page of an example/problem, the problem givens were displayed. The learners could read them and then go to the next page where a first solution step was presented or the learners were required to determine a solution step on their own (or at least to attempt it). After inspecting or determining this solution step, the participants proceeded to the following page where the next solution step was added or required, and so on. When the whole solution of a problem was presented or required, the next page contained the first page of a new example/problem until the lesson was completed. In the case of omitted solution steps, the learners had to type in a solution attempt. Hence, the correctness of the problem solving attempts could be determined. Note that the correct step was always displayed when the learners went to the next page so that there was feedback on the correctness of the learners' problem solving attempts.

On the whole, there were two sets of four probability tasks. Each set consisted of four tasks with the same underlying structure (solution rationale) but different surface features (cover stories, numbers). In the fading group, the first task was a completely worked-out example. In the second task, the first solution step was omitted. In the third task the first two steps were omitted ("forward rationale" of omitting solution steps). The fourth task was essentially a problem-solving task (all three steps were missing). In the example-problem group, two such pairs (i.e., example-problem) were presented.

**Procedure.** The participants worked in group sessions lasting about 90 minutes. They worked individually in front of a computer. First, a pretest on prior knowledge in probability calculation was presented. In order to provide or re-activate basic knowledge that allowed the participants to understand the worked-out examples, an instructional text on basic principles of probability calculation was given to the participants. After reading this instructional text, the participants were to study the worked-out examples and problems provided by the computer program. In this phase, the experimental variation took place (fading vs. example-problem pairs). The time spent for learning was recorded. Finally, the participants worked on a posttest.

**Instruments.** A *pretest* was employed in order to assess prior knowledge. It consisted of nine relatively simple problems involving probability calculation (e.g., "When rolling a 6-sided die what is the probability that '2' or '4' will appear?"). For each correct solution, one point was dispensed (no partial credit). The overall score was divided by the theoretical maximum score (9) so that it represents the percentage of points in relation to perfect performance. We obtained a sufficient reliability of .73 (Cronbach's Alpha).

The learning outcomes were assessed by a *posttest* that included thirteen problems. Besides one very simple warm-up problem, which was ignored for further analysis, we employed six near transfer items and six far transfer items. As compared to the examples/problems studied during the learning phase, the near transfer problems had the same underlying

ing structure (solution rationale) but different surface features (cover story, numbers; e.g., "While preparing a batch of rolls at the local bakery, the baker's assistant forgot to add salt to 30% of the rolls and, independent of this event, he burned 40% of the rolls. If the head baker arrives to examine the quality of his assistant's work by randomly testing a roll, what is the probability that it is edible; that is, that it has the right amount of salt and is not burned?"). Far transfer problems differed with respect to both structure and surface features (e.g., "When driving to work, Mrs. Fast has to pass the same traffic light twice—once in the morning and once in the evening. It is green in 70% of the cases. What is the probability that she can pass through a green light in the morning but has to stop in the evening?").

For the totally correct solution on a posttest problem, which always included three solution steps, three points were dispensed. Partial credit was provided for partially correct solutions (1 or 2 points). The scores for both scales were finally divided by the theoretical maximum score (18) so that they represent the percentage of points in relation to perfect performance. We obtained sufficient reliabilities (Cronbach's Alphas) for both posttest scales: .91 for near transfer and .75 for far transfer.

## Results

Table 2 shows the means and the standard deviations of the two experimental groups for the pretest (prior knowledge), the time spent for studying the examples and problems (learning time), the proportion of correct solutions steps generated during learning, and posttest performance with regard to near transfer and to far transfer. The small difference between the pretest scores in favor of the example-problem group was not statistically significant ( $t(52) = -0.49; p > .10$ ). Hence, the groups were *a priori* comparable with respect to prior knowledge. In addition, the learning time did not significantly differ between groups ( $t(52) = 0.28; p > .10$ ). Thus, possible group differences with respect to learning could not be simply attributed to time-on-task.

Table 2: Group means (standard deviations in brackets) of the pretest, the learning time (min.), the correctness of solution steps during learning (in %), and the posttest.

	Fading	Example-problem pairs
Pretest	55.56 (23.67)	58.85 (25.93)
Learning time	31.15 (10.83)	30.37 ( 9.41)
Correctness of solution steps	66.42 (31.61)	51.81 (33.13)
Posttest: near transfer	53.91 (32.24)	43.83 (35.35)
Posttest: far transfer	38.68 (25.25)	43.42 (24.60)

With respect to treatment effects, we descriptively obtained substantially higher means in the fading group for the proportion of correct solution steps and for near transfer. We used an ANCOVA (controlling for prior knowledge) to make

comparisons between the experimental conditions that yielded a significant difference for near transfer performance ( $F(1,51) = 4.58; p < .05$ ), but not for far transfer ( $F < 1$ ). A third ANOVA revealed that there was also a significant difference between groups with respect to the proportion of correct solution steps ( $F(1,51) = 7.62; p < .05$ ).

In order to test the mediation hypothesis that fading fosters learning outcomes (at least near transfer) because less errors occur during learning, an additional ANOVA for near transfer performance was performed in which the proportion of correct solution steps was included as covariate in addition to prior knowledge. The mediation hypothesis would have been confirmed if the group effect (more or less totally) disappeared in this case (cf. Baron & Kenny, 1986). This proved to be true. The  $F$ -value for the group effect was not only smaller than 1, but was a negligible size of 0.23.

## Discussion

In the present lab experiment, we conceptually replicated the effectiveness of our fading procedure for near transfer. Both studies also yielded consistent results with respect to far transfer: No significant effect was found. We obtained these converging results even though the present study and our first investigation differed with respect to the type of learners ("low-track" students vs. university students), the learning domain (physics/electricity vs. mathematics/probability calculation), the learning setting (school lesson vs. computer-based learning in the lab), and the kind of fading out worked-out solution steps ("backward" vs. "forward"). We interpreted the stability of the findings despite these very different context conditions as an indicator that our fading procedure has a reliable effect.

Something that we did not expect in advance is that the effect of fading is restricted to near transfer. This differential effectiveness of the fading procedure may have something to do with the mediating mechanism that was identified in this study (amount of errors during learning). The analyses showed that the effect on near transfer is more or less totally mediated by the amount of errors committed during learning. Although we did not directly assess self-explanations, this result suggests that the fading procedure did not enhance learning outcomes via fostering self-explanation quality. This also helps to explain the differential effectiveness of fading. For far transfer performance (e.g., Renkl, 1997; see also Atkinson et al., in press), it is of special importance that the learners explain to themselves the rationale of solution steps in an active way so that they become aware of how domain principles can be applied in a domain and how certain goals can be achieved by certain operators. In other words, reflection about the more general aspects of specific problem solutions is necessary for far transfer. However, this process was obviously not elicited by the fading procedure. "Error-avoiding" instructional procedures such as Direct Instruction or drill-and-practice tutorials are known to effectively foster "low-level" level learning (near transfer). As our fading procedure is a method of avoiding errors during learning, it is understandable why it fosters "merely" near transfer performance.

## General Discussion

In the present study, the effectiveness of our fading rationale for designing the transition from example study to problem solving has been affirmed in an highly ecologically valid field experiment as well as in a well-controlled lab study. Thus, we have provided strong evidence that a fading procedure actually fosters near transfer. Nevertheless, there are at least three important questions left that should be addressed in further research:

(1) The results indicate that the effects of fading are more or less totally mediated by the low amount of errors during learning and not by the way in which the examples were processed (self-explanations). In order to obtain more direct evidence for this interpretation, self-explanations should be assessed in a subsequent study on fading in example-based learning. In such a study, the mediation effect involving the amount of errors should be replicated and it should be tested whether there are, as expected, no differences with respect to self-explanations.

(2) In the effort to successively optimize learning from worked-out examples, another issue related to self-explanations should be addressed. If it is true, as argued above, that the quality of self-explanations is especially important for far transfer, it should be tested whether a combination of fading and self-explanation training—such as the one developed and evaluated by Renkl, Stark, Gruber, and Mandl (1998)—can facilitate *both* near and far transfer learning.

(3) We employed two ways of fading out worked-out solution steps, a backward and forward procedure. As the context conditions in our two studies varied substantially, we could not compare the relative effectiveness of these two procedures. In addition, it may well be that other procedures are even more effective. For example, one could first omit the solution step that is the easiest one for the learners to determine, then the second easiest one and so on. Systematic experimentation on this issue is necessary in order to get information on whether different ways of fading have substantially different effects and, if so, which way of fading is the ideal one.

Taken together, this contribution has provided strong evidence for the effectiveness of our "new" rationale for the integration of example study and problem solving. However, in order for us to deeply understand the way this works and to optimize the employment of this rationale, further experiments are necessary.

## References

- Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. W. (in press). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173-1182.
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*, 145-182.
- Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction*. Hillsdale, NJ: Erlbaum.
- Gräsel, C., & Mandl, H. (1993). Förderung des Erwerbs diagnostischer Strategien in fallbasierten Lernumgebungen [Promoting the acquisition of diagnostic strategies in case-based learning environments]. *Unterrichtswissenschaft*, *21*, 355 – 369.
- Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology*, *84*, 429-434.
- Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, *21*, 1-29.
- Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary Educational Psychology*, *23*, 90-108.
- Stark, R. (1999). *Lernen mit Lösungsbeispielen. Einfluß unvollständiger Lösungsbeispiele auf Beispielelaboration, Motivation und Lernerfolg* [Learning by worked-out examples. The impact of incomplete examples on example elaboration, motivation, and learning outcomes]. Bern, Switzerland: Huber.
- Stark, R., Gruber, H., Renkl, A., & Mandl, H. (in press). Instruktionale Effekte einer kombinierten Lernmethode: Zahlt sich die Kombination von Lösungsbeispielen und Problemlöseaufgaben aus? [Instructional effects of a combined learning method: How effective is the combination of worked-out examples and problems to-be-solved?] *Zeitschrift für Pädagogische Psychologie*.
- Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, *2*, 59-89.
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251-296.
- Trafton, J. G., & Reiser, B. J. (1993). The contributions of studying examples and solving problems to skill acquisition. In M. Polson (Ed.), *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology*, *47*, 513-539.
- Van Merriënboer, J. J. G. (1990). Strategies for programming instruction in high school: Program completion vs. program generation. *Journal of Computing Research*, *6*, 265-285.
- Van Merriënboer, J. J. G., & De Crook, M. B. M. (1992). Strategies for computer-based programming instruction: Program completion vs. program generation. *Journal of Educational Computing Research*, *8*, 212-234.

# Measuring Verb Similarity

Philip Resnik and Mona Diab  
{resnik,mdiab}@umiacs.umd.edu  
Department of Linguistics and  
Institute for Advanced Computer Studies  
University of Maryland  
College Park, MD USA

## Abstract

The way we model semantic similarity is closely tied to our understanding of linguistic representations. We present several models of semantic similarity, based on differing representational assumptions, and investigate their properties via comparison with human ratings of verb similarity. The results offer insight into the bases for human similarity judgments and provide a testbed for further investigation of the interactions among syntactic properties, semantic structure, and semantic content.

## Introduction

The way we model semantic similarity is closely tied to our understanding of how linguistic representations are acquired and used. Some models of similarity, such as Tversky's (1977), assume an explicit set of features over which a similarity measure can be computed, and recent computational methods for measuring word similarity can be thought of as an update of this idea on a large scale, representing words in terms of distributional features acquired via analysis of text corpora (e.g., Brown, Della Pietra, deSouza, Lai, & Mercer, 1992; Schütze, 1993). Other methods, following in the semantic networks tradition of Quillian (1968), focus less on explicit features and more on relationships among lexical items within a conceptual taxonomy, sometimes going beyond taxonomic relationships to also take advantage of frequency information derived from corpora (e.g., Rada, Mili, Bicknell, & Blettner, 1989; Resnik, 1999).

Although some of these approaches are not explicitly designed as cognitive models, we have proposed that prediction of human similarity can provide a useful point of comparison for computational measures of similarity, noting that one must be aware that such comparisons can be quite sensitive to the specific choice of test items (Resnik, 1999). To date, we are only aware of comparisons having been done using noun similarity.

In this paper, we consider the problem of measuring the semantic similarity of verbs. Verb similarity is in many respects a different problem from noun similarity, because verb representations are generally viewed as possessing properties that nouns do not, such as syntactic subcategorization restrictions, selectional preferences, and event structure, and there are dependencies among these properties.<sup>1</sup> This means that particular

care must be taken in selecting items, as discussed below, and it also means that the same computational measures may be capturing different properties for verbs than for nouns. For example, the IS-A relationship in WordNet's verb taxonomy (Fellbaum, 1998), central in the computation of some measures, signifies generalization according to manner, as in *devour* IS-A *eat*; concomitantly, the verb taxonomy is considerably wider and shallower than WordNet's noun taxonomy. Similarly, measures based on syntactic dependencies may be sensitive to syntactic adjuncts, such as locative and temporal modifiers, that occur predominantly with verbs rather than with nouns.

In what follows, we first discuss several different measures of word similarity and their properties. We then describe an experiment designed to obtain human similarity ratings for pairs of verbs, discuss the fit of the alternative measures to the human ratings, and suggest some implications of these results for future work.

## Models of Verb Similarity

We consider three classes of similarity measure, corresponding to three kinds of lexical representation. In the first, verbs are associated with nodes in a semantic network. In the second, verbs are represented by distributional syntactic co-occurrence features obtained via analysis of a corpus. In the third, verbs are associated with lexical entries represented according to a theory of lexical conceptual structure. These classes of representation can be viewed as occupying three different points on the spectrum from non-syntactic to syntactically relevant facets of verb meaning.

## Taxonomic Models

Taxonomic models of lexical and conceptual knowledge have a long history. In this work we use WordNet version 1.5, a large scale taxonomic representation of concepts lexicalized in English. As a model of the lexicon, WordNet's verb hierarchy is limited by design to paradigmatic relations, in explicit contrast to attempts to organize semantically coherent verb classes through shared syntactic behavior.

The simplest and most traditional measure of semantic similarity in a taxonomy counts the number of edges in-

---

be part-of-speech *per se*; one could argue that some nouns carry similar kinds of participant information, observing, for example, that *x's gift of y to z* parallels *x gave y to z*. We are not attempting to address that issue here.

---

<sup>1</sup>Admittedly, the relevant contrast may turn out not to

tervening between nodes (“edge counting”). A distance in edges is converted to similarity by subtracting from the maximum possible distance in the taxonomy, giving the following measure of distance between verbs  $w_1$  and  $w_2$ :

$$\text{wsim}_{\text{edge}}(w_1, w_2) = (2 \times \text{MAX}) - \left[ \min_{c_1, c_2} \text{len}(c_1, c_2) \right] \quad (1)$$

where  $c_1$  ranges over  $s(w_1)$ ,  $c_2$  ranges over  $s(w_2)$ , MAX is the maximum depth of the taxonomy, and  $\text{len}(c_1, c_2)$  is the length of the shortest path from  $c_1$  to  $c_2$ , with  $s(w)$  denoting the set of concepts in the taxonomy that represent senses of word  $w$ . If all senses of  $w_1$  and  $w_2$  are in separate sub-taxonomies of the WordNet verb hierarchy their edge-count similarity is defined to be zero.

The simple edge-counting approach has well known problems, and arguments have been made for the following measure of semantic similarity between concepts in a taxonomy based on shared information content (Resnik, 1999):

$$\text{sim}_{\text{info1}}(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log p(c)], \quad (2)$$

where  $S(c_1, c_2)$  is the set of concepts that subsume both  $c_1$  and  $c_2$ , and  $-\log p(c)$  quantifies the “information content” of node  $c$ . This yields a measure of verb similarity

$$\text{wsim}_{\text{info1}}(w_1, w_2) = \max_{c_1, c_2} [\text{sim}_{\text{info1}}(c_1, c_2)], \quad (3)$$

where  $c_1$  ranges over  $s(w_1)$  and  $c_2$  ranges over  $s(w_2)$ , and  $p(c)$  is estimated by observing frequencies in a corpus.<sup>2</sup> Intuitively, the quantity defined in (3) measures the maximum overlap in information between the words being compared. When two words are not very similar, the information content of their most informative subsumer (the node  $c$  maximizing  $-\log p(c)$ ) is low: that subsumer resides high in the taxonomy and thus has high probability, implying low information content. In the most extreme case, the most informative subsumer is just the TOP node of the taxonomy, in which case the probability is 1 and the shared information content (and hence similarity) is 0. When two words are similar, that means there is a node lower in the taxonomy that subsumes them both; being lower in the taxonomy its probability is lower and therefore its information content is higher. Crucially, structural notions such as “lower” and “higher”, and the number of intervening arcs between nodes, play no actual role in this model of similarity. As a result, unlike edge counting, this measure does not fall prey to the rampant variation in density within any realistic conceptual taxonomy, where a single IS-A link could represent a tiny semantic distance (e.g. *ballpoint-pen* IS-A *pen*) or a very large semantic distance (e.g. *toy* IS-A *artifact*).<sup>3</sup>

Lin (1998) argues for an alternative information-based measure of similarity that, when applied to a taxonomy,

<sup>2</sup>For taxonomic measures described in this section, probabilities of nodes in WordNet 1.5 were estimated on the basis of word frequencies in the Brown Corpus (Francis & Kučera, 1982).

<sup>3</sup>Examples are from WordNet 1.5, where *artifact* signifies a man-made object.

closely resembles the measure just described. It differs in normalizing the shared information content using the sum of the *unshared* information content of each item being compared:

$$\text{sim}_{\text{info2}}(c_1, c_2) = \frac{2 \times \log p(\bigcap_i C_i)}{\log p(c_1) + \log p(c_2)} \quad (4)$$

where the  $C_i$  are the “maximally specific superclasses” of both  $c_1$  and  $c_2$ . As a result of this normalization, the measure possesses some desirable properties, such as a fixed range from 0 to 1. Word similarity  $\text{wsim}_{\text{info2}}$  is defined analogously to Definition (3).

## Distributional Co-Occurrence Model

Information-based measures of similarity can be applied to representations other than taxonomic structures. Indeed, Lin demonstrates the generality of the idea by showing how such a measure can be used to measure not only taxonomic distance but also string similarity and the distance between feature sets *à la* Tversky. The latter approach is illustrated by representing words as collections of syntactic co-occurrence features obtained by parsing a corpus. For example, both the noun *duty* and the noun *sanction* would have feature sets containing the feature *subj-of(include)*, but only *sanction* would have the feature *adj-mod(economic)*, since “economic sanctions” appears in the corpus but “economic duties” does not. Because these features include both labeled syntactic relationships and the lexical items filling argument roles, the underlying representational model can be thought of as capturing both syntactic and semantic components of verb meaning.

Lin computes the quantity of shared information as the information in the intersection of the distributional feature sets for the two items being compared. This yields the following measure:

$$\text{wsim}_{\text{distrib}}(w_1, w_2) = \frac{2 \times I(F(w_1) \cap F(w_2))}{I(F(w_1)) + I(F(w_2))} \quad (5)$$

where  $F(w_i)$  is the feature set associated with word  $w_i$ , and where  $I(\mathcal{S})$ , the quantity of information in a feature set  $\mathcal{S}$ , is computed as  $I(\mathcal{S}) = -\sum_{f \in \mathcal{S}} \log p(f)$ .<sup>4</sup> In the experiments described here, we use similarity values obtained for verb pairs using Lin’s implementation of his model, with his feature sets and probabilities obtained via analysis of a 22-million-word corpus of newswire text.

## Semantic Structure Model

Our third method for assessing the semantic similarity of verbs relies on elaborated representations of verb semantics according to the theory of lexical conceptual structure, or LCS (Dorr, 1993; Jackendoff, 1983). LCS representations make an explicit distinction between *semantic structure*, which characterizes the grammatically relevant facets of verb meaning, from *semantic content*, which characterizes idiosyncratic information associated with the verb but not reflected in its syntactic behavior.

<sup>4</sup>Note the assumption that features are independent, permitting the summation of log probabilities.

This difference between semantic structure and semantic content plays an important role in current research on lexical representation (e.g. Grimshaw, 1993; Pinker, 1989; Rappaport, Laughren, & Levin, 1993). We take advantage of this distinction here to derive a measure that focuses exclusively on similarity of semantic structure as disentangled from semantic content.

To illustrate with a simple example, within an LCS representational system *roll* and *slide* might both have semantic structure indicating a change of location, e.g.,

$$\begin{aligned} & (g_{\text{loc}} \text{ x} \\ & \quad (\text{to}_{\text{loc}} \text{ x} (\text{at}_{\text{loc}} \text{ x} y)) \\ & \quad (\text{from}_{\text{loc}} \text{ x} (\text{at}_{\text{loc}} \text{ x} z)) \\ & \quad (\text{manner} \langle \text{M} \rangle)), \end{aligned}$$

and differ only in the value  $\langle \text{M} \rangle$  — an element of semantic content within the semantic structure — indicating the manner of motion (either  $\langle \text{SLIDING} \rangle$  or  $\langle \text{ROLLING} \rangle$ ). Such regularities in semantic structure are argued to provide an explanation for systematic relationships between meaning and syntactic realization (Levin & Rappaport Hovav, 1998).

If those regularities are a part of verb lexical representations, then they also plausibly influence ratings of verb similarity, and the question is how to assess similarity between two such structured representations. Lin’s work provides one plausible answer: decomposing complex representations into (pseudo-)independent feature sets and then comparing feature sets.<sup>5</sup> Our method of decomposition was particularly simple, recursively creating an independent feature from each primitive component of the representation and the “head” of its subordinates. So, for example, the feature set representation of *roll* would contain six features:

$$\begin{aligned} & [g_{\text{loc}} \text{ to}_{\text{loc}} \text{ from}_{\text{loc}} \text{ manner}] \\ & [\text{to}_{\text{loc}} \text{ x} \text{ at}_{\text{loc}}] \\ & [\text{at}_{\text{loc}} \text{ x} y] \\ & [\text{from}_{\text{loc}} \text{ x} \text{ at}_{\text{loc}}] \\ & [\text{at}_{\text{loc}} \text{ x} z] \\ & [\text{manner} \langle \text{ROLLING} \rangle]. \end{aligned}$$

The features of *slide* would be identical but for the last feature, which would instead be  $[\text{manner} \langle \text{SLIDING} \rangle]$ , and the nearly complete overlap between the feature sets for the two verbs captures the fact that the semantic distinction between this particular pair of verbs rests entirely on semantic content and not semantic structure.

Since we had available to us a large lexicon of LCS representations for verbs in English (Dorr & Olsen, 1996, 1997), containing thousands of lexical entries, we estimated the probability of each feature by counting feature occurrences within the lexicon. We define the similarity of two LCS lexicon entries  $e_1$  and  $e_2$  using the shared information content of their feature sets:

$$\text{sim}_{\text{LCS}}(e_1, e_2) = I(F(e_1) \cap F(e_2)) \quad (6)$$

<sup>5</sup>We are grateful to Dekang Lin for suggesting this approach to us.

using  $I(S)$  as in (5), and we compute  $\text{wsim}_{\text{LCS}}(w_1, w_2)$  as the maximum value of  $\text{sim}_{\text{LCS}}$  taken over the cross product of all the words’ lexical entries.<sup>6</sup>

It is worth emphasizing that this similarity measure considers *only* semantic structure, not semantic content, and therefore only syntactically relevant components of meaning enter into the computation. For example, in the comparison of LCS entries for *slide* and *roll*,  $F(e_1) \cap F(e_2)$  will never contain either  $[\text{manner} \langle \text{ROLLING} \rangle]$  or  $[\text{manner} \langle \text{SLIDING} \rangle]$ , and therefore any potential similarities or differences between the content elements — the *physical* aspects of sliding motion versus rolling motion based on real-world knowledge — are excluded from the model.

## Experiment

In order to assess alternative computational models of similarity, we collected human ratings of similarity for pairs of verbs, following a design after that of Miller and Charles (1991). Considering the additional complexities in the verb lexicon, however, the selection of materials required considerable care: we were careful to pay close attention to syntactic subcategorization, thematic grids, and aspectual class information, as described below, in order to limit the possible dimensions across which the two verbs in a pair could differ and to focus on *semantic* similarity. We also designed two versions of the task, with and without presentation of verbs in context, in order to investigate the extent to which contextual narrowing of verbs’ senses affects ratings of similarity.

**Participants.** Participants were 10 volunteers, all native speakers of English, ranging in age from 24 to 53, without significant background in psychology or linguistics. All participated by e-mail.

**Materials.** In constructing the set of verb pairs for similarity ratings, we began with the set of verbs in a large lexicon of LCS entries, containing entries for 4900 verbs. Verb entries in the lexicon contain information about both aspectual features (dynamicity, durativity, telicity; Olsen, 1997) and thematic grid (identifying whether or not a verb takes an agent, theme, goal, etc.) — for example, the verb *broil* requires both an agent and a theme, and is marked as both durative and telic but not dynamic. For subcategorization information, we referred to the Collins Cobuild dictionary (Sinclair, 1995), using the subcategorization frame for the first listed verb sense.

To construct verb pairs, we began by eliminating all verbs whose thematic grid did not require a theme, in order to limit the range of variation in thematic grids.<sup>7</sup>

<sup>6</sup>Although our probability estimate counts features within a set of types (entries in a large lexicon) rather than tokens (verb instances in a large corpus), inspection of the estimated probabilities suggests that frequent features are suitably discounted, having low information content, and rare features are highly informative. Corpus-based estimates are a matter for future work.

<sup>7</sup>All verbs require an agent, so the remaining variation is in the presence or absence of oblique roles such as GOAL.



We then grouped the full set of verbs into eight lists corresponding to the eight possible combinations of the three aspectual features, and restricted our attention to the four most numerous lists.<sup>8</sup> Within each of those four lists, we created 12 pairs of verbs subject to the constraint that the verbs’ associated subcategorization frames had to match, so as to avoid effects of purely syntactic similarity. Items were selected to span the range from low- to high-similarity verb pairs.

In summary, a set of 48 verb pairs was constructed so that (i) both verbs in every pair require a theme, (ii) both verbs have the same subcategorization frame, and (iii) both verbs come from the same aspectual class. Verbs on the list were all given in the past tense. In order to avoid ordering effects, half the subjects in each condition saw items in a random order, and the other half saw the items in the reverse order.

To assess the effects that contextual narrowing of verb senses might have on similarity ratings, the materials as just described were duplicated in order to create *No Context* and *Context* conditions. The conditions were identical except that in the *Context* condition, each item was accompanied by an example sentence for each verb illustrating the verb’s intended sense. Each example sentence came from the corresponding verb entry in the Collins Cobuild dictionary. For example, the example sentence for *loosen* was “He loosened his seat belt.”

**Procedure.** The 10 subjects were split evenly into *Context* and *No Context* groups. Subjects in the *No Context* group were given the set of 48 verb pairs, without example sentences, and asked to compare their meanings on a scale of 0–5, where 0 means that the verbs are not similar at all and 5 indicates maximum similarity. Subjects were explicitly asked to ignore similarities in the sound of the verb and similarities in the number and type of letters that make up the verb. Subjects were also asked explicitly to rate similarity rather than relatedness, with the instructions giving an example of the distinction. (For example, *pay* and *eat* are related in that they are things we do in restaurants, but they are not particularly similar.) Since some verbs in the set have low frequency, a “don’t know” box was included for subjects to mark if they were unsure of the meaning of either verb. There was no time limit on the task, which tended to take approximately 20 minutes.

Subjects in the *Context* group were given exactly the same task, but using the *Context* materials, i.e. with each verb accompanied by an example sentence illustrating the intended sense. As in the previous condition, two orders of presentation were used within this condition to avoid ordering effects.

Each computational similarity measure took the set of verb pairs as input, without context, and computed a similarity score for each.

<sup>8</sup>These were {durative}, {durative,dynamic}, {dynamic,telic}, {durative,dynamic,telic}. Verbs could and did appear on multiple lists.

Table 1: Comparing sets of ratings

wsim	<i>Context</i>	<i>No Context</i>
edge	.720	.675
info1	.779	.658
info2	.768	.668
distrib	.453	.433
lcs	.313	.385
Combined	.872	.785
Inter-rater	.793	.764

**Results and Discussion.** In order to judge the degree to which sets of similarity ratings are predictive of each other, we use a similarity coefficient computed as Pearson’s *r*. Table 1 provides a summary showing *r* for each computational model as compared to the mean of the human subject ratings in the *Context* and *No Context* conditions.<sup>9</sup>

The *Combined* row of the table shows the value of multiple *R* when the five computational measures are compared with human ratings using a multiple regression (see below), and the *Inter-rater* row of the table shows human average inter-rater agreement, measured by *r*, using leave-one-out resampling (Weiss & Kulikowski, 1991).

Examining these figures, we first consider each computational model separately. It is unsurprising that the similarity measure based on LCS representations fares worst, given the design of the experiment: the verb pairs were selected so as to eliminate differences of subcategorization frame, aspectual class, and thematic grid, ruling out *a priori* pairs that differ interestingly with respect to semantic structure. The distributional measure based on syntactic co-occurrence features may be a victim of its dependence on a particular corpus, and of data sparseness — for example, glaring divergences with human ratings include some verb pairs containing some lower-frequency words, such as *embellish/decorate* and *dissolve/dissipate*. Turning to the taxonomic methods, the information-based approaches appear superior to edge counting in the *Context* condition, consistent with previous work on noun similarity, though in the *No Context* condition there are no clear differences. We suspect a difference will emerge with a larger set of items, but this remains to be seen. Our inspection of by-item

<sup>9</sup>From the full set of items, 10 verb pairs were excluded because some participant did not know the meaning of one or the other verb. Moreover, in preparation of the final version of this paper, we discovered that 11 verb pairs inadvertently had been included despite failing to strictly match the criteria described in the Materials section or having other minor errors of presentation, and these are now excluded, as well. Although this is a large number of excluded items, we consider them quite unlikely to have affected participants’ judgments since the excluded pairs were distributed almost perfectly evenly over the four verb lists and varied across degrees of similarity, and since the pattern of results was unaffected. We report all quantitative results in the paper based on only the 27 non-excluded verb pairs.

ratings of the information measures suggests strongly that the differences between the unnormalized and normalized information-based measures are small in comparison to the role played by the structure of the WordNet verb taxonomy.

Comparison of human raters yields several interesting observations. First, a comparison of the *Context* and *No Context* mean ratings by human participants yields  $r = .89$ , which provides some reassurance that subjects in the *No Context* condition are generally interpreting the verbs in the same sense as are subjects in the *Context* condition — where, recall, the context sentence encouraged interpretation according to the first listed verb sense in the Collins Cobuild dictionary. Second, however, average inter-rater agreement in the two conditions (.79 and .76) is much lower than that obtained in a noun ratings experiment using the same method, where leave-one-out resampling yielded an estimate of  $r = .90$  (Resnik, 1999). This may reflect the small sample size in each group ( $N = 5$ ), but we suspect that in actuality it is evidence that word similarity is harder for subjects to quantify for verbs than for nouns. Third, we find that subjects in the *No Context* condition have a very strong tendency to assign higher similarity ratings to the same pair as compared to subjects in the *Context* condition, as determined using a paired  $t$ -test ( $N = 27, t(26) = 4.49, p < .0002$ ).

This last observation is consistent with the idea that subjects in the *No Context* condition are accommodating verb comparisons — allowing for more flexible interpretations of verb meaning — in a way not available to subjects in the *Context* condition because their interpretations are constrained by the context sentence. For example, the verb pair *compose/manufacture* has a mean rating of 2.8 in the *Context* condition, and the context sentences are *He sees the whole, not the various lines that compose it* and *Many factories were manufacturing desk calculators*. In the *No Context* condition, the mean rating for this pair is 4.0, likely indicating that in the process of comparison, subjects focused on available semantic elements of *compose*'s meaning that are closest to *manufacture* (e.g., the notion of composing as creating, *She composed satirical poems for the New Statesman*).

As a preliminary step toward combining models, we performed a multiple regression predicting human ratings using the ratings of the five computational models as independent variables, with the results shown in Table 1 as *Combined*. Although we have not extensively analyzed these data, regressions using all  $2^5 - 1 = 31$  combinations of models show that the highest multiple  $R$  is obtained when all five models are combined, that the two different information-based measures are making essentially the same contribution to the combined model (consistent with our observation that WordNet structure plays the dominant role, rather than details of the measure), and that the LCS measure contributes little for this set of items. Taking these observations into account, the improvement in predictive power when combining models comes from distributional and information-based

models being sensitive to at least some different information.

## General Discussion

The experimental results reflect the fact that similarity measures model different aspects of verb representation and use. Taxonomic similarity measures place little emphasis on verbs' argument structure, emphasizing relationships of semantic content; for example, *drag* and *tug* appear quite close in the taxonomy (under *displace*) although they differ significantly in semantic structure (e.g. in “the tailpipe dragged” and “the donkey tugged” the syntactic subjects have different thematic roles). Conversely, semantic structure is emphasized in the measure based on LCS representations to the exclusion of real-world knowledge, such as the similarity of the physical motions of dragging and tugging. Distributional similarity based on syntactic co-occurrence features is a combination, capturing elements of semantic structure by means of the syntactic relationships (one-versus two-participant relationships), and also indirectly capturing elements of semantic content by means of the lexical items co-occurring in those syntactic positions (*tug* being weighted more heavily against inanimate subjects than *drag*, for example). Based on the performance of the models, and improved predictive power of the multiple regression, we interpret our results as evidence that human ratings of similarity are sensitive to both paradigmatic and syntagmatic facets of verb representation, and we believe the computational models are capturing relevant aspects of verb representation in order to make predictions about similarity judgments.

On a somewhat speculative note, it is interesting to briefly examine cases where the computational models fail to capture similarities identified by the human raters. Consider, for example, items *unfold/divorce*, *chill/toughen*, *initiate/enter*. Based on the WordNet taxonomy, the verbs in these pairs have no common subsumer, so the shared information content is zero; nor do the distributional or LCS measures predict that they are at all similar. The human mean ratings are low (averaging 1.6, 1.4, and 3.2, respectively, in the *No Context* condition), but why are they not zero — and why are they in fact higher than the ratings for some other pairs, such as *open/inflate* (0.6), where one could also identify reasons for believing the meanings have something in common? It would appear that in these cases subjects are finding similarities of meaning according to dimensions that we have not yet formalized. The apparent sense extensions verge on the metaphorical: one can describe divorce as the unfolding of a marriage, observe a person chill and toughen in response to an insult, enter a group by being initiated into it. Capturing those dimensions of similarity in our models will require a better understanding than we have at present of how word meanings are represented and organized.

Even for the time being, however, the work described in this paper offers a method and a testbed for investigating lexical issues that can go well beyond the present experiments. We chose here to tightly control aspect and

syntactic subcategorization while allowing our test items to differ on thematic grids and vary widely with respect to semantic content. Having validated the approach — performance being consistent with what one would predict of the alternative models given the design of the task — the initial work opens the door to other configurations, controlling variation among subcategorization frames, aspectual features, thematic grids, and semantic content in other combinations. What is crucial is that implemented models of similarity, drawing on such theoretical constructs, yield testable predictions that can be verified through careful experimentation.

### Acknowledgments

We are grateful to Dekang Lin and Amy Weinberg for valuable discussions, to Dekang Lin for his kindly computing values of distributional similarity (Definition 4) for the verb pairs in our experiment, and to three anonymous reviewers for their helpful comments. This work was supported in part by DARPA/ITO Contract N66001-97-C-8540.

### Appendix: Verb Pairs

bathe	kneel	loosen	open
chill	toughen	neutralize	energize
compose	manufacture	obsess	disillusion
compress	unionize	open	inflate
crinkle	boggle	percolate	unionize
displease	disillusion	plunge	bathe
dissolve	dissipate	prick	compose
embellish	decorate	swagger	waddle
festoon	decorate	unfold	divorce
fill	inject	wash	sap
hack	unfold	weave	enrich
initiate	enter	whisk	deflate
lean	kneel	wiggle	rotate
loosen	inflate		

### References

- Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4), 467–480.
- Dorr, B. J. (1993). *Machine Translation: A View from the Lexicon*. The MIT Press, Cambridge, MA.
- Dorr, B. J., & Olsen, M. B. (1996). Multilingual Generation: The Role of Telicity in Lexical Choice and Syntactic Realization. *Machine Translation*, 11(1–3), 37–74.
- Dorr, B. J., & Olsen, M. B. (1997). Deriving Verbal and Compositional Lexical Aspect for NLP Applications. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pp. 151–158 Madrid, Spain.
- Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Lexical Database*. MIT Press.
- Francis, W. N., & Kučera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston.
- Grimshaw, J. (1993). Semantic structure and semantic content in lexical representation. unpublished manuscript, Center for Cognitive Science, Rutgers University, New Brunswick, New Jersey.
- Jackendoff, R. (1983). *Semantics and Cognition*. The MIT Press, Cambridge, MA.
- Levin, B., & Rappaport Hovav, M. (1998). Building Verb Meanings. In Butt, M., & Geuder, W. (Eds.), *The Projection of Arguments: Lexical and Compositional Factors*, pp. 97–134. CSLI Publications, Stanford, CA.
- Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML-98)* Madison, Wisconsin.
- Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1–28.
- Olsen, M. B. (1997). *A Semantic and Pragmatic Model of Lexical and Grammatical Aspect*. Garland, New York.
- Pinker, S. (1989). *Learnability and Cognition*. MIT Press, Cambridge, MA.
- Quillian, M. R. (1968). Semantic memory. In Minsky, M. (Ed.), *Semantic Information Processing*. MIT Press, Cambridge, MA.
- Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics*, 19(1), 17–30.
- Rappaport, M., Laughren, M., & Levin, B. (1993). Levels of lexical representation. In Pustejovsky, J. (Ed.), *Semantics and the Lexicon*. Kluwer.
- Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research (JAIR)*, 11, 95–130. <http://www.cs.washington.edu/research/jair/abstracts/resnik99a.html>
- Schütze, H. (1993). Word space. In Hanson, S. J., Cowan, J. D., & Giles, C. L. (Eds.), *Advances in Neural Information Processing Systems 5*, pp. 895–902. Morgan Kaufmann Publishers, San Mateo CA.
- Sinclair, J. (Ed.). (1995). *Collins Cobuild English Dictionary*. Collins. Patrick Hanks, managing editor.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352.
- Weiss, S. M., & Kulikowski, C. A. (1991). *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann, San Mateo, CA.

# The Advantages and Disadvantages of Semantic Ambiguity

Jennifer Rodd (jenni.rodd@mrc-cbu.cam.ac.uk)

MRC Cognition and Brain Sciences Unit  
15 Chaucer Road, Cambridge, UK

Gareth Gaskell (g.gaskell@psych.york.ac.uk)

Department of Psychology  
University of York, York, UK

William Marslen-Wilson (william.marslen-wilson@mrc-cbu.cam.ac.uk)

MRC Cognition and Brain Sciences Unit  
15 Chaucer Road, Cambridge, UK

## Abstract

There have been several reports of faster lexical decisions for words that have many meanings (e.g., *ring*) compared with words with few meanings (e.g., *hotel*). However, it is not clear whether this advantage for ambiguous words arises because they have multiple unrelated meanings, or because they have a large number of highly related word senses. All current accounts of the ambiguity advantage assume that it is unrelated meanings that produce the processing benefit. We report two experiments that challenge this assumption; in visual and auditory lexical decision experiments we found that while multiple senses did produce faster responses, multiple meanings produced a disadvantage. We discuss how models of word recognition could accommodate this new pattern of results.

## Introduction

Many words are semantically ambiguous, and can refer to more than one concept. For example, *bark* can refer either to a part of a tree, or to the sound made by a dog. To understand such words, we must disambiguate between these different interpretations, normally on the basis of the context in which the word occurs. However, ambiguous words can also be recognised in isolation; when presented with a word like *bark* we are able to identify an appropriate meaning rapidly, and are often unaware of any other meanings.

Words can be ambiguous in different ways. The two meanings of a word like *bark* are semantically unrelated, and seem to share the same written and spoken form purely by chance. Other words are ambiguous between highly related senses, which are systematically related to each other. For example, the word *twist* can refer to a bend in a road, an unexpected ending to a story, a type of dance, and other related concepts.

The linguistic literature makes a distinction between these two types of ambiguity, and refers to them as homonymy and polysemy (Lyons, 1977; Cruse, 1986). Homonyms, such as the two meanings of *bark*, are said to be different words that by chance share the same orthographic and phonological form. On the other hand, a polysemous word like *twist* is considered to be a single word that has more than one sense.

All standard dictionaries respect this distinction between word meanings and word senses; lexicographers routinely decide whether different usages of the same spelling should correspond to different lexical entries or different senses within a

single entry. Many criteria (e.g., etymological, semantic and syntactic) have been suggested to operationalise this distinction between senses and meanings. However, it is generally agreed that while the distinction appears easy to formulate, it is difficult, to apply with consistency and reliability. People will often disagree about whether two usages of a word are sufficiently related that they should be taken as senses of a single meaning rather than different meanings. This suggests that these two types of ambiguity may be best viewed as the end points on a continuum. However, even if there is not a clear distinction between these two different types of ambiguity, it is important to remember that words that are described as ambiguous can vary between these two extremes.

In this paper we will review the evidence on how lexical ambiguity affects the recognition of isolated words, and will argue that the distinction between these two qualitatively different types of ambiguity has not been addressed. We then report two experiments that confirm the importance of the sense-meaning distinction, and show that in both the visual and the auditory domains the effects of word meanings and word senses are very different.

## The Ambiguity Advantage

In early studies of semantic ambiguity, Rubenstein, Garfield, and Millikan (1970) and Jastrzembski (1981) reported faster visual lexical decisions for semantically ambiguous words than for unambiguous words. However, these studies did not control for the subjective familiarity of the words, and Gernsbacher (1984) found no effect of ambiguity over and above familiarity. Since then, however, Kellas, Ferraro, and Simpson (1988), Borowsky and Masson (1996) and Azuma and Van Orden (1997) have all reported an ambiguity advantage in visual lexical decision experiments using stimuli that were controlled for familiarity.

Although there does seem to be a consensus in the literature that lexical ambiguity can produce faster lexical decision times, it is not at all clear what type of ambiguity is producing the effect. Is it multiple meanings, or multiple senses that produces the advantage? One way of trying to answer this question is to examine the dictionary entries of the words used in these experiments. As described above, dictionaries make a distinction between words whose meanings are sufficiently unrelated that they are given multiple entries, and those that

have multiple senses within an entry. This provides a convenient way in which to categorise words as being ambiguous between multiple meanings or between multiple senses.

Rodd, Gaskell, and Marslen-Wilson (1999) analyzed the stimuli used in the three studies that report a significant ambiguity advantage in this way, and found that for all three studies the high-ambiguity words have more word senses than the low-ambiguity words. Further, only in the Borowsky and Masson (1996) stimuli did the two groups differ in the number of meanings. Therefore, it appears that it may be multiple senses rather than multiple meanings that are producing the ambiguity advantage. Despite this, all current explanations of the ambiguity advantage assume that the processing benefit arises because of the presence of unrelated meanings.

## Models of the Ambiguity Advantage

One way that the ambiguity advantage has been explained has been to assume that ambiguous words have multiple entries within a lexical network. For example, (Kellas et al., 1988) suggest that the benefit arises because, while the multiple entries for an ambiguous word do not inhibit each other, they both act independently to inhibit all other competing entries, and this increased inhibition of competitors produces the faster recognition times.

Others have assumed that the benefit arises within this type of model by assuming that there is some level of noise or probabilistic activation (Jastrzemski, 1981). Because words with multiple meanings are assumed to have multiple entries, these words might benefit from having more than one competitor in the race for recognition; on average, by a particular point in time, one of these competitors is more likely to have reached the threshold for recognition than a word that has only one entry in the race.

Both these approaches to explaining the ambiguity advantage predict that the effect will occur whenever the different meanings of the ambiguous words are sufficiently unrelated to have separate entries in the mental lexicon; they make no specific predictions about what should happen for words with multiple senses, as it is not clear whether word senses would correspond to separate entries within the network.

An alternative view of word recognition is that words compete to activate a representation of their meaning. There have been several recent models of both spoken and visual word recognition that have taken this approach (Hinton & Shallice, 1991; Plaut & Shallice, 1993; Joordens & Besner, 1994; Gaskell & Marslen-Wilson, 1997; Plaut, 1997). These models use distributed lexical representations; each word is represented as a unique pattern of activation across a set of orthographic/phonological and semantic units.

Within models of this type, the orthographic pattern *bark* must be associated with two different semantic patterns corresponding to its two meanings. When the orthographic pattern is presented to the network, the network will try to instantiate the word's two meanings across the same set of semantic units simultaneously. These competing semantic representations will interfere with each other, and this interference is likely to increase the time it takes for a stable pattern of activation to be produced. Therefore, it appears that these models predict that lexical ambiguity should delay recognition, and not produce the faster response times seen in the literature.

In response to this inconsistency between the ambiguity advantage literature and the predictions of semantic competition models, there have been several attempts to show that, given particular assumptions, this class of model can overcome the semantic competition effect, and show an advantage for ambiguous words (e.g. Joordens and Besner (1994), Borowsky and Masson (1996) and Kawamoto, Farrar, and Kello (1994)). Importantly, these models assume that the effect to be modelled is an advantage for those words with multiple unrelated meanings.

Thus, the ambiguity advantage has been interpreted within a range of models of word recognition. However, all these accounts have implicitly assumed that the ambiguity advantage literature demonstrates that there is a processing advantage for words with more than one, unrelated, meaning. As discussed above, it is not clear that this is the case; the ambiguity advantage may be a benefit for words with multiple *senses* rather than multiple *meanings*. In order to understand fully the implications of semantic ambiguity for models of word recognition, we need to determine which of these explanations is correct.

## Experiment 1: Visual Lexical Decision

### Method

**Experimental Design** This experiment attempts to separate out the effects of lexical ambiguity and multiple word senses by using a factorial design (see Table 1). Groups of ambiguous and unambiguous words were selected to have either few or many senses on the basis of their dictionary entries.

Table 1: Experiment 1: Experimental Design

Ambiguity	Senses	Example
Ambiguous	Few	pupil
Ambiguous	Many	slip
Unambiguous	Few	cage
Unambiguous	Many	mask

**Participants** The participants were 25 members of the MRC Cognition and Brain Sciences Unit subject panel. All had English as their first language, and had normal or corrected-to-normal vision.

**Stimuli** The word stimuli were selected to conform to a 2 x 2 factorial design, where the two factors were ambiguity and number of senses. Words were classed as being unambiguous if they had only one entry in The Online Wordsmyth English Dictionary-Thesaurus (Parks, Ray, & Bland, 1998), and as ambiguous if they had two or more entries. Two measures of the number of senses were used. These were the total number of word senses listed in the Wordsmyth dictionary for all the entries for that word, and the total number of senses given in the WordNet lexical database (Fellbaum, 1998).

Thirty-two stimuli were selected to fill each cell of the factorial design, such that the number of word meanings was matched across each level of number of word senses, and the total number of word senses was matched across each level of the number of word meanings.

The four groups of words were matched for frequency in the CELEX lexical database (Baayen, Piepenbrock, &

Van Rijn, 1993), number of letters, number of syllables, concreteness and familiarity. Concreteness and familiarity scores were obtained from rating pre-tests in which all the words were rated on a 7-point scale by participants who were members of the MRC Cognition and Brain Sciences Unit subject panel, and who did not participate in the lexical decision experiment.

The groups were not explicitly matched for neighbourhood density; however, the number of words in CELEX that differed from each word by only one letter ( $N$ ; Coltheart, Davelaar, Jonasson, & Besner, 1977) was calculated for each word. An analysis of variance (ANOVA) showed that the words in the four groups did not differ significantly on this measure;  $F(3, 124) = 1.02, p > .3$ .

The non-word distractors were pseudohomophones, such as *brane*, with a similar distribution of word lengths to the word stimuli. Pseudohomophones were used because both (Azuma & Van Orden, 1997) and (Pexman & Lupker, 1999) found stronger effects of semantic ambiguity when these non-words were used. In this first experiment, we wanted to maximise the chance of finding significant effects of ambiguity.

**Procedure** All the stimulus items were pseudo-randomly divided into four lists, such that each list contained approximately the same number of words from each stimulus group. Participants were presented with the four lists in a random order, with a short break between lists. Within the lists, the order in which stimulus items were presented was randomised for each participant. All participants saw all of the stimulus materials. A practice session, consisting of 64 items not used in the analysis, was given to familiarise participants with the task. Each block began with 10 stimuli not included in the analysis.

For each of the word and non-word stimuli, the participants were presented with a fixation point in the centre of a computer screen for 500 msec, followed by the stimulus item. Their task was to decide whether each item was a word or a non-word; recognition was signalled with the dominant hand, non-recognition with the other hand. As soon as the participant responded, the word was replaced with a new fixation point.

## Results

The data from two participants were removed from the analysis, because of error rates greater than 10%. The latencies for responses to the word and non-word stimuli were recorded, and the inverse of these response times ( $1/RT$ ) were used in the analyses to minimize the effect of outliers (Ulrich & Miller, 1994; Ratcliff, 1993). Incorrect responses were not included in the analysis. The overall error rate for responses was 3.6%.

Mean values were calculated separately across participants and items. The participant means were subjected to an ANOVA, and the item means were subjected to an analysis of covariance (ANCOVA) with frequency, familiarity, concreteness and length entered as covariates. The mean response times are given in Figure 1.

The ANCOVA revealed significant effects of frequency, familiarity, length and neighbourhood density (all  $p < .05$ ). The effect of concreteness was non-significant ( $p > .5$ ), so this variable was removed from the ANCOVA. The response

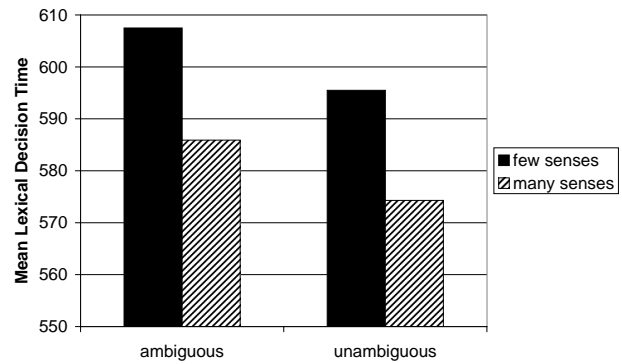


Figure 1: Experiment 1, mean lexical decision times

time data revealed a main effect of the number of senses ( $F_1(1, 22) = 14.22, p < .001$ ;  $F_2(1, 120) = 4.51, p < .05$ ). Words with many senses were responded to faster than words with few senses. The effect of ambiguity was marginal in the participants analysis ( $F_1(1, 22) = 3.77, p < .07$ ), but non-significant in the items analysis ( $F_2(1, 120) = 1.67, p > .2$ ). Ambiguous words were responded to more slowly than unambiguous words. There was no significant interaction between these two variables ( $p > .2$ ).

The error data also showed a significant effect of the number of senses; fewer errors were made for words with many senses ( $F_1(1, 22) = 12.2, p < .005$ ;  $F_2(1, 120) = 5.19, p < .05$ ). In the error data neither the effect of ambiguity nor the interaction between the two variables reached significance (all  $p > .4$ ).

## Discussion

This experiment shows that words with many senses were responded to faster and with fewer errors than words with few senses. This advantage for multiple senses is in contrast with a *disadvantage* for words with multiple meanings. Although this disadvantage was not significant, it is clear that contrary to the accepted view in the literature, there is no processing advantage for words with multiple meanings. Moreover, Rodd et al. (1999) did find a significant disadvantage in visual lexical decision for words with more than one meaning, compared with unambiguous words, when the stimuli were selected to minimise the effect of word senses. Thus, previous reports of an ambiguity advantage must be the result of the multiple senses of the high-ambiguity stimuli rather than their multiple meanings.

Therefore, the results of this experiment together with the results of Rodd et al. (1999) show that the two types of lexical ambiguity have opposite effects on visual word recognition; while ambiguity between multiple meanings may delay recognition, ambiguity between multiple senses is beneficial.

The following experiment will investigate whether this pattern is also seen in the auditory domain. If the above pattern of data is telling us something interesting about the way in which word meanings are stored and processed, we should expect to find the same pattern independent of the input modality.

This experiment will also allow us to establish that these

effects of semantic ambiguity are not contingent on the type of non-word distractors used. In Experiment 1, pseudohomophones such as *brane* were used. There is still debate about how pseudohomophones affect lexical processing (see Pexman & Lupker, 1999 for a review). One possibility is that they simply increase the difficulty of the task, and so increase sensitivity to relatively small effects. However, an alternative explanation is that pseudohomophones strategically effect the way that participants make use of orthographic, phonological and semantic information. The following experiment, which does not use pseudohomophones will attempt to demonstrate that these effects are not due to strategic effects induced by these particular non-words.

Finally, this experiment will also allow us to try and replicate the significant ambiguity disadvantage seen by Rodd et al. (1999).

## Experiment 2: Auditory Lexical Decision

### Method

**Experimental Design** The experimental design was identical to Experiment 1.

**Participants** The participants were 26 students at Cambridge University who had not participated in the first experiment. All had English as their first language, and had normal or corrected-to-normal vision.

**Stimuli** 23 stimuli were selected to fill each cell of the factorial design, such that the number of word meanings was matched across each level of number of word senses. The words were selected on the basis of dictionary entries as in Experiment 1. The number of words in each cell is smaller than was used in Experiment 1, because of the additional constraints used to match the groups. 77% of the words were also used in Experiment 1.

The four groups of words were matched for frequency, number of phonemes, the phoneme at which the word becomes unique, actual length of the words in msec, concreteness and familiarity. Concreteness and familiarity scores were obtained from the same rating pre-test as in Experiment 1. All the words had only one syllable.

The non-word stimuli were created to be as word-like as possible, and to have a similar distribution of word lengths to the word stimuli.

**Procedure** The procedure used was the same as that in Experiment 1, except that now the stimuli were spoken words. Each item appeared 1000 ms after the participants' response to the preceding item. If the participant did not respond within 3000 ms of the onset of a word, the next item was presented.

### Results

The data from four participants were removed from the analysis, because of error rates greater than 10%. Incorrect responses were not included in the analysis. The overall error rate for responses was 5.8%.

As in Experiment 1, inverse response times were used in all analyses. Mean values were calculated separately across participants and items. The participant means were subjected to an analysis of variance (ANOVA), and the item means were

subjected to an analysis of covariance (ANCOVA), with familiarity and length entered as covariates. The mean response times are given in Figure 2.

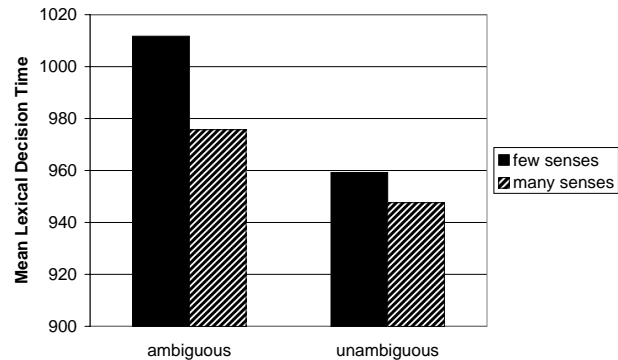


Figure 2: Experiment 2, mean lexical decision times

The ANCOVA revealed significant effects of familiarity ( $r = -.26, p < 0.05$ ) and length ( $r = -.75, p < 0.001$ ). Concreteness, frequency, number of phonemes and uniqueness point were not significant predictors of response times ( $p > .2$ ), so these variables were not included in the ANCOVA.

The main effect of the number of word senses was significant in both the participants and items analysis ( $F_1(1, 21) = 16.9, p < .001$ ;  $F_2(1, 86) = 4.4, p < .05$ ). Words with many senses were responded to faster than words with few senses. The effect of ambiguity was also significant in both the participants analysis and the items analysis ( $F_1(1, 21) = 27.8, p < .001$ ;  $F_2(1, 86) = 7.4, p < .005$ ). Ambiguous words were responded to more slowly than unambiguous words. The interaction between these two variables was marginal in the subjects analysis but did not approach significance in the items analysis ( $F_1(1, 21) = 3.8, p < .1$ ;  $F_2(1, 86) = 0.4, p > .5$ ).

The error data showed a similar pattern of results to the response time data. Fewer errors were made for words with many senses, although this difference was significant only in the subjects analysis but not in the items analysis; ( $F_1(1, 21) = 10.5, p < .005$ ;  $F_2(1, 86) = 2.7, p = .1$ ). Fewer errors were also made for unambiguous words, although this difference was only marginal in the subjects analysis and did not approach significance in the items analysis; ( $F_1(1, 21) = 4.2, p < .06$ ;  $F_2(1, 86) = 0.7, p > .4$ ). The interaction between the two variables was not significant in either analysis ( $p > .5$ ).

### General Discussion

Both the experiments reported here have shown an advantage for words with many word senses. This advantage for multiple senses was seen alongside a disadvantage for words with multiple meanings. This suggests that the ambiguity advantage reported in earlier studies must have been produced by the high number of related word senses of high-ambiguity stimuli, and not by their unrelated meanings.

What are the implications of this new pattern of results for models of word recognition? Previously, these models had

been required to produce an advantage for words with multiple meanings, but our data suggests they must accommodate exactly the reverse effect. In fact, this is less problematic than might be expected.

The ambiguity disadvantage can easily be explained by models in which words compete for the activation of semantic representations (Hinton & Shallice, 1991; Plaut & Shallice, 1993; Joordens & Besner, 1994; Gaskell & Marslen-Wilson, 1997; Plaut, 1997). As discussed earlier, in these models competition between the different meanings of ambiguous words would delay their recognition. As noted by Joordens and Besner (1994), an ambiguity advantage can only be produced by these models if an additional mechanism is present to overcome this semantic competition. These results suggest that no such mechanism is required.

The other class of model that may be able to accommodate this new pattern of results is those models in which words compete to activate abstract word nodes within a lexical network. Earlier, we discussed how these models could produce an ambiguity advantage by assuming either that ambiguous words are more efficient at inhibiting competitors, or that they benefit from having multiple competitors in the race for recognition.

Surprisingly, these models can just as easily accommodate a disadvantage for words with multiple meanings. As in all experiments of this type, the ambiguous words and unambiguous words in these experiments were matched on total frequency. This means that the frequency of each meaning of the ambiguous words is on average half that of the unambiguous word. This frequency difference could produce faster lexical decisions for the unambiguous words. Similarly, if lateral inhibition were present between all word nodes, including the nodes corresponding to the different meanings of an ambiguous word, this would act to slow the recognition of ambiguous words.

Therefore, it appears that both classes of models considered here can be modified to accommodate the finding of slower responses to words with more than one unrelated meaning. However, Rodd et al. (1999) have shown that at least in the visual domain, the ambiguity disadvantage is modulated by the rated relatedness of the two meanings of the ambiguous words; words whose meanings are sufficiently different to be considered meanings rather than senses but whose meanings are mildly related are responded to more quickly than those whose meanings are highly unrelated. This suggests that semantic representations are actively involved in the process that produces the ambiguity disadvantage, and that the effect cannot be explained solely as the result of a frequency bias for unambiguous words or lateral inhibition between abstract word nodes. Therefore, the ambiguity disadvantage may more easily be explained as the result of semantic competition which is maximal when the competing representations are unrelated.

It is therefore apparently straightforward to explain the observed ambiguity disadvantage. The intriguing question that remains is what causes the advantage for words with many senses?

One possibility is to explain this effect in terms of the attractor basins that develop in a distributed semantic network. The different senses of a word correspond to a set of

highly correlated patterns of semantic activation. As noted by Kawamoto (1993), for a word with many related senses, these senses will create a broad and shallow basin of attraction, containing more than one stable state corresponding to each different sense. It is plausible that within certain architectures, settling into the correct attractor may be quicker for such a broad attractor, compared with the attractor of a word with few senses, or that the multiple stable states within the attractor may lead to faster settling times. This suggestion needs to be assessed by performing the appropriate simulations.

A second possible explanation of the sense effect would be to consider the difference between words with many and few senses as reflecting a difference in the amount of semantic information associated with the two types of words. In other words, a word with many senses may be considered to be semantically rich. This is essentially the same argument that Plaut and Shallice (1993) put forward to account for the processing benefit of concrete words over abstract words. In their computational account of the concreteness effect, the difference between abstract and concrete words is reflected in the number of semantic features in a distributed semantic representation; abstract words are given fewer semantic features than concrete words. This results in concrete words activating more stable representations than abstract words. These stable representations lead in turn to faster settling times for words with more semantic features.

It is not yet possible to distinguish between these (and other) possible explanations of the sense effect reported here. A combination of network simulations and further experiments is required to determine how existing models of word recognition should be modified to accommodate the benefit for words with many word senses. What is clear is that the distinction we have emphasised between word meanings and word senses is critical. In the past, ambiguity has been treated as a unitary property of words; we have shown that this has masked an informative pattern of results that can be used to constrain models of how words are recognised.

More generally, these experiments emphasise how word recognition is inextricably linked with word meanings. Data of this kind places an increasing demand on models of word recognition to incorporate richer semantic representations that reflect the complex structures of the meanings of words.

## References

- Azuma, T., & Van Orden, G. C. (1997). Why safe is better than fast: The relatedness of a word's meanings affects lexical decision times. *Journal of Memory and Language*, *36*, 484–504.
- Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). *The CELEX Lexical Database*. CD-ROM. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.
- Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning Memory and Cognition*, *22*, 63–85.
- Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.



- Cruse, A. D. (1986). *Lexical semantics*. Cambridge, England: Cambridge University Press.
- Fellbaum, C. (Ed.). (1998). *Wordnet: An electronic lexical database*. Cambridge, MA: MIT Press.
- Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*, 613–656.
- Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General, 113*, 254–281.
- Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98*, 74–95.
- Jastrzemski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology, 13*, 278–305.
- Joordens, S., & Besner, D. (1994). When banking on meaning is not (yet) money in the bank - explorations in connectionist modeling. *Journal of Experimental Psychology: Learning Memory and Cognition, 20*, 1051–1062.
- Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: a parallel distributed processing account. *Journal of Memory and Language, 32*, 474–516.
- Kawamoto, A. H., Farrar, W. T., & Kello, C. T. (1994). When two meanings are better than one: Modeling the ambiguity advantage using a recurrent distributed network. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1233–1247.
- Kellas, G., Ferraro, F. R., & Simpson, G. B. (1988). Lexical ambiguity and the timecourse of attentional allocation in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 601–609.
- Lyons, J. (1977). *Semantics*. Cambridge, England: Cambridge University Press.
- Parks, R., Ray, J., & Bland, S. (1998). *Wordsmyth English dictionary-thesaurus*. [ONLINE]. Available: <http://www.wordsmyth.net/> [1999, February 1], University of Chicago.
- Pexman, P. M., & Lupker, S. J. (1999). Ambiguity and visual word recognition: Can feedback explain both homophone and polysemy effects? *Canadian Journal of Experimental Psychology, 323–334*.
- Plaut, D. C. (1997). Structure and function in the lexical system: insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes, 12*, 765–805.
- Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10*, 377–500.
- Ratcliff, R. (1993). Methods for dealing with reaction-time outliers. *Psychological Bulletin, 114*(3), 510–532.
- Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (1999). Semantic competition and the ambiguity disadvantage. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society* (pp. 608–613). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9*, 487–494.
- Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction-time analysis. *Journal of Experimental Psychology-General, 123*(1), 34–80.

# Attention to Action: Securing Task-Relevant Control in Spoken Word Production

Ardi Roelofs (ardi@mpi.nl)

Max Planck Institute for Psycholinguistics  
Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands

## Abstract

The Stroop phenomenon is the finding that color naming is inhibited by incongruent color words but word reading not by incongruent colors. When the stimulus onset asynchrony (SOA) is manipulated, maximal inhibition of incongruent words on color naming is obtained when the words are presented within 100 msec of the colors, whereas facilitation of preexposed congruent words is constant. These findings are obtained both with and without task certainty. Whereas existing models explain the basic Stroop effects, they fail to account for the time course findings and for performance under task uncertainty. In this paper, I extend and apply the WEAVERTM model of spoken word production (Roelofs, 1992, 1993, 1997c; Levelt, Roelofs, & Meyer, 1999) to performance on the Stroop task and show that the model accounts for the key findings.

## Introduction

Performance on the Stroop task is of direct relevance to theories of language production and comprehension. The basic modes of language use, namely speaking, listening, reading, and writing, all seem to make use of overlapping sets of basic processing components (e.g., Caplan, 1992; Levelt, 1989; Shallice, 1988). Whereas language perception occurs automatically, hearing or reading a word does not automatically lead to its production but this is under the control of a language user. Similarly, seeing an object does not automatically lead to the naming of it. Furthermore, words do not occur in isolation, but are typically part of a spoken discourse, a text on a page, or appear on objects in the real world. This points to the need to deal with the issue of selectivity. It is generally assumed that performance on the Stroop task can provide evidence on how language is controlled, that is, how a speaker secures task-relevant control over the basic language processes underlying naming and oral reading (e.g., Allport, 1993).

Since Stroop's (1935) experiments in the 1930s, over 700 articles have appeared using his task (reviewed by MacLeod, 1991), which established the following basic empirical picture. Color naming is inhibited by incongruent color words, but word reading not by incongruent color patches. For example, saying "red" to a red color patch on which the word "green" is superimposed proceeds slower than saying "red" in a control condition consisting of a string of Xs. The stimulus onset asynchrony (SOA) between color patch and word has an important effect. Maximal inhibition of incongruent words on color naming is obtained when the words are presented within 100 msec of the colors (e.g., Glaser & Glaser, 1982). Preexposed congruent words yield facilitation, which is constant over SOAs. Whereas color naming is affected by incongruent words,

reading aloud words (e.g., "red") is not influenced by incongruent colors (e.g., green). The asymmetry in effect is not due to a difference in processing speed between words and colors (i.e., reading is some 200 msec faster than color naming), as evident from manipulating the SOA. When a color patch is presented 300 or 400 msec before the word to be read, still no effect of incongruent colors is obtained (e.g., Glaser & Glaser, 1982).

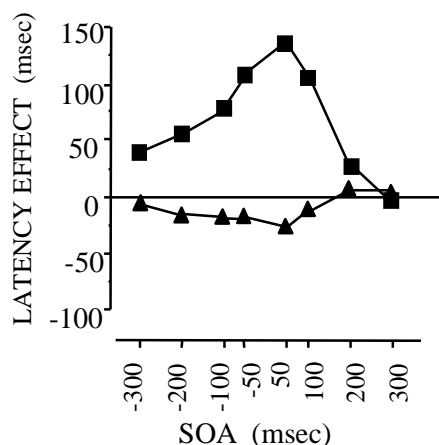


Figure 1: Time course of the Stroop effects (relative to control) in color naming under task uncertainty as measured by Glaser and Glaser (1989): ■ = incongruent, ▲ = congruent

In a standard Stroop experiment, participants are certain about what task to perform. Typically, one group of participants is asked to name the color and to ignore the word, and another group of participants is asked to read aloud the word and to ignore the color patch. Thus, trials are blocked by task. However, in examining the effect of task uncertainty, Glaser and Glaser (1989) asked a single group of participants to perform both tasks. They instructed their participants to respond to the second stimulus component (in the condition with negative SOAs) or the first component (in the condition with positive SOAs). Words had to be read aloud and colors had to be named. Participants can perform this task up to differences in presentation time of 50 msec. The experiment was run with SOAs of -300, -200, -100, -50, 50, 100, 200, and 300 msec (a minus sign indicates preexposure of the irrelevant stimulus). With task uncertainty, Glaser and Glaser obtained the normal patterns of inhibition and facilitation observed with task certainty, for example, with the instruction to name the color and ignore the word. Inhibition increased when the SOA be-

came less negative, peaking at SOAs between -100 and 100 msec. And there was a flat pattern of facilitation at the negative SOAs. Figure 1 shows the SOA curves.

The Stroop phenomenon is not restricted to naming colors and reading color words but appears in many other verbal domains. For example, numerals interfere with the naming of numerosity (e.g., saying “two” to two 6s), but there is no reverse effect (e.g., Flowers, Warner, & Polansky, 1979). Alternating between tasks that exhibit the Stroop conflict (i.e., between color naming and numerosity naming trials) does not yield a greater task switch cost than alternating between tasks that do not yield the Stroop conflict (i.e., between word reading and numeral reading trials), as observed by Allport, Styles, and Hsieh (1994).

## **The Challenge Posed by Task Uncertainty**

### **The Stroop Conflict as Task Conflict**

Recent proposals in the literature (e.g., Rogers & Monsell, 1995) have suggested that the essence of the Stroop conflict is competition between tasks *per se* (i.e., word reading and color naming). The Stroop conflict is explained as inhibition of the “weaker” color naming task by the supposedly “stronger” reading task, while the reverse does not occur. However, task competition fails to explain why the congruent condition (where the same tasks compete) yields facilitation. Furthermore, the fact that task uncertainty has no influence on the SOA patterns poses a challenge. Also, task competition would predict a greater cost for switching between conflict trials with color and numerosity naming than between nonconflict trials with word and numeral reading, but Allport et al. (1994) found no cost difference.

The findings on the time course of the effects and on task uncertainty also pose a challenge to models that do not conceive of the Stroop conflict as a task conflict *per se*. Whereas existing models account for the basic Stroop effects obtained with SOA = 0 msec, they fail to explain the time course findings and they cannot cope with task uncertainty.

### **The Model of Cohen, Dunbar, and McClelland (1990)**

Among the most influential models of the Stroop phenomenon is the connectionist model of Cohen, Dunbar, and McClelland (1990). The model assumes a feedforward network with parallel reading and color naming pathways, which differ in strength. Task relevant control is achieved in the model by task nodes for color naming and reading. These task nodes provide extra input to the color and reading pathway, depending on the task. Each response node in the network is connected with an evidence accumulator. Before the beginning of a simulated Stroop trial, all evidence accumulators are set to zero. A task node is activated and the model is run until the activation of all nodes stabilizes. This allows the system to settle into a “ready state” for the task. Next, the components of a Stroop stimulus are presented with the appropriate SOA. A response is selected when one of the accumulators exceeds a fixed response threshold.

The model of Cohen et al. does well in accounting for the basic Stroop effects obtained with SOA = 0 msec, but there are two major problems. First, as simulations by Cohen et al. (1990, p. 344) showed, the amount of evidence accumulated for the irrelevant stimulus is a positive function of its preexposure time. That is, more evidence is collected at more negative SOAs. Thus, the inhibition in the incongruent condition peaks at the most negative SOA and decreases when the SOA becomes less negative. Similarly, the amount of facilitation in the congruent condition peaks at the most negative SOA and decreases when the SOA becomes less negative. The problem is that these patterns are exactly contrary to the empirical results, where maximal impact of incongruent words is observed when the words appear within 100 msec of the colors and facilitation from preexposed congruent words is constant. The model also predicts a small Stroop effect at negative SOAs in reading aloud, contrary to the real data. The second major problem with the model is that it cannot handle task uncertainty. Before the beginning of a trial, a task node is activated and the model is run until the activation of all nodes stabilizes, which allows the system to settle into a ready state for the task. But with task uncertainty, the task is not known beforehand so that such task-dependent settling of activation is not possible.

### **The Model of Phaf, Van der Heijden, and Hudson (1990)**

Another influential model of the Stroop phenomenon is the connectionist model of Phaf, Van der Heijden, and Hudson (1990), called SLAM (for SeLective Attention Model), which has been developed within the framework of Van der Heijden’s (1992) general theory of attention. The model assumes an interactive-activation network. Input nodes for colors are connected to corresponding hidden nodes for colors, which in their turn are linked to word output nodes. Input nodes for words are directly connected to these output nodes. Thus, unlike the model of Cohen et al. (1990), the model assumes asymmetrical pathways for reading and color naming. Processing occurs through activation spreading from color input via hidden to output nodes, and directly from word input to output nodes, whereby nodes change their activation with time in a continuous, nonlinear manner. There are excitatory links between nodes representing compatible information and there are inhibitory links between nodes standing for incompatible information. All nodes of a particular type within a layer inhibit each other. Selective attention to the color naming and reading tasks is achieved by adding extra external activation to all hidden color nodes for color naming and all output nodes for word reading. The task activation is given from trial onset onward. On each simulated trial, word and color input is given to the network and activation cycles around from one unstable pattern to another until a stable pattern of activation is reached. The excitatory and inhibitory connections push activation of the response nodes into one stable state depending on the inputs provided to the layer (e.g., color and task input). To choose one response or another, activation of the response layer is

input to a sampling and recovery procedure that stochastically favors the most highly activated response node.

The Phaf et al. (1990) model successfully accounts for the basic Stroop effects with SOA = 0 msec, but, again, there are the same two major problems. First, the model does not adequately account for the time course of the Stroop phenomenon that has been observed by Glaser and Glaser (1982) and others. As simulations by Phaf et al. (1990, p. 324) showed, the model predicts that the amount of inhibition of words in color naming does not vary with SOA but remains constant for negative SOAs, contrary to the empirical findings. The reason for predicting a constant SOA effect in color naming is that after perceiving the word, the system quickly settles into a stable state of activation for the response corresponding to the word. By definition, the stable state does not vary with time, and hence making the SOA more or less negative has no effect, until an SOA is used that is too short for the distractor to reach an attractor basin. Consequently, the amount of time it takes for the color name to overcome the inhibition from the word is constant. The second major problem is that the model cannot cope with task uncertainty.

In line with Phaf et al. (1990), Cohen and Huston (1994) discuss an attractor version of the model proposed by Cohen et al. (1990). The behavior of this model is similar to that of Phaf et al. (1990). The amount of inhibition at negative SOAs is constant (see Figure 18.11 of Cohen and Huston, 1994), contrary to the real data. And the new version of the model also cannot cope with task uncertainty.

An alternative to connectionist task control is provided by “production rule system” models (e.g., Anderson, 1983; Anderson & Lebiere, 1998). Below, I show that the WEAVER++ model of word production (Levelt et al., 1999; Roelofs, 1992, 1993, 1997c), which falls into this general class of model, accounts for the findings on task uncertainty. The relevant features of the model are: (1) words are retrieved by spreading activation and (2) task-relevant control is achieved by production rule application.

## Control in the WEAVER++ Model

### Planning Stages

In WEAVER++, naming a perceptual entity such as a color involves a number of processing stages, illustrated in Figure 2. First, there is the conceptual identification of the color based on perceptual input (e.g., red) and its designation as goal concept (i.e., RED(X)). Second, the lemma of the corresponding word is retrieved (i.e., *red*), in the Stroop literature often referred to as response selection (except that it involves here lemmas, which is new). A lemma is a representation of the syntactic properties of a word, crucial for its use in sentences (cf. Roelofs, Meyer, & Levelt, 1998). Third, the form of the word is encoded (i.e., [rɛd]), called response programming. Lemma retrieval and word-form encoding are discrete processes in that only the form of a *selected* lemma becomes activated and selected (Levelt, Schriefers, Vorberg, Pechmann, Meyer, & Haviga, 1991). And finally, the name is articulated, called response execution.

A perceived written word activates its lemma and its

output form in parallel. Oral reading is achieved by a shallow form-to-form route (e.g., from the orthographic form *red* to [rɛd]) or may involve an extra step of lemma retrieval (i.e., from *red* via *red* to [rɛd]), roughly corresponding to what is traditionally called the “semantic” route (e.g., Caplan, 1992; Shallice, 1988). I refer to Levelt et al. (1999) and Roelofs, Meyer, and Levelt (1996) for an extensive discussion.

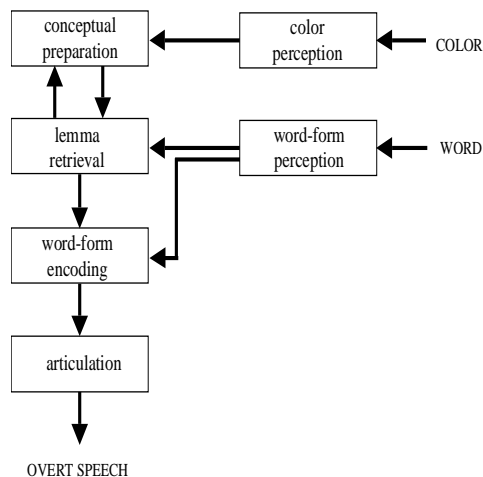


Figure 2: Stages of spoken word planning in WEAVER++

### Network Structure

The model assumes that the mental lexicon is a huge network with information about words, a small fragment of which is illustrated in Figure 3.

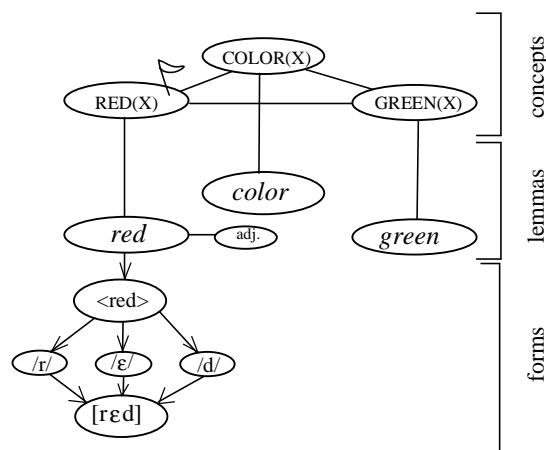


Figure 3: Fragment of the lexical network of WEAVER++

The network comprises three major strata: a conceptual, a syntactic, and a form stratum. The conceptual stratum represents concepts as nodes in a semantic network, following Collins and Loftus (1975), and many others. For example, the concept RED is represented by the node RED(X). The syntactic stratum contains lemma nodes, such as *red*, which are connected to nodes for their syntactic

class (e.g., adjective). And finally, the form stratum contains nodes representing morphemes, segments, and motor programs. For an extensive discussion of the theoretical and empirical motivation of these assumptions, I refer to Levelt (1989), Levelt et al. (1999), Roelofs (1992, 1993, 1996a,b,c, 1997a,b,c, 1998, 1999, 2000, submitted), and Roelofs and Meyer (1998).

### Spreading Activation and Production Rule Application

Information is retrieved from the network by the spreading of activation. For example, a perceived color (e.g., red) activates the corresponding concept node (i.e., RED(X)) in the network. Activation then spreads through the network following a linear activation rule with a decay factor. Each node sends a proportion of its activation to the nodes it is connected to. For example, RED(X) sends activation to other concepts such as GREEN(X) and also to its lemma node *red*. Selection of nodes is accomplished by production rules (i.e., condition-action pairs). A rule is triggered when its nodes become active. A lemma retrieval production rule selects a lemma if the connected concept is flagged as goal concept. For example, *red* is selected for RED(X) in case it is the goal concept and *red* has reached a critical difference in activation compared to other lemmas. The actual moment in time of the firing of a production rule whose condition is satisfied is determined by the ratio of activation of the relevant lemma node and the sum of all the others. Thus, how fast a node is selected depends on how active the other nodes are.

### Performing the Stroop Task

In color naming, a production rule like P1 controls general aspects of the task and a rule like P2 achieves the actual lemma selection (and sets a subgoal to encode the word's form, which is omitted here). Word reading is accomplished by a task rule like P3 that maps the orthographic code of a word onto the corresponding articulatory program. Earlier (Roelofs, 1992) I proposed an "intersection" mechanism to achieve selective attention in response selection, which has recently been dropped and replaced by the task production rules (see Roelofs, 2000, submitted).

- (P1) IF the goal is to say the name of the color  
and the concept is the color of the stimulus  
THEN select the concept  
and flag the concept as goal concept  
and enhance its activation
- (P2) IF RED(X) is flagged as goal concept  
and the activation of *red* exceeds threshold  
THEN select *red*
- (P3) IF the goal is to say the name of the word  
and the morpheme is the name of the stimulus  
THEN select the morpheme  
and flag the morpheme as goal morpheme

With task uncertainty, the task itself has to be set during

each trial. This is achieved by production rules like P4 and P5 (in the negative SOA condition).

- (P4) IF the first stimulus is a color  
THEN the goal is to name the word
- (P5) IF the first stimulus is a word  
THEN the goal is to name the color

To assess the Stroop performance of the model, computer simulations were run. The simulations employed a basic set of eight parameters, whose values were the same as in all earlier simulations (e.g., Levelt et al., 1999; Roelofs, 1992, 1993, 1996a, 1997c) except for two parameter values, which were changed slightly to fine-tune the fit of the model to the data. The "distractor duration" was set to 100 msec and the response threshold to 1.6. The distractor duration determines the gain of the distractor input relative to the target input. Roelofs (submitted) gives all the details of the simulations and applies the model to the key findings from over half a century of Stroop research (e.g., reviewed by MacLeod, 1991).

### Illustration of a Simulated Trial

Assume that the task is to name the second stimulus, which may be a color patch or a word. Assume that on a particular trial a red color patch is presented on which the word "green" is superimposed, with the word presented 100 msec before the color patch (i.e., the SOA is -100 msec). The simulation starts with the lemma node of "green" receiving external activation (for 100 msec, the distractor duration). This triggers production rule P5, which sets the goal to naming the color. Activation spreads through the network, with the node *green* sending a proportion of its activation to GREEN(X), and this node in its turn spreads activation to the other concept nodes. After the number of time steps that is the equivalent of 100 msec (the SOA), the concept node RED(X) receives external input from the color. Next, production rule P1 fires, RED(X) becomes flagged as goal concept, and its activation level is selectively enhanced. After the response threshold of the lemma *red* is exceeded, production rule P2 fires and *red* is selected as response.

### Simulation Results

The key finding to account for in this paper is that color naming is similarly affected by color words under task certainty and task uncertainty. Maximal inhibition of incongruent words on color naming is obtained when the words are presented within 100 msec of the colors, whereas the facilitation of preexposed congruent words is constant. Whereas color naming is affected by words, reading aloud is not affected by colors. Again, this holds both for task certainty and for task uncertainty.

Figure 4 shows how WEAVER++ performs. The figure shows the SOA curves of the Stroop effects for color naming under task uncertainty. The curves for task certainty (not shown) exhibit the same patterns. Maximal impact of incongruent words occurs in the model when the words are

presented within 100 msec of the color patches, exactly as empirically observed. For reading aloud in the model, no inhibition and facilitation is obtained at any SOA (also not shown), both for task certainty and task uncertainty, as empirically observed. Thus, WEAVER++ accounts for the time course findings and for the effect of the task certainty manipulation.

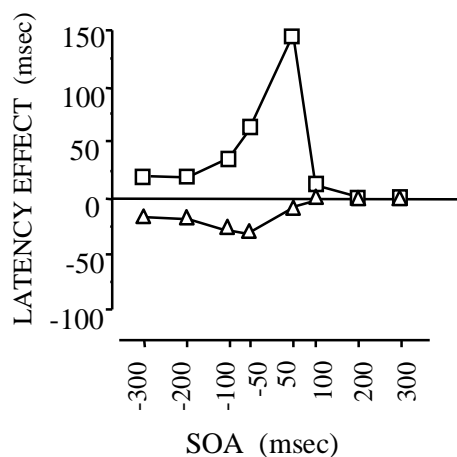


Figure 4: Time course of the Stroop effects (relative to control) in color naming under task uncertainty from WEAVER++ simulations:  $\square$  = incongruent,  $\triangle$  = congruent

Why is there no inhibition from colors on word reading? In WEAVER++, lemma retrieval and word-form encoding are discrete. Only the form of a selected lemma becomes activated and selected. Thus, activation does not spread automatically from lemmas to forms but this is under task control. Furthermore, color naming requires both lemma retrieval and form encoding, whereas word reading requires form encoding only. In reading “red” superimposed on a green color patch, the lemma but not the form of “green” becomes active: Because the task is reading and not color naming, the lemma of “green” (corresponding to the color) is not selected and the form of “green” does not become active. The task rule P3 for reading achieves direct selection of the morpheme <red> from the orthographic form *red* rather than indirect selection of <red> by first selecting the lemma *red* and next selecting <red> via the lemma. Thus, selecting <red> from the orthographic form *red* controls the response. Since the form of “green” is not active, planning the form of “red” is unaffected by the color patch.

Independent empirical support for the assumption of a discreteness of stages comes from double-task experiments. Levelt et al. (1991) asked participants to name pictured objects. On one third of the trials (the critical ones), a spoken probe was presented, and participants had to perform a lexical decision on this probe. Peterson and Savoy (1998) also asked participants to name pictures, but on the critical trials in their study written words were presented, which had to be read aloud. The lexical decision and reading latencies showed that in naming a perceptually given entity,

there is no form activation for non-synonymous semantic relatives (i.e., fellow category members) of the target. For example, in naming a cat, there is lemma activation for “cat” and “dog” and word form activation for “cat”, but the word form of “dog” is not activated. By extrapolation, in naming a red color patch, the lemmas of “red” and “green” become active and this also holds for the word form of “red”, but the word form of “green” is not activated. Only the form of a *selected* lemma becomes activated.

O’Seaghdha (1999) argues that the form of “dog” is activated during the naming of a cat, but that the experiments of Levelt et al. (1991) and Peterson and Savoy (1998) were insufficiently powerful to measure this. In support, he refers to a study by O’Seaghdha and Marin, who ran six experiments using word reading with word-word stimuli and an SOA of -500 msec. The effects in the experiments ranged from -2 to +5 msec and were not significant. However, by pooling the observations from the 248 participants in all six experiments, an overall effect of +3 msec was obtained, which reached significance by participants but not by items. By standard criteria, however, such an effect is nonsignificant. Moreover, with large negative SOAs (i.e., -500 msec) expectancy-based priming cannot be excluded (e.g., Neely, 1991). Thus, the findings of O’Seaghdha and Marin do not challenge the discreteness assumption.

## Summary and Conclusions

I have argued that performance on the Stroop task provides evidence on how speech production is controlled, that is, how a speaker exerts task-relevant control over the basic language processes underlying naming and oral reading. Color naming is inhibited by incongruent color words but word reading not by incongruent colors. Maximal impact of incongruent words on color naming is obtained when the words are presented within 100 msec of the colors, whereas the effect of preexposed congruent words is constant. The key observation for the current paper is that these findings are obtained both *with* and *without* task certainty. Whereas existing models (e.g., Cohen et al., 1990; Phaf et al., 1990) explain the basic Stroop effects, they fail to account for time course of the findings and for performance under task uncertainty. In this paper, I have extended and applied the WEAVER++ model of word production to performance on the Stroop task, and I have shown that the model accounts for the findings on the time course as well as for the performance under task uncertainty.

## References

- Allport, A. (1993). Attention and control: Have we been asking the wrong questions? A critical review of 25 years. In D. Meyer & S. Kornblum (Eds.), *Attention and Performance XIV: A silver jubilee* (pp. 183-218). Cambridge, MA: MIT Press.
- Allport, D. A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umiltà & M. Moscovitch (Eds.), *Attention and Performance XV: Conscious and nonconscious information processing* (pp. 421-452). Cambridge, MA: MIT

- Press.
- Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. London: Erlbaum.
- Caplan, D. (1992). *Language: Structure, processing, and disorders*. Cambridge, MA: MIT Press.
- Cohen, J., Dunbar, K., & McClelland, J. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, 97, 332-361.
- Cohen, J. D., & Huston, T. A. (1994). Progress in the use of interactive models for understanding attention and performance. In C. Umiltà & M. Moscovitch (Eds.), *Attention and Performance XV: Conscious and nonconscious information processing* (pp. 453-476). Cambridge, MA: MIT Press.
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407-428.
- Flowers, J. H., Warner, J. L., & Polansky, M. L. (1979). Response and encoding factors in "ignoring" irrelevant information. *Memory and Cognition*, 7, 86-94.
- Glaser, M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. *Journal of Experimental Psychology: Human Perception and Performance*, 8, 875-894.
- Glaser, W. R., & Glaser, M. O. (1989). Context effects in Stroop-like word and picture processing. *Journal of Experimental Psychology: General*, 118, 13-42.
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, 1-38.
- Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, Th. & Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review*, 98, 122-142.
- MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109, 163-203.
- Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264-336). Hillsdale, NJ: Erlbaum.
- O'Seaghdha, P. G. (1999). Parsimonious feedback. *Behavioral and Brain Sciences*, 21, 51-52.
- Peterson, R. R. & Savoy, P. (1998). Lexical selection and phonological encoding during language production: Evidence for cascaded processing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24, 539-557.
- Phaf, R. H., Van der Heijden, A. H. C., & Hudson, P. T. W. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology*, 22, 273-341.
- Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, 42, 107-142.
- Roelofs, A. (1993). Testing a non-decompositional theory of lemma retrieval in speaking: Retrieval of verbs. *Cognition*, 47, 59-87.
- Roelofs, A. (1996a). Computational models of lemma retrieval. In T. Dijkstra, & K. De Smedt (Eds.), *Computational psycholinguistics: AI and connectionist models of human language processing* (pp. 308-327). London: Taylor & Francis.
- Roelofs, A. (1996b). Serial order in planning the production of successive morphemes of a word. *Journal of Memory and Language*, 35, 854-876.
- Roelofs, A. (1996c). Morpheme frequency in speech production: Testing WEAVER. In G. E. Booij and J. van Marle (Eds.), *Yearbook of Morphology 1996* (pp. 135-154). Dordrecht: Kluwer Academic Publishers.
- Roelofs, A. (1997a). A case for nondecomposition in conceptually driven word retrieval. *Journal of Psycholinguistic Research*, 26, 33-67.
- Roelofs, A. (1997b). Syllabification in speech production: Evaluation of WEAVER. *Language and Cognitive Processes*, 12, 657-693.
- Roelofs, A. (1997c). The WEAVER model of word-form encoding in speech production. *Cognition*, 64, 249-284.
- Roelofs, A. (1998). Rightward incrementality in encoding simple phrasal forms in speech production: Verb-particle combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 904-921.
- Roelofs, A. (1999). Phonological segments and features as planning units in speech production. *Language and Cognitive Processes*, 14, 173-200.
- Roelofs, A. (2000). Control of language: A computational account of the Stroop asymmetry. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the Third International Conference on Cognitive Modeling* (pp. 234-241). Veenendaal, The Netherlands: Universal Press.
- Roelofs, A. (submitted). *Attentional control of verbal action: Stroop phenomena and their time course*.
- Roelofs, A., & Meyer, A. S. (1998). Metrical structure in planning the production of spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 922-939.
- Roelofs, A., Meyer, A.S., & Levelt, W.J.M. (1996). Interaction between semantic and orthographic factors in conceptually driven naming: Comment on Starreveld and La Heij (1995). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 246-251.
- Roelofs, A., Meyer, A.S., & Levelt, W.J.M. (1998). A case for the lemma-lexeme distinction in models of speaking: Comment on Caramazza and Miozzo (1997). *Cognition*, 69, 219-230.
- Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, 124, 207-231.
- Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge, UK: Cambridge University Press.
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643-662.
- Van der Heijden, A. H. C. (1992). *Selective attention in vision*. London: Routledge.

# Non-Linguistic Constraints on the Acquisition of Phrase Structure

Jenny R. Saffran (jsaffran@facstaff.wisc.edu)

Department of Psychology; 1202 W. Johnson Street  
Madison, WI 53706 USA

## Abstract

To what extent is linguistic structure learnable from statistical information in the input? One set of cues which might assist in the discovery of hierarchical phrase structure given serially presented input are the dependencies, or predictive relationships, present within phrases. In order to determine whether adult learners can use this statistical information, subjects were exposed to artificial languages which either contained or violated the kinds of dependencies which characterize natural languages. The results suggest that adults possess learning mechanisms which detect and utilize statistical cues to phrase and hierarchical structure. A second experiment contrasted the acquisition of these linguistic systems with the same grammars implemented as non-linguistic input (sequences of non-linguistic sounds or shapes). These findings suggest that constraints on the mechanisms which highlight the statistical cues which are most characteristic of human languages are not specifically tailored for language learning.

## Introduction

While the idea that surface distributional patterns point to pertinent linguistic structures holds a distinguished place in linguistic history (e.g., Bloomfield, 1933; Harris, 1951), statistical learning has only recently re-emerged as a potential contributing force in language acquisition (though see Maratsos & Chalkley, 1980). This renewed interest in statistical learning has been fueled by developments in computational modeling, by the widespread availability of large corpora of child-directed speech, and most recently by empirical research demonstrating that human subjects can perform statistical language learning tasks in laboratory experiments. For example, computational algorithms can use the co-occurrence environments of words to discover form classes in large corpora (e.g., Cartwright & Brent, 1997; Finch & Chater, 1994; Mintz, 1996; Mintz, Newport, & Bever, 1995). Similarly, individual verb argument structures can be induced by models which track the co-occurrences of verbs and their arguments in the input (e.g., Schütze, 1994; Seidenberg & MacDonald, 1999). Extensive modeling work has also examined the statistical cues available for the discovery of word boundaries in continuous speech (e.g., Aslin, Woodward, LaMendola, & Bever, 1996; Brent & Cartwright, 1996; Cairns, Shillcock, Chater, & Levy, 1997; Christiansen, Allen, & Seidenberg, 1998; Perruchet & Vintner, 1998).

These models provide invaluable explorations of the extent to which statistical information is available, in princi-

ple, to language learners equipped with the right distributional tools. But are humans such learners? A wealth of statistical cues are useless unless humans can detect and use them. In fact, recent research suggests that humans are extremely good at some statistical language learning tasks, such as word segmentation (e.g., Aslin, Saffran, & Newport, 1998; Goodsitt, Morgan & Kuhl, 1993; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996).

These results suggest that humans possess powerful statistical language learning mechanisms, which are likely to provide important contributions to the language learning process. At the same time, it is important to recognize that these mechanisms would not be useful in language acquisition unless they are somehow constrained or biased to perform only certain kinds of computations over certain kinds of input. The pertinent generalizations to be drawn from a linguistic corpus are awash in irrelevant information. Any learning device without the right architectural, representational, or computational constraints risks being sidetracked by the massive number of misleading generalizations available in the input (e.g., Gleitman & Wanner, 1982; Pinker, 1984). There are an infinite number of linguistically irrelevant statistics that an overly powerful statistical learner could compute: for example, which words are presented third in sentences, or which words follow words whose second syllable begins with *th* (e.g., Pinker, 1989).

One way to avoid this combinatorial explosion would be to impose constraints on statistical learning which perform only a subset of the logically possible computations. It is clear that learning in biological systems is limited by internal factors; there are species differences in which specific types of stimuli serve as privileged input (e.g., Garcia & Koelling, 1966; Marler, 1991). External factors also strongly bias learning, because input from structured domains consists of non-random information. In order for statistical learning accounts to succeed, learners must be similarly constrained: humans must be just the type of statistical learners who are best suited to acquire the type of input exemplified by natural languages, focusing on linguistically relevant statistics while ignoring the wealth of available irrelevant computations. Such constraints might arise from various sources, either specific to language or from more general cognitive and/or perceptual constraints on human learning.



We have recently begun to explore the possibility that statistical learning itself is constrained. This line of research focuses the acquisition of hierarchical phrase structure. While words are spoken and perceived serially, our representations of sequences of words are highly structured. Consider the sentence *The professor graded the exam*. This sequence of words cannot be grouped as follows – (*The*) (*professor graded the*) (*exam*) – because words that are part of the same phrase are separated. For example, determiners like *the* require nouns; separating these two types of words violates the dependency relations which are part of native speakers’ knowledge of English. The correct groupings, (*The professor*) (*graded the exam*), reflect English phrase structure, which generates a non-linear hierarchically organized structure. Hierarchical phrase structure represents a fascinating learning problem, because the child must somehow arrive at non-linear structure which is richer than is immediately suggested by the serial structure of the input. How do children make this leap? Innate knowledge is one possibility; prosodic regularities may also serve to chunk the input into phrasal units (e.g., Morgan, Meier, & Newport, 1987).

Another type of potentially useful information in the input suggests a statistical learning solution (see also Morgan & Newport, 1981). Linguistic phrases contain dependency relations: the presence of some word categories depends on others. For example, English nouns can occur without determiners like *the* or *a*. However, if a determiner is present, a noun almost always occurs somewhere downstream. This type of predictive relationship, which characterizes basic phrase types, may offer a statistical cue that highlights phrasal units for learners. Research using artificial languages with phrase structure grammars suggests that adult and child learners can exploit predictive dependencies to discover phrases (Saffran, 2000).

These studies suggest that people are skilled statistical learners. But what about the constraints required for the successful acquisition of languages? A particularly useful type of constraint would bias statistical learning mechanisms to preferentially acquire the types of structures observed in natural languages. To address this issue, Experiment 1 assessed the extent to which adults’ ability to acquire an artificial grammar is affected by the availability of predictive dependencies as cues to linguistic phrase structure.

## Experiment 1

**Participants.** 40 monolingual English speaking undergraduates at the University of Rochester participated in this study, and were each paid \$6. Subjects were randomly assigned to the two experimental conditions.

**Materials.** The artificial grammars were adapted from the language used by Morgan & Newport (1981). One of the languages used in this study was a small phrase structure grammar (Language P, for predictive), in which dependencies

between word categories afforded predictive cues to phrases, as in natural languages (e.g., if D is present, A must be present). Importantly, attempts to impose English predictive structure onto the input would mislead learners, as the phrase structure of Language P was head-final while English is head-initial. The second language was equally complex in terms of its size and formal characteristics, but contained a phrase structure unlike natural languages (Language N, for non-predictive). This language did not contain predictive dependencies marking phrases. Rather, it was characterized by overarching optionality: the presence of one word type never predicted the presence of another, which generates statistical properties unlike natural languages (note, however, that this language still possesses phrase structure of a sort – the absence of one word type predicts the presence of another; e.g., if A is not present, D must be present). Each form class (A, C, etc.) included 2 - 4 nonsense words (e.g., the words for the A category were BIFF, RUD, HEP, and MIB).

Table 1. Phrase structure grammars for Experiments 1 - 2. Letters refer to word classes; items in parentheses are optional. In Language N, one member of each phrase type must be present; if both are present, they must be in the order described by the grammar.

<u>Language P</u>	<u>Language N</u>
S → AP + BP + (CP)	S → AP + BP
AP → A + (D)	AP → (A) + (D)
BP → CP + F	BP → CP + F
CP → C + (G)	CP → (C) + (G)

The language generated by Language N is no larger than the language generated by Language P. In fact, Language N contained fewer sentence types (nine) than Language P (twelve). Language N also had shorter sentences on average, presumably making it less daunting to the learner: Language P generated 60% more five word sentences than Language N, and only 40% as many three word sentences. For both languages, only sentence types with five or fewer words were used (eight types for Language P, nine for Language N). Both languages contained the same number of grammatical categories and vocabulary items.

Because the languages were so similar in terms of their non-structural attributes, comparison of learning outcomes is valid. Language P is larger, and contains longer sentences, which could make it more difficult to acquire. However, if predictiveness affects learning, then the structure of Language N might have hindered its acquisition. A trained speaker recorded a corpus of 50 sentences from each language, with uniformly descending prosody but no grouping cues to phrase structure. Subjects were randomly assigned to hear either Language P or Language N sentences. Following approximately 30 min. of auditory exposure to one of the

two languages (the corpus was repeated eight times during exposure), all participants received the same forced-choice test consisting of novel grammatical and ungrammatical sentences, in order to assess acquisition of the rules of the two languages. Importantly, attempts to impose English syntax on either language would hinder performance. No cues other than the statistical information mirroring the underlying phrase structure of the language were available to learners.

**Results.** Each group's overall performance was significantly better than would be expected by chance: for Language P, the total score was 22.8 out of a possible 30:  $t(19) = 10.46, p < .0001$ ; for Language N, the total score was 20.55:  $t(19) = 6.62, p < .0001$  (see Figure 1). The principal hypothesis of interest concerns differences in learning as a function of structural differences between the two languages. To address this question, the scores for the two language groups for items testing each of the five rules were submitted to an ANOVA. The main effect of Language (P versus N) was significant:  $F(1, 38) = 4.2, p < .05$ .

These findings suggest that humans may be constrained to learn most readily via exactly the types of cues present in languages. To the extent that this is the case, the structure of natural languages may have been shaped by the nature of human learning (e.g., Bever, 1970; Christiansen, 1994; Christiansen & Devlin, 1997; Morgan, Meier, & Newport, 1987; Newport, 1990). According to the constrained statistical learning hypothesis, the mechanisms underlying language acquisition are biased to assist learners in detecting the 'right' statistical properties of the input. On this view, human languages have been sculpted by human learning and processing mechanisms – thereby creating input which contains the types of properties most useful for human learners, and rendering a close match between constraints on human learning and constraints on natural language structure.

If learners are biased to preferentially acquire structures where one item predicts another, is this constraint on learning particularly tailored for linguistic input? Biases in learning mechanisms may develop tightly coupled with the particular structure they are designed to acquire. Alternatively, constraints to use predictive statistics may be more generally applied to other types of sequentially presented information, as suggested by the constrained statistical learning hypothesis. Constraints on statistical learning which are not specific to language acquisition, but rather on the acquisition and processing of serial information, may have shaped the structure of natural languages. Experiment 2 thus utilized non-linguistic stimuli from two different modalities: visual shapes and complex sounds. An additional condition included visual linguistic stimuli (written words). As in Experiment 1, we contrasted the acquisition of Language P and N.

## Experiment 2

**Participants.** 154 monolingual English speaking undergraduates at the University of Wisconsin - Madison participated in this study. Forty-four subjects were randomly assigned to the non-linguistic auditory condition, forty subjects to the non-linguistic visual condition, and thirty subjects to the linguistic visual condition. Within each exposure condition, half of the subjects were assigned to Language P and half were assigned to Language N.

**Method.** For the non-linguistic visual condition, we translated the Language P and N grammars shown above into languages of shapes (for a similar methodology, see Goldowsky, 1995). For example, consider the phrase structure rule:  $AP \rightarrow A + (D)$ . In the linguistic version of this language, the category A consisted of 4 nonsense words. In the visual version, the category A consisted of 4 distinct shapes (such as a red circle with stripes). Category membership could not be induced by shape similarity, unlike prior studies by Morgan & Newport (1981). Participants observed the language on a computer monitor: each shape was presented in the middle of the screen, one at a time, with the same timing parameters as the auditory linguistic stimuli used in Experiment 1. Following exposure, participants were tested using a forced-choice test analogous to the linguistic task, in which they saw two shape sequences, one after the other, and decided which shape sequence more closely approximated the exposure stimuli. The linguistic visual condition was identical to the non-linguistic visual condition except that the nonsense words from Experiment 1 were shown typed on the computer screen. In the non-linguistic auditory condition, we translated Language P and N into non-linguistic sounds drawn from the digitized bank of alert sounds provided with Windows 98. Each word corresponded to a different sound, chosen to be maximally discriminable (an ascending buzz, a chord, chimes, etc.). Sound "sentences" generated by Language P and N were presented auditorily at the same rate as the linguistic and visual stimuli. Following exposure, participants received the same forced choice test, translated into non-linguistic sounds. Neither of the two non-linguistic conditions contained any linguistic information.

**Results.** Each group's overall performance was significantly better than would be expected by chance: for Language P Non-linguistic auditory, Nonlinguistic visual, and Linguistic visual,  $p < .0001$ ; for Language N Nonlinguistic visual,  $p < .001$ ; for Language N Nonlinguistic auditory,  $p < .001$ ; and for Language N Linguistic visual,  $p < .05$  (see Figure 1). As in Experiment 1, the principal hypothesis concerns

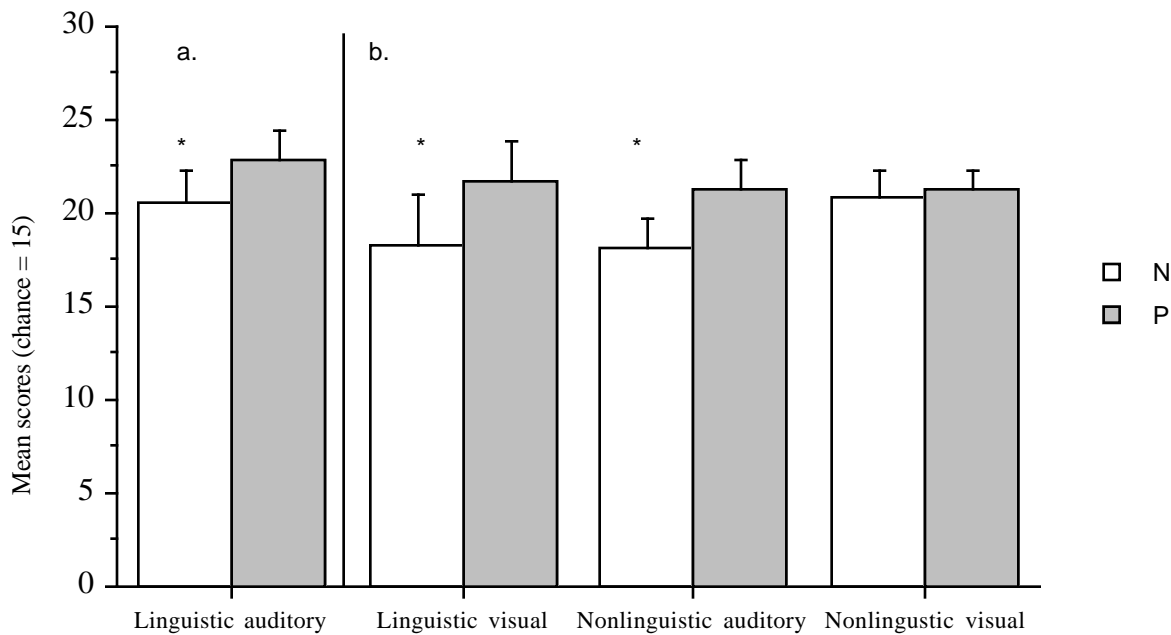


Figure 1: a. Mean scores from Experiment 1. b. Mean scores from Experiment 2.

differences in learning as a function of structural differences between Language P and Language N. To address this question, the scores for the two language groups for items testing each of the five rules were submitted to an ANOVA. The main effect of Language P versus N was significant for the Nonlinguistic auditory [ $F(1, 42) = 7.72, p < .01$ ] and the Linguistic visual condition [ $F(1, 28) = 4.56, p < .05$ ], but not for the Nonlinguistic visual condition [ $F(1, 38) = .23, n.s.$ ].

In order to ask whether the linguistic or non-linguistic status of the input influenced performance differentially as a function of the availability of linguistic dependencies, we performed a two-way between-subjects ANOVA contrasting Language (P versus N) and Linguistic Status (language versus non-language materials), including the auditory linguistic data from Experiment 1. There was a significant main effect of Language:  $F(1, 150) = 15.17, p < .0001$ . Neither the main effect of Linguistic Status [ $F(1, 150) = 1.09, n.s.$ ] nor the interaction between Language and Linguistic Status [ $F(1, 150) = .71, n.s.$ ] were significant. These analyses indicate that the linguistic status of the input – that is, whether the grammars were implemented in linguistic or non-linguistic tokens – did not affect overall performance. Instead, the dominant factor was whether the input was derived from Language P, which contained predictive dependencies as a statistical cue to phrase structure, or Language N, which did not. This overall non-effect of linguistic status occurred despite the fact that performance on the visual non-linguistic task did not show the predicted difference between Language P and N (see Figure 1). We are currently testing hypotheses concerning why the visual nonlinguistic task patterned differently from the other three conditions included in Experiments 1 and 2.

## General Discussion

These studies ask whether predictive dependencies serve a learnability function in the acquisition of language. The results of Experiment 1 suggest that adult learners are better able to acquire an artificial language which contains predictive dependencies as a cue to phrase structure than a comparable language which does not. Experiment 2 extends these results to demonstrate that the use of predictive dependencies in learning phrase structure is not limited to language learning tasks. These findings mirror prior results suggesting that transitional probability computation in word segmentation tasks can occur when ‘words’ are created from non-linguistic tones (Saffran, Johnson, Newport, & Aslin, 1999) or visuo-motor sequences (Hunt & Aslin, 1998).

Predictive dependencies are a hallmark of natural languages. However, it is of interest to note that these general organizational principles are by no means unique to language. Lashley (1951) observed that hierarchical organization characterizes an enormous variety of behaviors: “the coordination of leg movements in insects, the song of birds, the control of trotting and pacing in a gaited horse, the rat running the maze, the architect designing a house, and the carpenter sawing a board present a problem of sequences of action which cannot be explained in terms of successions of external stimuli” (p. 113). Such observations suggest that learners may be biased to process information in a particular fashion, enabling a learning process which results in phrases and hierarchically structured representations.

The kinds of structure at issue here serve to organize and package serial information into manageable chunks, which then enter relationships with one another. This process presumably maximizes cognitive economy, facilitating the

transmission of more complex information than could be transmitted otherwise. Pinker and Bloom (1990) argue that "hierarchical organization characterizes many neural systems, perhaps any system, that we would want to call complex...Hierarchy and seriality are so useful that for all we know they may have evolved many times in neural systems" (p. 726). When applied to syntax, this kind of argument suggests that grammars look the way they do because these kinds of organizational principles are the human engineering solution to the problem of serial order.

It is conceivable that this type of packaging of serial inputs into higher-order organization facilitates not only language production and processing, but also language acquisition. Systems which are highly organized are more learnable than systems which are not -- as long as the system of organization is consistent with the learner's cognitive structure. We anticipate that future research will be extremely useful in further clarifying the extent to which the constraints observed during the process of language acquisition subserve other learning processes as well.

With respect to linguistic structure, one potential theoretical implication of this research concerns an alternative to the traditional innate universal grammar explanation for the pervasiveness of particular linguistic features cross-linguistically.. If human learners are constrained to preferentially acquire certain types of structures, then some of the universal structures of natural languages may have been shaped by these constraints (see also, e.g., Bever, 1970; Christiansen, 1994; Christiansen & Devlin, 1997; Newport, 1982, 1990). Perhaps languages fit our learning abilities so neatly precisely because languages have no choice. If the pertinent learning mechanisms preceded the advent of languages, then there must have been intense pressure for languages to be learnable, with learnability dictated by the structure of human learning mechanisms. On this view, languages evolve to fit the human learner. To the extent that this type of view is correct, then the striking similarities of human languages may be in part the direct reflections of constraints on human learning abilities.

The present research begins the task of recharacterizing language universals in terms of constraints on learning by recasting the distributional features and dependencies inherent in hierarchical phrase structure into cues detected during the learning process. In the case of the constraint to interpret predictive relations as signaling a linguistic unit, the phrase, we find the beginnings of an explanation for why languages ubiquitously contain the within-phrase dependencies initially characterized by structural linguists. Future research will continue to pursue the hypothesis that constraints on learning play an important role in shaping the structure of natural languages. For example, recent computational research suggests that universal word order typologies may in fact reflect the ease with which different types of systems are learned (Christiansen & Devlin, 1997).

With respect to statistical learning, the present research runs counter to the assumption that statistical language learning accounts -- and any other type of theory which assigns an important role to linguistic input -- are necessarily underconstrained. As animal research has amply demonstrated, learning in biological systems is highly constrained (e.g., Garcia & Koelling, 1966; Marler, 1991). There is every reason to believe that statistical learning is similarly constrained; the purported intractability of statistical learning need not be asserted *prima facie*. What exactly these constraints will turn out to be, and whether they will confer sufficient explanatory power, remain empirical questions. Nevertheless, there are grounds for optimism. Learners are not, and never have been, blank slates. The more we learn about the mechanisms engraved upon that slate, the more we learn about learning.

## Acknowledgments

This research was supported by NIH Training Grant 5T32DC0003 to the University of Rochester, by NIH grant DC00167 to Elissa Newport, and by NIH grant 144HN72 to Jenny Saffran.

## References

- Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax*. Hillsdale, NJ: Erlbaum.
- Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9, 321-324.
- Bever, T. (1970). The cognitive basis for linguistic structures. In J. Hayes (Ed.), *Cognition and the development of language*. New York: Wiley.
- Bloomfield, L. (1933). *Language*. New York: Henry Holt.
- Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61, 93-125.
- Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, 33, 111-153.
- Cartwright, T. A., & Brent, M. R. (1997). Early acquisition of syntactic categories: A formal model. *Cognition*, 63, 121-170.
- Christiansen, M. H. (1994). *Infinite languages, finite minds: Connectionism, learning and linguistic structure*. Unpublished Ph.D. dissertation, University of Edinburgh.
- Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13, 221-268.

- Christiansen, M. H., & Devlin, J. T. (1997). Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In *Proceedings of the Nineteenth Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Finch, S. P., & Chater, N. (1994). Distributional bootstrapping: From word class to proto-sentence. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Garcia, J. & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science*, 4, 123-124.
- Gleitman, L. R., & Wanner, E. (1982). Language acquisition: The state of the state of the art. In E. Wanner and L. R. Gleitman (Eds.), *Language acquisition: The state of the art*. Cambridge: Cambridge University Press.
- Goldowsky, B. (1995). *Learning structured systems from imperfect information*. Unpublished Ph.D. dissertation, University of Rochester.
- Goodsitt, J. V., Morgan, J. L., & Kuhl, P. K. (1993). Perceptual strategies in prelingual speech segmentation. *Journal of Child Language*, 20, 229-252.
- Harris, Z. S. (1951). *Methods in structural linguistics*. Chicago: University of Chicago Press.
- Hunt, R. H., & Aslin, R. N. (1998). Statistical learning of visuomotor sequences: Implicit acquisition of sub-patterns. In *Proceedings of the Twentieth Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon Symposium*. New York: Wiley.
- Maratsos, M., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. Nelson (Ed.), *Children's language, Vol. 2*. New York: Gardner Press.
- Marler, P. (1991). The instinct to learn. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition*. Hillsdale, NJ: Erlbaum.
- Mintz, T. H. (1996). *The roles of linguistic input and innate mechanisms in children's acquisition of grammatical categories*. Unpublished Ph.D. dissertation, University of Rochester.
- Mintz, T. H., Newport, E. L., & Bever, T. G. (1995). Distributional regularities of form class in speech to young children. *Proceedings of NELS 25*. Amherst, MA: GLSA.
- Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology*, 19, 498-550.
- Morgan, J. L., Meier, R. P., & Newport, E. L. (1989). Facilitating the acquisition of syntax with cross-sentential cues to phrase structure. *Journal of Memory and Language*, 28, 360-374.
- Morgan, J. L., & Newport, E. L. (1981). The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior*, 20, 67-85.
- Newport, E. L. (1982). Task specificity in language learning? Evidence from speech perception and American Sign Language. In E. Wanner and L. R. Gleitman (Eds.), *Language acquisition: The state of the art*. Cambridge: Cambridge University Press.
- Newport, E. L. (1990). maturational constraints on language learning. *Cognitive Science*, 14, 11-28.
- Perruchet, P., & Vintner, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, 39, 246-263.
- Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: MIT Press.
- Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.
- Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13, 707-784.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926-1928.
- Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70, 27-52.
- Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35, 606-621.
- Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, 8, 101-195.
- Schütze, H. (1994). A connectionist model of verb sub-categorization. *Proceedings of the 16th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Seidenberg, M. S., & MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, 23, 569-588.

# Attentional Biases in Artificial Noun Learning Tasks: Generalizations Across the Structure of Already-Learned Nouns

Larissa K. Samuelson (samuelso@indiana.edu)

Department of Psychology and  
Program in Cognitive Science  
Indiana University  
Bloomington, IN 47405

## Abstract

Even though learning noun meanings in a first language should be a difficult task young children learn nouns quickly and often with little effort. Previous research suggests that the task of learning words is made easier by constraints or biases that reduce the problem of finding the correct word-referent mapping to a solvable size. The research presented here examines the relation between the attentional biases young children demonstrate in laboratory noun learning tasks and the pattern of word learning seen outside the laboratory. The comparison suggests that attentional biases in laboratory noun learning tasks are a generalization across the nouns young children have already learned. Further, changing the nouns young children know changes not only the attentional biases they demonstrate in the laboratory, but also their vocabulary development outside the laboratory.

## Introduction

Young children typically say their first word at 1-year-of-age. However, conservative estimates suggest that by 5-years-of-age children have as many as 10,000 words in their productive vocabulary. How do children learn so many words so fast? One suggestion is that the task of learning words is made easier by biases or constraints which reduce the problem of finding the correct word-referent mapping to a solvable size (e.g. Landau, Smith, & Jones, 1988; Markman, 1992; Soja, Carey, & Spelke, 1991). There is strong experimental evidence for the existence of a number of these word learning biases. The research presented here concentrates on two: the shape bias for learning names of solid objects and the material bias for learning names of non solid substances.

Evidence for these two attentional biases comes from artificial noun learning experiments. In these experiments, a young child is presented with a novel object. This exemplar object is then named, i.e. "this is a dax". The child is then presented with novel test objects that match the exemplar in one perceptual dimension, for example in shape only, color

only, or material only. The child is then asked which of these test objects can be called by the same name as the exemplar.

Numerous studies have shown that when the exemplar object is made of a solid, rigid material such as wood or hardened clay, children 24-months-of-age and older generalize novel names to other objects that match the exemplar in shape. This "shape-bias" has been demonstrated in numerous laboratories, with stimuli ranging from real, 3-dimensional objects specially constructed for the experiment (Imai & Gentner, 1997; Landau, et al., 1988), to pictures of familiar objects (Imai, Gentner, & Uchida, 1994). However, when the exemplar object is made from a non solid substance such as hair gel or face cream, children generalize the novel name to test objects made from the same material as the exemplar (Dickinson, 1988; Soja, 1992; Soja, et al., 1991). This "material-bias" has also been demonstrated in numerous studies and laboratories, however this bias does not appear to be robust until after 30-months-of-age (Samuelson & Smith, 1999).

In a series of recent studies, I have examined the relation between these attentional biases, demonstrated in laboratory tasks, and the pattern of noun vocabulary growth seen outside the laboratory (Samuelson & Smith, 1999). These experiments suggest that the attentional biases seen in laboratory word learning tasks may be generalizations across the category structure of already learned nouns. This paper reviews these findings and the suggested hypothesis. Two experiments testing this hypothesis are then presented. The results show that changing the nouns young children know changes the development of attentional biases seen in laboratory word learning tasks, and that this change further alters the trajectory of vocabulary development outside the laboratory

## Attentional Biases and the Nouns Children Know

If the shape and material biases are to help children learn nouns, then these biases need to match the kinds of nouns that young children learn early. That is, if the shape bias helps children learn names for solid objects by directing their attention to within category similarity in shape, then many

of the nouns young children learn early should refer to solid things in categories well organized by shape. Likewise, if the material bias helps children learn names for non solid things by directing their attention to similarity in material substance, then there should also be many names for non solid things in categories well organized by material substance among the nouns children learn early. An important question, then, is what kind of nouns do young children learn early?

To answer this question, I examined the category structures of a corpus of early-learned nouns. The corpus of 312 nouns studied was taken from the Mac Arthur Communicative Development Inventory (MCDI), a parental checklist of 680 words and phrases commonly found in the productive vocabulary of children between 16- and 30-months-of-age. In a series of yes/no judgments, thirteen adult native speakers of English were asked to think of examples of each noun in the studied corpus and say whether the examples were solid, non solid, similar in shape, and similar in material. An 85% agreement criterion was then used to determine the structure of the category referred to by each noun. For example, 85% of the adults agreed that crayons were solid, similar in shape, and similar in material. Thus, CRAYON was classified as referring to a category of solid things similar in both shape and material.

A summary of the findings across the entire corpus can be seen in Figure 1. In the figure, each square represents the 312 nouns studied. The area of each circle represents the proportion of those nouns that fell in each classification. And, the overlapping area of the circles represents the proportion of nouns that fell in the intersection of the classifications. As can be seen in the figure, many of the nouns children learn early name solid things and things in categories well organized by shape. And, there is a large amount of overlap between these classifications; many of the nouns children learn early name solid things in shape-based categories. In contrast, few of the nouns children learn early name non solid substances or things in categories well organized by similarity in material substance. And, there is not much overlap between these classifications; young children do not learn many names for non solid substances

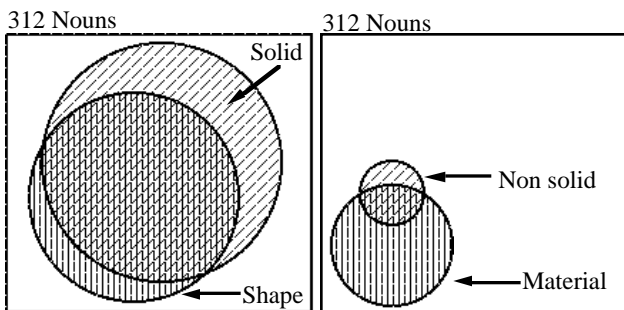


Figure 1: Summary of the category structure of the corpus of 312 early-learned nouns.

in material-based categories (Samuelson & Smith, 1999).

Thus, many of the nouns children commonly learn by 30-months-of-age fit the shape-bias. And, by 30-months-of-age children have not learned many nouns that fit the material-bias. These facts fit with previous findings that children demonstrate a shape-bias in artificial noun learning tasks by 24-months-of-age, but do not reliably demonstrate a material bias until 36-months-of-age. However, this study does not address the developmental relation between these findings. That is, do children demonstrate a shape bias because they know many names for solid things in shape-based categories. Or, do children learn many names for solid things in shape based categories because they have a shape bias? To address this question, I compared artificial noun learning with solid and non solid stimuli in children with a range of vocabulary sizes. Specifically, fifty-eight children between 17- and 31-months of age completed a forced choice artificial noun learning task in which half the exemplars and choice stimuli were made from solid materials such as wood and Styrofoam, and the other half were made of non solid materials such as hair gel and face cream. I also measured each child's productive noun vocabulary via parental report on the MCDI.

Figure 2 presents the key results. As can be seen in the figure, I found that children did not generalize novel names for solid objects to other solid objects by shape at levels reliably above chance until they already had 150 nouns in their productive vocabulary. And, children in the vocabulary range I studied did not generalize novel names for non solid stimuli to other non solid stimuli at levels reliably above chance. Thus, it appears that the shape-bias emerges only after children have already learned many names for shape-based categories. And, the material-bias does not emerge within the vocabulary range I studied (Samuelson and Smith, 1999).

These results suggest that the attentional biases young children demonstrate in artificial noun learning tasks might be the product of their previous noun learning. More

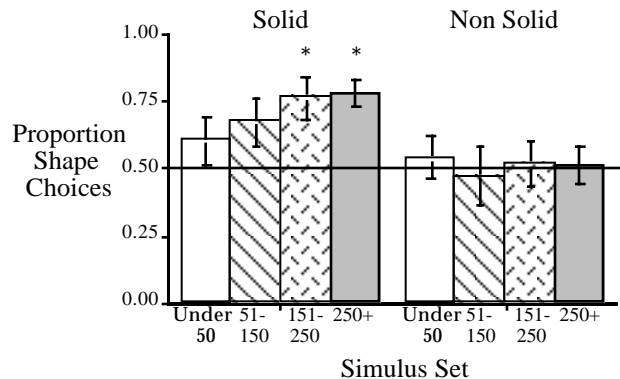


Figure 2: Proportion shape choices by noun vocabulary size for the solid and non solid stimulus sets. Chance responding is .50, \* =  $p < .05$  difference from chance.

specifically, what children do in artificial noun learning tasks is attend to whatever perceptual property has mattered most in learning the nouns they already know. Attentional biases thus appear to be a generalization across the structure of already learned nouns. This hypothesis is tested in the following experiment.

## Experiment 1

Previous results suggest that children's performance in artificial noun learning tasks is a generalization across the structure of already learned nouns. The specific question addressed in this experiment is whether changing the nouns children know changes their attentional biases in artificial noun learning tasks. The idea was to intensively teach nouns to young children who do not have many nouns in their vocabulary and do not yet demonstrate systematic attentional biases in artificial noun learning tasks. If children's attentional biases in noun learning are a product of the nouns they already know, then changing the nouns they know should change their attentional biases.

To this end, two groups of children participated in a nine week longitudinal study. Children in one group were taught twelve names for categories of solid objects well organized by similarity in shape. Another group of children were taught twelve names for categories of non solid substances well organized by similarity in material substance. Two dependent measures were examined: artificial noun learning with both solid and non solid stimuli, and productive vocabulary via parental report. Both measures were taken early in training and, again, later in training. In addition, a follow-up report of productive vocabulary was obtained one month after the experiment was complete.

## Methods

**Participants.** Twenty children between 15- and 21-months-of-age participated (mean 18m 28d, range 15m 20d to 21m 3d). Children were recruited from the child participants file at Indiana University and contacted by phone. All children were learning English as their first language. Children were randomly assigned to either the Shape Nouns or Material Nouns condition such that the mean age and vocabulary across conditions did not differ. Four additional children began but did not complete the experiment. All children received a small prize at each experimental visit and copies of experimental videotapes and T-shirts at the completion of the study.

**Materials.** Conditions differed only in the twelve nouns taught to the children over the course of the longitudinal study. All twenty-four nouns are nouns not usually learned until after 26 months of age. The noun category training sets for each condition consisted of three examples of each of the twelve nouns.

In the Shape Nouns condition, children were taught twelve names for solid things in categories well organized by shape,

for example, bucket, pear, and ladder. In this condition, example items for each category were the same in shape but differed in size, color, and the material they were made from. For example, one ladder was wide and made of red wood, one was taller and made of white plastic and one was short and wide and made of pink metal.

In the Material Nouns condition, children were taught twelve names for non solid things in categories well organized by material, for example, glitter, lotion, and Jell-O. In this condition, example items for each category were made from the same material but differed in amount, color, and shape. For example, the Jell-O was either red, orange or blue and was either presented as a large pile, a couple small piles, or in the shape of a teddy-bear, and these shapes and amounts changed as the child ate the Jell-O.

Eight sets of artificial noun learning stimuli were also constructed—four made from solid materials such as wood and Styrofoam and four made from non solid materials such as hair gel and face cream. Each set consisted of an exemplar object and four test objects. In each set, two test objects were the same shape as the exemplar but were different colors and made from different materials, and two test objects were made from the same material as the exemplar but were different in shape and color. Eight unique nonsense words were created for use in the artificial noun learning task. The pairing of names to stimulus sets was counterbalanced across children.

Twenty unique sets of practice stimuli were also assembled for use in practice artificial noun learning trials. These sets consisted of small toys familiar to most 15-month-olds such as balls, toy cars, and cups. Each set consisted of two identical toys and a third toy that differed in color, shape and size (for example, two purple plastic eggs and a red wooden block).

**Procedure.** Children and their parents visited the lab once a week for nine consecutive weeks. These nine weeks were broken into three blocks of three weeks each. Each block consisted of two weeks of noun category training. On the third week children were tested in artificial noun learning. Productive vocabulary was measured via parent report on the Mac Arthur Communicative Development Inventory at the beginning and end of the experiment and at a follow-up appointment one month after the final experimental session.

During all experimental sessions, the child sat across a large table from the experimenter with his or her parent. Experimental sessions began with two practice trials of the artificial noun learning task. These practice trials were used both to engage the child in the experimental session, and to encourage their participation in the artificial noun learning task. In these practice trials, the experimenter gave the child one set of practice stimuli to examine. After the child had examined the items the experimenter retrieved the toys, put one of the matching pair and the non-matching item on a tray, held up the other matching item and said, "See this,



this is my (name of toy).” She then pushed the tray towards the child saying “Can you get your (name of toy)”. If the child picked-up or gestured towards the matching toy she was praised heavily. If she picked the incorrect toy the experimenter said “Is that the (name of toy)? No! Get the (name of toy)” until the child picked the correct toy.

On noun category training weeks, noun training followed the practice trials. During noun training, the child, experimenter, and parent played with and named the examples of four noun categories. The three examples of each category were played with as a set for approximately three minutes each. The experimenter then put these items away and brought out the examples of the next noun category. The experimenter named each noun category at least 20 times for each child, and encouraged the child to say each noun at least once.

On artificial noun learning weeks, the artificial noun learning task followed the practice trials. This task was identical to the procedure used during the practice trials. The child was given the exemplar, one shape-match test object, and one material-match test object to examine. The experimenter then placed the test objects on the tray, held up the exemplar and said, “See this, this is my bing”, (for example). The experimenter then pushed the tray towards the child and said “Can you get your bing?”. If the child did not respond she was prompted again. The experimenter then proceeded to the next trial for that stimulus set. The parent was asked not to refer to the stimuli during this task but to encourage the child to respond. There were four trials for each stimulus set (each shape-match test object with each material-match test object) and one solid and one non solid stimulus set at each artificial noun learning task. Children never saw the same stimulus set twice. Order of solid and non solid sets was counterbalanced across artificial noun learning tests and order of stimulus sets was counterbalanced across children.

All experimental sessions were video taped for later coding of naming instances and artificial noun learning responses. Three coders blind to the experimental hypothesis coded all artificial noun learning sessions. Coders indicated which test object the child picked on each trial. Twenty percent of the trials were coded by two coders and reliability was greater than 90%.

## Results and Discussion

Figure 3 shows the mean proportion of shape choices in the artificial noun learning task with solid and non solid stimuli for children in the Shape Nouns and Material Nouns conditions at weeks three and nine. A Condition (Shape Nouns v. Material Nouns) X Stimulus Set (Solid v. Non Solid) X Week (3 v. 9) ANOVA revealed a significant main effect of Condition,  $F = 6.854$ ,  $p < .02$ , and significant Week by Condition and Stimulus Set by Week interactions,  $F = 4.695$ ,  $p < .05$ ,  $F = 10.436$ ,  $p < .01$  respectively. As can be seen in the figure, at week three there were no

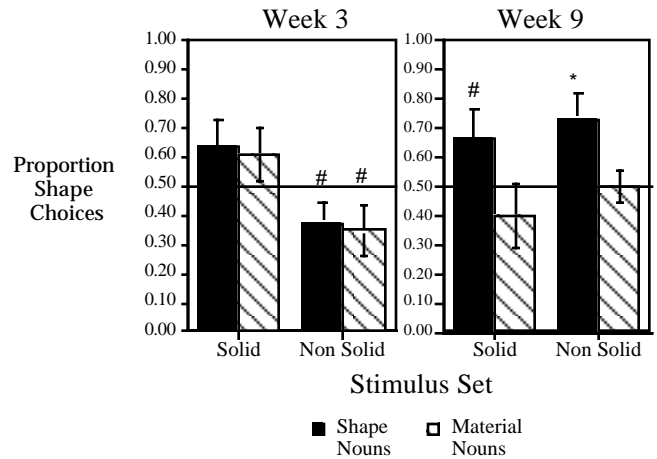


Figure 3. Results of Experiment 1. Chance responding equals .50. \* =  $p < .05$ , # =  $p = .06$  difference from chance.

differences between children in the two conditions in their responding to the solid or non solid stimuli. However, by week nine, children in the Shape Nouns condition picked shape matching test objects more than children in the Material Nouns condition for both the solid and non solid sets, Tukey’s HSD  $p < .05$ . Further, only children in the Shape Nouns condition at week nine picked shape matching test objects at levels significantly above chance. Thus, children who were taught twelve names for solid objects in categories well organized by shape also learned a shape-bias but children who were taught twelve names for non solid substances in categories well organized by material substance did not learn a material-bias.

Importantly, this learned shape-bias demonstrated in the laboratory also influenced children’s vocabulary development outside of the laboratory. Figure 4 presents the mean number of words in the total productive vocabulary of children in each condition at the first experimental session, at the last experimental session, and at the follow-up appointment one month after the experiment had ended. As can be seen in the figure, during the course of the experiment children in both conditions learned new words outside the laboratory at rates that did not differ. However, after the experiment ended, children in the Shape Nouns condition acquired significantly more words by the follow-up appointment. It appears that the twelve shape-biased nouns they were taught in the laboratory somehow accelerated their learning of other words outside the laboratory. It is also possible, however, that the difference in the vocabularies of children in the two conditions at the follow-up appointment was actually due to a suppression of the vocabulary development of children in the Material Nouns condition. Perhaps teaching these children twelve names for non solid things in categories well organized by material substance—categories that even 30-month-old children do not know many of—actually harmed their vocabulary development. This possibility was tested in Experiment 2.

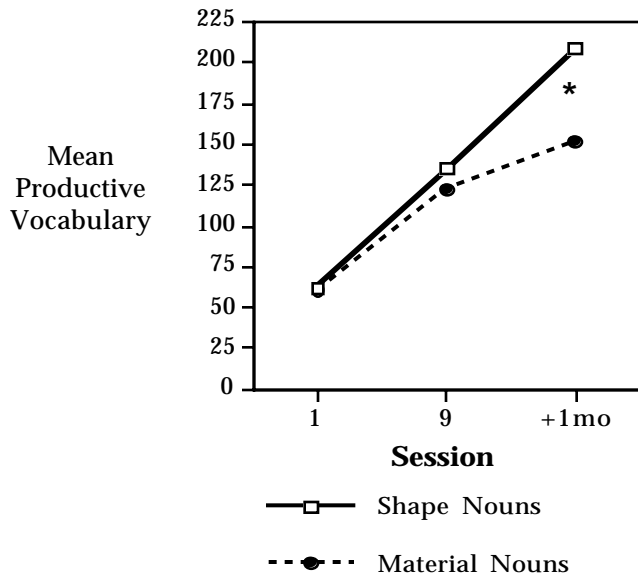


Figure 4. Mean productive vocabulary of children in experiment 1.

### Experiment 2

This experiment provides a control for the possibility that the vocabulary development of children in the Material Nouns condition of Experiment 1 was harmed by the unusual kind of noun categories they were taught. Ten children visited the laboratory nine consecutive weeks but did not receive any noun category training. Children participated in weekly artificial noun learning task practice trials as well as the full artificial noun learning task every third week. As in Experiment 1, the children’s vocabulary was measured at the beginning, end, and one month follow-up appointments. Thus, this experiment provides a measure of the typical vocabulary development of a matched set of children who repeatedly visit the laboratory and participate in the artificial noun learning task.

#### Methods

**Participants.** Ten children between 15- and 21- months-of-age participated (mean 18m 2d, range 15m 9d to 21m 7d). Children were recruited from the child participants file at Indiana University and contacted by phone. All children were learning English as their first language. Children were matched to children from Experiment 1 such that the mean age and vocabulary across experiments did not differ. All children received a small prize each experimental visit and copies of experimental videotapes and T-shirts at the completion of the study.

**Materials.** The same eight sets of artificial noun learning stimuli and 20 sets of practice stimuli used in Experiment 1 were used.

**Procedure.** The procedure was the same as Experiment 1 with the exception that children were not taught any noun categories.

#### Results and Discussion

The key result is pictured in Figure 5. There were no significant differences in the productive vocabularies of children from this experiment and children from the Material Nouns condition of Experiment 1. Thus, teaching children twelve names for non solid substances in categories well organized by material substance did not harm the vocabulary development of children in the Material Nouns condition of Experiment 1. And, thus, teaching children twelve names for solid objects in categories well organized by shape in Shape Nouns condition of Experiment 1 did accelerate their vocabulary growth.

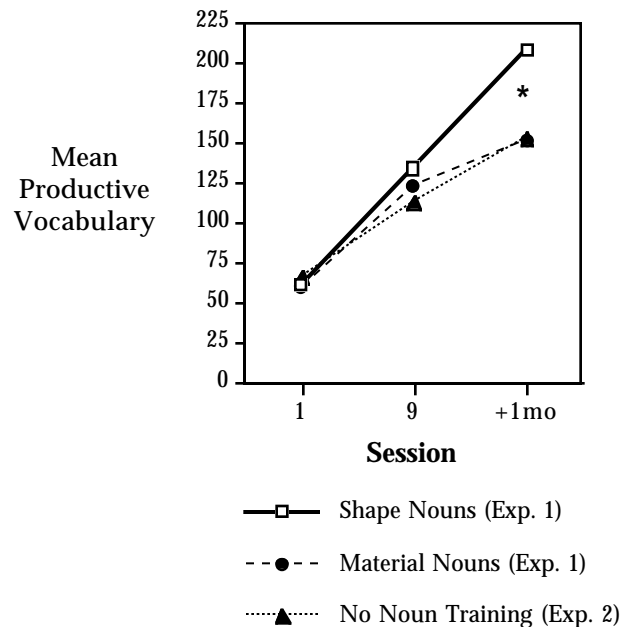


Figure 5. Mean productive vocabulary of children in Experiment 2 (no noun training). Data from Experiment 1 are included for comparison

#### Conclusions

Results from the present experiments comparing early vocabulary growth and the attentional biases seen in laboratory artificial noun learning tasks suggest a clear and sensible developmental story. Early in vocabulary development, children learn many names for solid objects in categories well organized by shape. This learning changes children: they begin to attend to shape when learning novel names in the context of novel solid objects. Thus, it appears that the attentional biases young children demonstrate in artificial noun learning tasks are a generalization across the nouns that they have already learned. In fact, teaching very young children, children who do not yet demonstrate

systematic attentional biases in artificial noun learning tasks, names for solid objects in shape-based categories teaches them a generalizable shape bias. This in turn, promotes the more rapid learning of other words.

There are two important questions that remain unanswered by these results. First, why is there such an advantage for solid-shape-based categories over non solid-material-based categories? One possibility is that this a reflection of the structure of the language children hear (but see Sandhofer, Smith, & Luo, 1999).

The second unanswered question is how do children ever get a material-bias? We know that by three-years-of-age children reliably demonstrate a bias to attend to material substance when generalizing a novel name for a novel non solid substance. However, this bias is not reliable before 30-months-of-age (Samuelson & Smith, 1999). The developmental story for the material-bias may be the same as that for the shape-bias, just more protracted in time. That is, while children are rapidly learning names for solid things, they are also encountering a smaller number of non solid substances and their names. Each of these few substances may have to be individually learned, as an exception, without the boost from past learning given to names for solid objects. But, as vocabulary grows, children may learn enough names for non solid substances that the correlation between the perceptual cues of non solidity and naming by substance cohere to form a generalizable expectation about how substances are named.

### **Acknowledgments**


The research presented in this paper was supported in part by NIMH grant F31MH12069-02. Thanks go to John Spencer for comments on an earlier draft.

### **References**

- Dickinson, D. K. (1988). Learning names for material: Factors constraining and limiting hypotheses about word meaning. Cognitive Development, *3*, 15-35.
- Imai, M., & Gentner, D. (1997). A cross-linguistic study of early word meaning: universal ontology and linguistic influence. Cognition, *62*, 169-200.
- Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. Cognitive Development, *9*(1), 45-75.
- Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. Cognitive Development, *3*, 299-321.
- Markman, E. M. (1992). In M. R. Gunnar & M. Maratsos (Eds.) Constraints on word learning: Speculations about their nature, origins, and domain specificity. Hillsdale, NJ: LEA.
- Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category organization and syntax correspond? Cognition, *73*(1), 1-33.
- Sandhofer, C., Smith, L. B., & Luo, J. (1999). Counting nouns and verbs in the input: Differential frequencies, different kinds of learning. Journal of Child Language.
- Soja, N. N. (1992). Inferences about the meanings of nouns: The relationship between perception and syntax. Cognitive Development, *7*, 29-45.
- Soja, N. N., Carey, S., & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meaning: Object terms and substance terms. Cognition, *38*, 179-211.

# Phrases as Carriers of Coherence Relations

Holger Schauer and Udo Hahn

 Computational Linguistics Division

Freiburg University

D-79085 Freiburg, Germany

<http://www.coling.uni-freiburg.de/>

## Abstract

Coherence relations have mainly been studied as a mechanism for the representation of text structure based on the analysis of clauses and larger text fragments. A closer look at textual data reveals, however, that adjuncts, typically cued by prepositions, also have a coherence establishing function. We discuss empirical evidence for this claim, and outline a framework that integrates the semantic interpretation and recognition of coherence relations covert in prepositional phrases.

## Introduction

Single sentences encode one or more propositions, the semantic content of an utterance. When sentences are grouped together to form a text, this does not just constitute a bag of such propositions. Rather texts are characterized by the property of being coherent.

The glue out of which coherent texts are made is typically attributed to so-called *coherence relations*. Basically, these relations link low-level propositions by causal connections, motivational links (e.g., relating a goal to a sequence of actions intended to accomplish that goal), property descriptions, and argumentative roles. This linkage between several propositions is crucial not only for simple fact retrieval from memory but also for other high-level cognitive tasks. Hence, making coherence relations available lies at the heart of any cognitively plausible approach to modeling human text comprehension and automatic text understanding, as well.

Given the importance of coherence relations for adequate text understanding, the question arises how these relations can be determined by explicit criteria and how this may be achieved by automated systems. The currently dominating approach derives coherence relations directly from particular cue words (i.e., sentence connectives such as ‘*because*’, ‘*alternatively*’, etc. [Knott and Dale, 1994, Marcu, 1998]).

In line with one of the most prominent approaches to coherence, Rhetorical Structure Theory [Mann and Thompson, 1988], such approaches typically take clauses as the elementary coherence bearing units, ignoring the role of more smaller units, i.e., phrases. We will argue that such *interclausal* coherence analysis should be complemented by an analysis of *intraclausal* coherence, in order to obtain more accurate results, both with respect to completeness as well as with respect to correctness of the analyses performed.

When *phrases* are considered as the origin or target of coherence relations, it becomes evident that intraclausal coherence relations are explicitly cued (by prepositions or adverbs), but they also depend on implicit inferences at the semantic level, with references to the underlying common-sense or domain knowledge. While this finding coincides with commonly held views in the cognitive science community [Black, 1985, Meyer, 1985], usually no concrete specifications are supplied for how to compute coherence relations under these assumptions. In this paper, we will first present empirically supported arguments for phrases as the smallest units of coherence analysis, and then discuss explicit nonlexical, i.e., inferential criteria for deriving coherence relations from them.

## Arguing for Phrases as Discourse Units

### Intraclausal Coherence Phenomena

The main claim we make is that coherence relations not only have to be addressed at the *interclausal* but also at the *intraclausal* level of discourse analysis. Unless this finer grain size for discourse units is chosen, we will argue in the following that some coherence relations will not be identified at all, or some of them will be identified but are invalid. Accordingly, we will treat at least some phrase types, *viz.* prepositional and adverbial phrases, as discourse units. Consider the following example:<sup>1</sup>

- (1) a. *Mit dem P6LXZ-A wird Elitegroup aber kaum neue Kunden gewinnen.*  
[With the P6LXZ-A – Elitegroup will hardly attract new customers.]
- b. *Mit einem PCI-Slot bietet das Motherboard zu wenig Platz für Erweiterungen.*  
[With one PCI slot – the motherboard provides not enough space for extensions.]

A straightforward coherence analysis with relations from Rhetorical Structure Theory [Mann and Thompson, 1988]<sup>2</sup> takes (1-b) as a single unit and links it to (1-a), probably via an

<sup>1</sup>In the translations, important phrases keep the syntactic position of the original German sentences and are therefore separated with dashes.

<sup>2</sup>Relations referring to Rhetorical Structure Theory (henceforth referred to as RST) will appear *emphasized* and *Capitalized*.

Evidence	The Satellite, the not so important unit, provides evidence for the situation in the important unit, the Nucleus.
Explanation	The Satellite, which is typically independent of the will of an animate object, explains the Nucleus.
Cause	The situation in the Satellite causes the situation in the Nucleus.
Interpretation	The situation presented in the Satellite interprets (presents a different perspective on) the Nucleus and constitutes the opinion of the writer of the situation in the Satellite.
Means	The Satellite explains the means by which the Nucleus was done.

Table 1: Relations from Rhetorical Structure Theory.

*Evidence* relation, see Table 1.<sup>3</sup> Paraphrasing sentence (1-b) reveals, however, a plausible decomposition into two basic discourse units:

- (2) a. *The motherboard has but one PCI slot,*  
b. *so it does not provide enough space for extensions.*

Obviously, (2-a) gives an *Explanation* for (2-b).<sup>4</sup> From a methodological point of view it cannot be justified to analyze Sentence (2) as being composed of two elementary units, while the prepositional phrase “*with one PCI slot*” should be an indistinguishable part of the whole Sentence (1-b).

Besides *missing* essential coherence relations by not looking at phrases as discourse units, we also have indications that even *wrong* analyses may result. Consider the following sentences:

- (3) a. *Floptical Disks lassen sich nicht wie Festplatten ansprechen.*  
*[Floptical disks cannot be addressed in the same way as ordinary hard disks.]*  
b. *Diese Beschränkung ist aufgrund technischer Unterschiede notwendig.*  
*[This restriction is – because of technical particularities – necessary.]*

One might argue, granting the interpretative force of ‘*because of*’, that (3-b) gives a *Cause* for (3-a). On a closer look, however, this seems to be mistaken, because (3-b) can be said to *Interpret* (3-a). Its main assertions consist of an assessment of (3-a) as being a “*restriction*” and as being “*necessary*”. Obviously, the embedded prepositional phrase (“*because of* . . .”) specifies just the *Cause* for the necessity of the restriction, and is not related to sentence (3-a).

### Criteria for Phrases as Discourse Units

Given that, on the one hand, at least some phrases should be analyzed as discourse units in their own right and that, on the other hand, certainly not all of them figure as discourse units, the question arises which criteria should be set up in order to single out true candidates for discourse units from spurious

<sup>3</sup>In order to avoid a lengthy introduction to RST, the definitions are taken from the manual coming with the tool that we used for our analyses [Marcu et al., 1999]. It makes available an extension of the original RST relations [Mann and Thompson, 1988].

<sup>4</sup>This analysis reflects the impact of the cue word “*so*” in (2-b). More generally, whenever an implicit coherence relation can be made explicit by a paraphrase incorporating a specific cue word, then this coherence relation is always assumed to hold [Martin, 1992, p.184].

ones. While [Grote et al., 1997] recognize that “prepositional phrases are the most compact form” to establish a coherence relation, [Marcu et al., 1999] are among the first who propose to consider those phrases as elementary discourse units that “are unequivocally the nucleus or the satellite of a rhetorical relation that adds some significant information to the text.” However, the restrictions provided by this criterion proved to be too liberal for the choice of possible candidates.

Focusing on the role of prepositional phrases (PPs) in our paper, we propose a mix of two criteria. First, the syntactic criterion requires only those PPs to be candidates for discourse units, which are not syntactically mandatory *complements* of a governing syntactic head, for which we assume a subcategorization frame or a valency list. Phrases which do not match such a schema of their governing syntactic head are syntactic elements we refer to as *adjuncts*.

For example, the PP starting Sentence (1-b), “*with one PCI slot*”, figures as an adjunct. It gives optional information, since the remainder still forms a complete grammatical sentence, “*the motherboard provides not enough space for extensions*”. This stands in contrast to example (4), which contains a true complement:

- (4) *We have to stop pointing our fingers at these kids, he said.*

In Sentence (4), neither should the PP “*at these kids*” be treated as a discourse unit, nor should any other mandatory phrase, such as the subject “*we*”.

At the semantic level we formulate the second major criterion. It is based on the assumption that semantic specifications of lexemes, independent of the specific semantic theory one subscribes to, are confined to “typical” properties, e.g., events are characterized by agents, patients, instruments, locations, time frames, etc. Since any straightforward semantic interpretation must account for these attributes, they should not be part of analyses targeting on coherence relations. Whenever nontypical, unpredictable information pieces have to be accounted for, coherence relations may capture their value-adding semantics. Therefore, only those PPs should be considered as discourse units

- whose straightforward semantic interpretation is precluded because they refer to nontypical properties;
- or whose semantic interpretation partially refers to typical properties, but the intended meaning is not fully covered by them; only additional computations – inferences taking the preliminary semantic interpretation as a starting point – completely account for the intended meaning.

We will illustrate the main criteria which determine whether a PP should be treated as a discourse unit or not by contrasting the sample sentences (1-a) and (1-b). In Sentence (1-a), the PP specifies an instrument for attracting new customers. As it seems entirely reasonable to consider “instrument” as a typical property of “attraction” events, this example should straightforwardly be dealt with by standard semantic interpretation — the conceptual correlate of *P6LXZ-A* will be assigned as the value of a corresponding “instrument” attribute. In particular, this analysis need not take recourse to any notion of coherence relation, although the proponents of RST might consider a *Means* relation as being appropriate.

This typicality consideration does not carry over to an “explanation” of events, which is our interpretation of “with one *PCI slot*” from Sentence (1-b). Rather than being missed at the representational level, accounting for this information adds valuable, ‘heavy’ knowledge. Such a relation, however, can only be computed by additional inferences relating to the underlying domain knowledge base.

### From Prepositional Phrases to Coherence Relations

We now briefly sketch a coherence analysis based on the considerations discussed in the previous section. To make this discussion more concrete, it is embedded in the framework of SYNDIKATE, a text analysis system under development in our lab [Hahn and Romacker, 1999b]. After being submitted to a syntactic analysis the dependency graph for Sentence (1-b) (cf. Figure 1) contains a prepositional adjunct (ppadj) subgraph which holds the phrase “Mit einem *PCI-Slot*”. (This analysis results from the valency specification for the main verb “*bietet*”.) In order to compute a semantic interpretation for Sentence (1-b) (assuming the framework of description logics [Woods and Schmolze, 1992]), the conceptual correlates of its content words are checked for role compatibility.

In this case, the major interpretation constraints derive from the main verb “*bietet*” (*provide*) which is represented as the concept PROVIDE (cf. Figure 2). It has three major conceptual roles, PROVIDE-PATIENT, PROVIDE-CO-PATIENT, and INSTRUMENT. The PROVIDE-PATIENT and PROVIDE-CO-PATIENT roles can be filled by some instance of MOTHERBOARD and SIZE, respectively, in the semantic interpretation phase. This causes conceptual interpretation processes to be triggered linking SIZE and MOTHERBOARD (cf. Figure 2) via the role SIZE-OF.

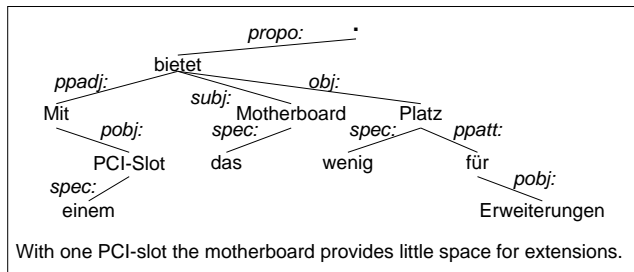


Figure 1: Dependency Analysis for (1-b)

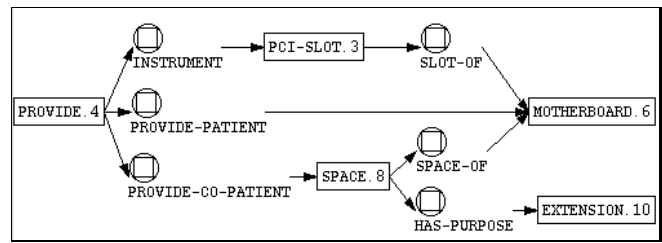


Figure 2: Semantic Interpretation for (1-b)

Focusing on the analysis of the PP, each preposition specifies *semantic constraints*, see [Hahn and Romacker, 1999a]. In the case of “*mit*” (*with*) they allow an interpretation of the dependency relation *ppadj* in terms of the conceptual INSTRUMENT role, so the corresponding role of SHOW-FEATURE is filled with *PCI-SLOT* during semantic interpretation. Conceptual interpretation, in addition, triggers the computation of a specialization of the PART-OF relation (SLOT-OF) between *PCI-SLOT* and *MOTHERBOARD*.

At this stage, we check whether the preposition might give rise to the computation of coherence relations. Corresponding *discourse constraints* of a preposition specify a set of possible coherence relations it may signal. These constraints were determined empirically, see Section (4). The constraints of permitted coherence relations are checked, taking the already computed semantic interpretation as a starting point. For “*mit*” (*with*) an *Explanation* may be signaled whenever the filler of the INSTRUMENT role stands in a PART-OF relation to the PROVIDE-PATIENT. As SLOT-OF is one of the subroles of the PART-OF relation, an *Explanation* relation is established.

Figure 2 also shows a PURPOSE relation linking instances of PHYSICAL-SIZE and EXTENSION that is due to semantic interpretation, in line with considerations which will be presented in the next section.

### Evaluation of Coherence Data

The basic claim we try to back up by empirical analysis is that focusing on intraclausal coherence leads to more adequate analyses, with respect to both completeness and correctness. In the following we will set out to validate the principal assumptions and criteria and not their implementation. For this task it is necessary to a) closely consider how many and which PPs can be seen as discourse units in their own right (i.e., checking the proposed criteria), b) how many of them have been missed in mainly clause-based analyses, and c) how many of these analyses could be judged as incorrect (similar to example (3)).

### Distribution of Prepositional Phrases in the Corpora

The textual data for our study were taken from two sources – a German-language corpus of test reports from the information technology domain (31 texts, with approximately 7,700 text tokens), and a small set of English texts from the MUC corpus [MUC-6, 1995] (9 texts, with approximately 5,100 text tokens) for comparison purposes.

For our empirical study we used RSTTOOL, a workbench for annotating texts in terms of their underlying coherence relations. The tool makes available an extension of the set of original RST relations. Both the tool and the English texts were kindly supplied by D. Marcu, see [Marcu et al., 1999].

The English texts were already analyzed and contained 795 discourse units connected by 379 relations. We re-analyzed these texts only with regard to prepositional phrases, modifying the original discourse analyses where appropriate. As the German texts were all analyzed with such a focus, we provide the distribution of units and relations in the next subsection.

Our analyses were performed in joint work by one of the authors and one student. During the discourse annotations, for each new clause to be segmented and related, we first determined the syntactic role of prepositional phrases, i.e. whether an identified PP should be seen as an adjunct or mandatory complement of its governing head. Next, when a coherence relation was unequivocally identifiable, the PP was taken as an elementary discourse unit and related with the coherence relation. As a result, we determined for each preposition the set of coherence relations it may give rise to.<sup>5</sup> Otherwise, we just recorded its likeliest interpretation. Obviously, the annotators needed to know about the hypothesis that (prepositional) phrases might trigger coherence relations. Therefore, the data presented below needs to be validated further.

Overall, we determined a total of 611 PPs in the German and 501 PPs in the English corpus. Table 2 lists their syntactic distribution, distinguishing between adjuncts and complements. The leftmost column indicates the syntactic head of the PP, either a nominal or verbal phrase, or an adjective/adverb.

	Adjuncts		Complements	
	German	English	German	English
NP	192	98	60	154
VP	176	128	159	109
Adj	10	5	14	7

Table 2: Syntactic Distribution of PPs in the Corpora

Distinguishing certain and dubious judgements, Table 3 shows the distribution of PPs that were solely analyzed by semantic interpretation, i.e., either no coherence relation could be determined or a semantic interpretation seemed entirely sufficient. We found that in those cases in which a *Means* or *Manner* relation might be used, the interpretation of the PPs just amounted to the assignment of values to reasonable and typical properties, see Example (1-a). Hence we felt that these cases should be dealt with by proper semantic interpretation and not be counted as coherence relations at all, just like locative/spatial and temporal information.

With regard to *Attribute/Restriction*, we found that many PPs that are adjuncts of NPs can be interpreted as specifying

<sup>5</sup>This set can then be used to specify the discourse constraints mentioned in the previous section.

attributes (such as “*the Matrox Millenium graphics card with 4 MByte SDRAM*”) or as stating restrictions for the interpretation of the NP (such as “*a computer with a Pentium is fast enough*”, where the PP picks a specific set of “*computers*”).

Interpretation	German		English	
	Cert.	Dub.	Cert.	Dub.
Locative/Spatial	37	16	25	0
Temporal	16	0	31	1
Means	44	3	4	0
Manner	22	0	4	1
Attribute/Restriction	205	0	226	0
Others	181	1	167	0

Table 3: PPs Not Considered as Discourse Units

Those cases that could neither be addressed by one of the given categories nor be treated as discourse units are listed as *Others*. Mostly these are cases in which the PP is a mandatory complement whose preposition has an almost idiomatic, at least a highly collocational status, see example (4). These phenomena are more adequately dealt with by lexicalized encodings covering the particular reading of the preposition rather than being treated by the general interpretation mechanism for prepositions.

### PPs as Discourse Units in the Corpora

Table 4 summarizes the interpretation of PPs in terms of true coherence relations. For the German texts, we found 66 cases for which coherence relations were unequivocally identifiable, plus 20 dubious cases. In 63 cases, the prepositional phrase appears in the middle of a clause. In this case, two units result from the remainder of the clause that need to be related by an artificial *Same-Unit* relation. Overall, the 66 identified PPs are responsible for 129 relations, including 63 *Same-Unit* relations. For the German texts, a total of 1713 units connected by 869 relations were identified. This means that 14.8% of coherence relations were of the intra-clausal type.

Our re-analyses of the English texts consisted only of additions and modifications of coherence relations due to PPs. This results in 884 units connected by 421 relations. Overall, 40 PPs give unequivocally rise to 51 coherence relations, plus 2 dubious cases. The 40 certain cases of coherence bearing PPs account for 12.1% of the coherence relations.

From those 40 PPs we considered as discourse units in the English texts, only 3 phrases were also analyzed by Marcu. This indicates that the common focus on clauses and larger fragments tends to provoke a certain analytical bias, just as we expected. So, the completeness of coherence analysis seems to benefit from the focus on adjuncts.

With regard to the syntactic criterion, almost all certain cases of discourse units (61 out of 66 in the German texts, 39 out of 40 in the English texts) are due to PPs that we judged as being adjuncts. In contradistinction, most of the dubious cases (15 of 20 for the German texts, 0 out of 2 for the English texts) coincide with the PP in a syntactically mandatory

Relation	German				English			
	Adjuncts		Complements		Adjuncts		Complements	
	Certain	Dubious	Certain	Dubious	Certain	Dubious	Certain	Dubious
Analogy	0	0	0	0	1	0	0	0
Attribution	5	0	0	0	1	0	0	0
Background	0	0	0	0	3	0	0	0
Cause	0	0	0	0	2	0	0	0
Circumstance	5	0	0	1	11	0	0	0
Condition	25	1	1	5	8	0	0	0
Consequence	0	0	0	0	2	2	0	0
Elaboration	0	0	0	0	2	0	0	0
Evaluation	0	0	0	0	1	0	0	0
Explanation	6	0	1	0	1	0	0	0
Purpose	20	4	3	9	5	0	0	0
Reason	0	0	0	0	2	0	1	0
$\Sigma$	61	5	5	15	39	2	1	0

Table 4: Prepositional Phrases Treated as Discourse Units

position. So the distinction between adjuncts and complements can be seen as a valid indicator for phrases that can be analyzed as discourse units.

With regard to the second criterion, it is necessary to explain the dubious cases. These often coincide with syntactically mandatory complements. As a result, it is often not clear whether the PP should be analysed semantically, although an interpretation in terms of a coherence relation would be possible. As an example, consider the next example, in which the prepositional phrase could be analysed as stating a *Purpose* for the graphics card.

- (5) *Für die nächste Generation an Spielen sind Grafikkarten wie die Spea V7 nötig.*  
 [For the next generation of games – graphics cards such as the Spea V7 are required.]

In contradistinction, for those PPs we unequivocally considered as discourse units, an interpretation solely in semantic terms is usually hard to imagine, see Example (1).

Commenting on the correctness of the original English analyses, we found no cases of errors caused by overlooked PPs, contrary to our expectation. This may be explained by the fact that those cases in German are mainly triggered when the phrase occurs inside a clause – e.g., example (3-b). We found no corresponding example in the English data.

Finally, commenting on the quantitative distribution of coherence relations in Table 4, the large number of *Condition* and *Purpose* relations might largely be attributed to the chosen domain (information-technology). In this domain, judgments are often valid only under certain assumptions and conditions. Also, nearly all actions serve some purpose and are evaluated against it. One might be challenged then to treat *Condition* and *Purpose* as “typical” in this domain; hence they should probably even be treated by the standard semantic interpretation (as already assumed in Figure 2).<sup>6</sup>

<sup>6</sup>See [Linden and Martin, 1995] for an account of the *Purpose*

## Related Work

The notion of coherence relations is dealt with by a variety of approaches — structural ones in which linear text fragments are bracketed and organized in discourse trees by rhetorical relations [Mann and Thompson, 1988], logical ones in which metapredicates provide the inferential basis for linking basic predicate-argument structures [Lascarides et al., 1992], psychological ones in which the level of micropropositions is clustered in terms of conceptually coherent macropropositions [Kintsch and Dijk, 1978, Black, 1985]. Since none of these approaches incorporate syntactic considerations into their analyses (a syntactic analysis is assumed to deliver appropriate text chunks or propositions), they are unable to account for coherence relations encoded via PPs.

There are a few attempts to incorporate the role of cue words in computational approaches to determine coherence relations. These are based on the RST framework [Marcu, 1998], logical interpretations [Hobbs et al., 1993], or extensions of sentence grammars [Webber et al., 1999]. But in these approaches, the level of intraclausal analysis is not an issue. A recent study mentions the role of PPs as carriers of coherence relations [Grote et al., 1997], but only for the purposes of text generation.

Our distinction between semantic and discourse constraints looks similar to the semantic/pragmatic distinction found in [Sanders et al., 1992, Knott and Dale, 1994]. Their distinction, however, addresses the intended effects coherence cues have on the reader, while in our work discourse constraints establish interpretations beyond ‘typicality’ limits.

Another distinction relates to the role of empirical arguments related to coherence relations. Our study deals with the *quantitative distribution* of a set of coherence relations as encoded by various PPs, while in [Sanders et al., 1992] the *plausibility* of certain coherence relations fulfilling a set of

relation in instructional texts that also acknowledges the role of intraclausal coherence.



criteria is judged by a number of subjects.

## Conclusion

We have presented an approach in which the computation of coherence relations is made dependent on the semantic interpretation of a particular class of prepositional phrases, *viz.* adjuncts. The notoriously difficult distinction between complements and adjuncts has been resolved in a pragmatic way such that the syntactic notion of complements is associated with typicality considerations at the semantic level, and, similarly, adjuncts are associated with nontypical properties.

Text interpretation then proceeds via a two-step procedure. First, proper semantic interpretation is concerned with matching parsed utterances to (conceptual) representations in the lexicon. If a match is found (i.e., complements refer to typical properties/relations), one checks, in addition, whether inferential criteria for coherence relations are fulfilled. If no match can be found, an adjunct has been determined which, by definition, constitutes a possible discourse unit and has to be checked for more specific criteria for coherence relations.

One focus of our paper was on finding empirical evidence for the claim that PPs are important at all for coherence analysis. Indeed, we have detected a significant subset of coherence relations encoded as PPs (for the English data roughly 12%, for the German data 15%). These would have been lost if a cue-phrase-only approach were followed, since prepositions cannot be considered reliable predictors of (specific) coherence relations. They would, however, also have been lost with an inference-only approach, since each preposition may signal only some coherence relations. Therefore, they do not seem to be derivable from conceptual representations alone.

Given that this argument is valid, the computation of coherence relations must incorporate both the syntactic and semantic level, as well as inference rules which determine those knowledge structures which have to be superimposed by coherence relations.

**Acknowledgements.** Holger Schauer is a member of the Graduate Program *Human and Machine Intelligence* at Freiburg University, Germany, funded by DFG.

## References

- [Black, 1985] Black, J. (1985). An exposition on understanding expository text. In Britton, B. and Black, J., editors, *Understanding Expository Text*, pages 249–267. Hillsdale, NJ: L. Erlbaum.
- [Grote et al., 1997] Grote, B., Lenke, N., and Stede, M. (1997). Ma(r)king concessions in English and German. *Discourse Processes*, 24(1):87–118.
- [Hahn and Romacker, 1999a] Hahn, U. and Romacker, M. (1999a). Incrementality and locality of language comprehension. In *Proc. Cognitive Science 99*, pages 202–207.
- [Hahn and Romacker, 1999b] Hahn, U. and Romacker, M. (1999b). SYNDIKATE – generating text knowledge bases from natural language texts. In *IEEE SMC'99*, volume 5, pages V-918–V-923. IEEE Press.
- [Hobbs et al., 1993] Hobbs, J., Stickel, M., Appelt, D., and Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, 63(1/2):69–142.
- [Kintsch and Dijk, 1978] Kintsch, W. and Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Rev.*, 85:363–394.
- [Knott and Dale, 1994] Knott, A. and Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.
- [Lascarides et al., 1992] Lascarides, A., Asher, N., and Oberlander, J. (1992). Inferring discourse relations in context. In *Proc. ACL'92*, pages 1–8.
- [Linden and Martin, 1995] Linden, K. V. and Martin, J. H. (1995). Expressing rhetorical relations in instructional texts: A case study of the purpose relation. *Computational Linguistics*, 21(1):29–59.
- [Mann and Thompson, 1988] Mann, W. and Thompson, S. (1988). Rhetorical Structure Theory: toward a functional theory of text organization. *Text*, 8(3):243–281.
- [Marcu, 1998] Marcu, D. (1998). *The Rhetorical Parsing, Summarization and Generation of Natural Language Texts*. PhD thesis, U Toronto.
- [Marcu et al., 1999] Marcu, D., Amorrortu, E., and Romera, M. (1999). Experiments in constructing a corpus of discourse trees. In *Proc. ACL'99 Workshop 'Standards and Tools for Discourse Tagging'*, pages 48–57.
- [Martin, 1992] Martin, J. (1992). *English Text*. Philadelphia: J. Benjamins.
- [Meyer, 1985] Meyer, B. (1985). Prose analysis: purposes, procedures, and problems. In Britton, B. and Black, J., editors, *Understanding Expository Text*, pages 11–64. Hillsdale, NJ: L. Erlbaum.
- [MUC-6, 1995] MUC-6 (1995). *Proceedings of the 6th Message Understanding Conference*. Columbia, Maryland, November 6-8, 1995. San Mateo, CA: Morgan Kaufmann.
- [Sanders et al., 1992] Sanders, T., Spooren, W., and Noordman, L. (1992). Towards a taxonomy of coherence relations. *Discourse Processes*, 15:1–35.
- [Webber et al., 1999] Webber, B., Knott, A., Stone, M., and Joshi, A. (1999). Discourse relations: A structural and presuppositional account using lexicalised TAG. In *Proc. ACL'99*.
- [Woods and Schmolze, 1992] Woods, W. and Schmolze, J. (1992). The KL-ONE family. *Computers & Mathematics with Applications*, 23(2/5):133–177.

# Syntactic Priming in German Sentence Production

Christoph Scheepers (chsc@coli.uni-sb.de)

Department of Computational Linguistics  
Saarland University, 66041 Saarbrücken, Germany

Martin Corley (Martin.Corley@ed.ac.uk)

Department of Psychology and Human Communication Research Centre  
University of Edinburgh, Edinburgh EH8 9JZ, UK

## Abstract

Current theories of language production tend to differentiate between a (syntactic) *functional* level and a (surface) *positional* level in the generation of sentences, where functional selection precedes and constrains positional processing. In this paper, we present evidence from a syntactic priming study in German, where position, function, and type of constituent are orthogonally specified for monotransitive and ditransitive verbs. In contrast to findings for English (in which these factors are confounded) we show that previous generation of a ditransitive structure can *inhibit* the production of a further ditransitive when the order of potential arguments differs between prime and target. Our results suggest that positional processing must at the least interact with functional processing in production, and point to the importance of cross-linguistic evidence in the formation of models of language processing.

*Syntactic Priming* is the name given to the tendency that people have to re-use syntactic structure that they have just generated. For example, Bock (1986) demonstrated that, having read aloud a sentence such as *The rock star sold some cocaine to an undercover agent*, participants are more likely to describe a picture with a phrase such as *The girl handed a paintbrush to the man* rather than with the alternative *The girl handed the man a paintbrush*. Current interpretations of these findings tend to emphasise a functional level of sentence production, at which syntactic information (such as subcategorisation properties of verbs) is specified and syntactic roles (such as subject or object) are assigned. The eventual positions of constituents in the utterance are determined by subsequent processes which take as their input the representations built at the functional level (e.g., Bock & Levelt, 1994).

Evidence supporting the existence of a functional level in production has been found in a series of studies by Bock and colleagues (Bock, 1986, 1989; Bock & Loebell, 1990).

- (1a) The secretary baked a cake for her boss.
- (1b) The wealthy widow drove her Mercedes to the church.
- (1c) Susan brought a book to study.

Taking *X primes Y* to mean that utterance *Y* is more likely to be produced by participants who have just produced utterance *X*, it has been demonstrated that *The girl handed a paintbrush to the man* is primed by (1a) (individual lexical items

do not affect priming) and by (1b) (the priming of a prepositional phrase is not affected by its thematic role). However, (1c) does not prime *The girl handed a paintbrush to the man* (prosodic similarity does not affect production).

Starting with this evidence, Pickering and Branigan (1998, henceforth P&B) have recently argued for the specification of syntactic verb information within the production lexicon. Using a localised network model of the production lexicon derived from Roelofs (1992, 1993) they argue that lemma nodes for verbs are linked to additional nodes representing syntactic features such as tense and aspect. These nodes in turn link to ‘lexeme nodes’ on a separate stratum, which represent potential lexical forms of verbs. If the verb lemma <GIVE> and both a past tense node and a perfective aspect node are active, a likely articulation through the lexemic level would be *gave*. The syntactic feature nodes are unique, such that any verb which can be expressed in the past tense is linked to the same past tense node as is <GIVE>. Importantly, P&B also assume that verb lemmas are linked to ‘combinatorial’ nodes which express the constructions in which a verb can be used. <GIVE> would have links to (at least) two combinatorial nodes, representing ‘NP NP’ (give the dog a bone) and ‘NP PP’ (give the bone to a dog) combinations. It is worth noting that (at least) two types of information traditionally described as subcategorisation information are combined by these nodes, since they encode not only the types (syntactic category and case) of arguments used, but also the number of arguments (i.e., the verb’s valence).

Using standard assumptions about decaying activation,<sup>1</sup> the priming of *The girl handed the paintbrush to a man* by *The rock star sold some cocaine to an undercover agent* is accounted for by suggesting that the ‘NP PP’ node retains some activation and thus reaches threshold more easily when making the second utterance. P&B provide support for this model by adopting a novel methodology (see also Branigan, Pickering, Liversedge, Stewart, & Urbach, 1995), in which participants provide written completions for partial sentences. The prime sentences are pragmatically constrained such that the most likely completion is of a given form (e.g., *The racing*

---

<sup>1</sup>P&B’s model deviates from more traditional activation models in that some links, as well as nodes, retain activation over time. However the detail of the model has no bearing on the functional/positional dichotomy to which we address ourselves in this paper.

*driver gave the torn overall* \_\_\_\_ vs. *The racing driver gave the helpful mechanic* \_\_\_\_) but the target sentences end after the matrix verb (e.g., *The patient showed* \_\_\_\_). In line with P&B's predictions, subjects are more likely to produce target sentences with the same syntactic structures as the primes. Moreover, the priming effect becomes stronger when the verb is repeated between prime and target (when activation from both lemma and combinatorial nodes is assumed to contribute to the effect). Finally, syntactic priming is unaffected by differences between prime and target in the verb's tense, aspect, or number, supporting the idea that syntactic feature information is separate from the representations involved in syntactic priming (i.e. lemmas and combinatorial nodes).

However, P&B's evidence lends itself to alternative interpretations. Firstly, it might be possible to account for their findings in terms of *positional* rather than *functional* processing: in English, the positions of the two arguments of a ditransitive verb such as *give* are confounded with the different syntactic structures that are required to realise each possible sequence ('NP NP' vs. 'NP PP'). The same line of reasoning applies to Bock's research: it might even be argued that the irrelevance of thematic role assignment to PP-priming militates against a view where constituents are stipulated at a lexical (argument structure) level, and for a model in which particular constituents like 'NP' or 'PP' are more likely to be reproduced 'in the same linear position'. Evidence for the view that the order of constituents can be primed (where the underlying syntactic representation remains constant) has been recently demonstrated in Dutch (Hartsuiker, in preparation; Hartsuiker & Kolk, 1998b).

Secondly, P&B's experimental findings might be accounted for if they simply reflected the propensity of the production system to reuse particular types of syntactic constituents (for example, PPs). This hypothesis has the attraction of providing a more natural explanation for the priming of *The girl handed a paintbrush to the man* by sentences including optional arguments or modifiers (Bock & Loebell, 1990).<sup>2</sup> Because the experiments do not contain a baseline condition, it is impossible to tell whether both 'NP NP' and 'NP PP' primes have an effect on the sentence produced, or whether, for example, it is only an 'NP PP' prime which affects the standard distribution of responses (see Hartsuiker & Kolk, 1998a, for a similar argument). If previous findings can be accounted for by a mechanism which is simply more likely to use a particular type of constituent, then people may be equally likely to produce sentences where the verb has a different number of arguments but a particular constituent is reproduced, provided that there are no constraints on the verb produced in the target sentence. To make this concrete, consider a situation in which the prime sentence is *The man gave a toy to the child* (which has an 'NP PP' form). If syntactic priming simply reflects the probability of reusing particular constituents (say, a PP), then in the absence of a constraining verb in the target sentence, people may be as likely to produce *The man sang in the bath* as *The man put the soap in its*

*holder* (since both contain PPs).

German provides an interesting opportunity to explore the issues outlined above more fully. In German, ditransitive verbs such as *geben* (to give) take two case-marked arguments: the object given has accusative case, and the recipient has dative case. Importantly, the order of these arguments is (almost) arbitrary, so that *Ich gab dem Mann das Buch* and *Ich gab das Buch dem Mann* are both translated as "I gave the man the book". Therefore, it is possible to explore priming effects at the positional level (as in studies on Dutch: Hartsuiker, in preparation; Hartsuiker & Kolk, 1998b). A second feature of German is that *monotransitive* verbs (which take a single object) can subcategorise for either accusative or dative case objects, providing an opportunity to test whether certain types of arguments (designated by case rather than syntactic category) are reused over consecutive trials. Taken together, this results in a system where number of arguments (1 or 2), type of arguments (accusative or dative NP), and (for ditransitives) order of arguments are orthogonally specified.

## A Completion Experiment on the Internet

The aim of the current study is to exploit these features of German to provide a fuller investigation of syntactic priming, using the sentence completion method pioneered by Branigan et al. (1995) and Pickering and Branigan (1998). In this study, primes consist of ditransitives in each of the possible configurations (which we will refer to as *dat<acc* and *acc<dat*) as well as monotransitives which subcategorise for single accusative (*acc*) or dative (*dat*) arguments. As well as these four primes, we include a baseline condition (where the prime is unrelated to the type of target that can be generated, given experimental constraints). Finally, because we are interested not only in the order of arguments but also in the numbers and types of arguments generated, participants are left free to choose the verb for the target sentence fragment, in contrast to previous studies.

**Participants** The experiment was administered via the World Wide Web. Participants were recruited through advertisements in Usenet newsgroups as well as through links from other web pages. Fifty-eight participants from different regions of Germany, Austria, and Switzerland completed the experiment. All of them acquired German as their first language, and most of them (83%) were university graduates or students of different scientific subjects. Participants' average age was 28.5 years, ranging from 18 to 54 years. Thirty-six of them were male, 22 female.

**Materials** The experiment had a two-factor (5×2) design, using a syntactic priming paradigm in which both primes and targets consisted of sentence fragments for completion. Participants had to complete one of two types of target fragments

---

<sup>2</sup>P&B provide an alternative account of this finding by suggesting that 'combinatorial' nodes encode syntactic rules (such as VP ⇒ NP PP) rather than subcategorisation information.

after having completed one of five types of priming constructions.

The **targets** consisted of pairs of VP-head-final sentence fragments of the forms in (2) below (where \_\_\_ represents the missing material that was to be provided by the participant).

(2a) **accusative target** NP<sub>[nom]</sub> *hat* NP<sub>[acc]</sub> \_\_\_ *wollen*.

(2b) **dative target** NP<sub>[nom]</sub> *hat* NP<sub>[dat]</sub> \_\_\_ *wollen*.

Each target could be completed in one of two ways. The fragments could be completed using a *monotransitive* verb (i.e., a verb which takes a single object NP in the accusative or dative case for (2a) and (2b) respectively). Alternatively, the completion could consist of a second object NP followed by a *ditransitive* verb. For instance, an accusative target like *Der Mann hat den Freund \_\_\_ wollen* (The man has the<sub>[acc]</sub> friend \_\_\_ wanted, cf. (2a)) might be completed with *treffen* (to meet) which subcategorises for a single accusative object NP. Alternatively, a phrase like *seinem Kollegen vorstellen* (to introduce to his colleague) might be used, resulting in a ditransitive construction. Likewise, a dative target like *Der Mann hat dem Freund \_\_\_ wollen* (The man has [to] the<sub>[dat]</sub> friend \_\_\_ wanted, cf. (2b)) could legitimately be completed using *helfen* (to help), which takes a single dative object NP, as well as with a phrase like *seinen Kollegen vorstellen* (to introduce his colleague) as a ditransitive completion. Note that ditransitive completions of examples of the form of (2b) imply a canonical dat<acc ordering of the object NPs. For (2a), on the other hand, ditransitive completions result in less common (though acceptable) acc<dat sequences.

(3a) **acc<dat** NP<sub>[nom]</sub> *hat* NP<sub>[acc]</sub> \_\_\_ V<sub>[ppl; <dat,acc>]</sub>.  
or NP<sub>[nom]</sub> *hat* \_\_\_ NP<sub>[dat]</sub> V<sub>[ppl; <dat,acc>]</sub>.

(3b) **dat<acc** NP<sub>[nom]</sub> *hat* NP<sub>[dat]</sub> \_\_\_ V<sub>[ppl; <dat,acc>]</sub>.  
or NP<sub>[nom]</sub> *hat* \_\_\_ NP<sub>[acc]</sub> V<sub>[ppl; <dat,acc>]</sub>.

(3c) **acc** NP<sub>[nom]</sub> *hat* \_\_\_ V<sub>[ppl; <acc>]</sub>.

(3d) **dat** NP<sub>[nom]</sub> *hat/ist* \_\_\_ V<sub>[ppl; <dat>]</sub>.

(3e) **baseline** NP<sub>[nom]</sub> *war* \_\_\_ *als* NP<sub>[nom]</sub>.

The sets of five **priming** materials were constrained such that the most likely completion would be an object NP (in (3a–d)) or a comparative (in (3e)) (the latter, equivalent to the English ‘NP1 *was* \_\_\_ *than* NP2’, served as the baseline condition). Materials modelled on (3a)—where, in an equal proportion of trials, either the dative or the accusative object was missing—were constructed such that the most likely completion would result in a ditransitive construction with (non-canonical) acc<dat argument order. Condition (3b) was similar to (3a), but a canonical dat<acc ditransitive was the most likely outcome. (3c) and (3d) were most likely to be completed as monotransitive constructions with either a single accusative (3c) or a single dative (3d) object NP. A major constraining factor of the priming materials was the verbs,

which were selected on the basis of their subcat-specifications from the CELEX German Database: for (3a) and (3b), we selected strictly ditransitive predicates, like *gezeigt* (showed), which subcategorise for both a dative and an accusative object; in (3c), predicates which require a single accusative object, like *untersucht* (examined), were used; and for (3d), we chose predicates taking a single dative object like *begegnet* (came across).<sup>3</sup> The copula-verb baseline condition (3e) implied none of these verb frames.

Thirty different item-sets were generated, each comprising two target fragments (cf. (2)) and ten priming fragments (cf. (3)). There were two sentence fragments per priming condition and one sentence fragment per target condition in each item-set, so that the sentence fragments could be arranged in triplets of two primes of the same condition followed by one target. The sentence fragments used for each of the triplets were semantically unrelated.

All possible combinations of priming and target fragments were used, resulting in ten different triplets per item-set. The resulting 300 triplets were randomly allotted to ten treatments such that each treatment contained an equal number of triplets of each type. Each item-set appeared exactly once per treatment, but in a different condition than in the other treatments. A set of 90 filler fragments was also generated—these included intransitives, passives, or copula-verb constructions similar to (3e). The set of fillers was added to each treatment, resulting in a total of 180 sentence fragments per treatment.

Table 1 shows an example triplet consisting of two priming fragments of type (3a), followed by a target fragment of type (2b).

Table 1: Example material set corresponding to conditions (3a) and (2b).

<b>prime 1</b>	Die Mutter hat das Kind ___ anvertraut.
<b>prime 2</b>	Der Dekan hat ___ dem Professor vorgestellt.
<b>target</b>	Der Junge hat dem Mädchen ___ wollen.

**Procedure** Materials were presented using the WebExp experimental toolkit (Keller, Corley, Corley, Konieczny, & Todorascu, 1998).<sup>4</sup> Each sentence fragment was presented

<sup>3</sup>Some dative-object verbs in German select a form of *sein* (to be) rather than *haben* (to have) as their perfect tense auxiliary (cf. (3d)). While this kind of restriction can be useful to elicit the intended response in some of the priming constructions, it needs to be eliminated from the targets. Therefore, we used a modal auxiliary like *wollen* at the end of each target fragment, so that any type of infinitival main verb (instead of a participle) could be inserted. Informed by P&B, we considered the resulting syntactic feature differences between prime and target verbs irrelevant for the priming effects of interest.

<sup>4</sup>A demo is available at [http://www.hcrc.ed.ac.uk/web\\_exp/](http://www.hcrc.ed.ac.uk/web_exp/)

(via a web browser) in a text box, with a series of dashes representing the missing portion: participants were instructed to type one or more words into a second text box such that an *acceptable* sentence was formed from the fragment and the word or phrase that they supplied, where acceptable was defined as grammatically correct and reasonably plausible. There were no further restrictions on how participants completed the sentences, other than their being asked to avoid proper names if possible. Further instructions emphasised that participants should rely on first impressions rather than trying to create witty or original completions.

The WebExp software rotated through the ten sets of materials, so that each new participant saw a different treatment. Within each treatment, the materials were randomised such that each prime-prime-target triple was preceded by three fillers, drawn at random from the ninety available. Completion of an item (by pressing RETURN) resulted in the immediate display of the following item; participants were not able to re-inspect items or responses once they had been recorded. Responses were timed (on the participant's computer) by recording the time taken to make the first keystroke of any response, as well as the time to press RETURN at the end of a response. Completions were required for all 180 items in a given set of materials. At the close of the experiment, participants were thanked for their time and promised a debriefing once the experiment was complete (debriefs were later sent by email).

Two independent judges categorised the responses made to both prime and target fragments, recording the orders and cases of arguments, and the subcategorisation properties of the verbs chosen. The categories were later conflated into *correct* or *incorrect* for primes (reflecting whether the desired response had in fact been elicited) and into *monotrans*, *ditrans* or *other* for targets (reflecting mainly the subcategorisation properties of the verbs chosen). In the few cases where participants had selected a ditransitive verb without including an additional object NP (resulting in an 'implicit argument' construction), target responses were scored as *monotrans*. Grammatically incorrect responses (most of which included wrong case assignments) and responses involving prepositional complements were categorised as *other*.

**Analysis** Effects were examined by testing hierarchical log-linear models (see Howell, 1997, for an overview), adjusting observed cell counts to factor combinations of prime type (cf. (3)), target type (cf. (2)), completion type (*monotrans* vs. *ditrans* vs. *other*),<sup>5</sup> and either participants or items. The analyses including participants or items as random factors are reported as  $LR\chi^2_{(subj)}$  and  $LR\chi^2_{(item)}$  respectively; a further statistic,  $LR\chi^2_{(marg)}$ , refers to an analysis in which the effect itself (i.e., its constituting factor combination) serves as the saturated model, ignoring additional random factors. Technically, the first two statistics represent so-called *partial associations*, whereas the third refers to the *marginal association* of an effect. For main or simple effects only marginal associations ( $LR\chi^2_{(marg)}$ ) will be reported, as the partial associations

are redundant in these cases.

## Predictions

**Priming Effects** Due to the exploratory nature of this experiment, we will skip discussing hypothetical priming effects in favour of a discussion of the theoretical implications of the observed data at the end of this paper. Therefore, we turn our attention to predictions of baseline effects in the following section.

**Baselines** We assumed that the standard distribution of the target responses would be influenced by (at least) two factors: Firstly, the availability of different subcategorisation frames, and secondly, canonical argument ordering constraints. The former was assumed, as a rough estimation, to be a function of the relative sizes of different verb classes in German, given that participants have to choose from these to generate their responses. According to the CELEX German Database, about 63% of the 'common' verbs in German (i.e. verbs with a lemma frequency of at least 10 per million) are monotransitives requiring a single accusative object NP; 23% are ditransitives, taking both an accusative and a dative object NP; and only 4% are monotransitives subcategorising for a single dative object NP (the remaining 10% are either intransitives or verbs requiring other types of complements). Given this distribution of available verb frames (and because participants may avoid generating non-canonical orderings if possible) we predicted that *monotrans* completions should be predominant for accusative targets like (2a). For dative targets like (2b), however, we expected *ditrans* responses to be most frequent, as the set of verbs which take a single dative object NP is relatively small.

## Results

Data from seven participants were excluded from analysis: in five cases, because the proportion of *other* responses was greater than 25% (this may, in some cases, reflect dialectal variations), and in two cases, because median response latencies were extremely slow (at > 20 sec). Target data points from the remaining 51 subjects were excluded if: (a) the immediately preceding prime was categorised as incorrect; (b) the time between the onset of a response to the immediately preceding prime and the onset of the target response lay out with the participant-specific interquartile range. These criteria resulted in the exclusion of 13% of the trials from analysis. Consequently, the results reported are based on a total of 1334 data points.

The frequency of correct trials varied considerably across priming conditions (acc<dat: 68.5%; dat<acc: 88%; acc: 95%; dat: 90%; baseline: 95%), mainly reflecting the fact that participants were more reluctant to produce the intended completion in the non-canonical (acc<dat) priming condition. For the targets, there were 889 (66.6%) *monotrans*, 341

<sup>5</sup>Note that the 'dependent variable' is treated as a factor (as in a standard  $\chi^2$  test).

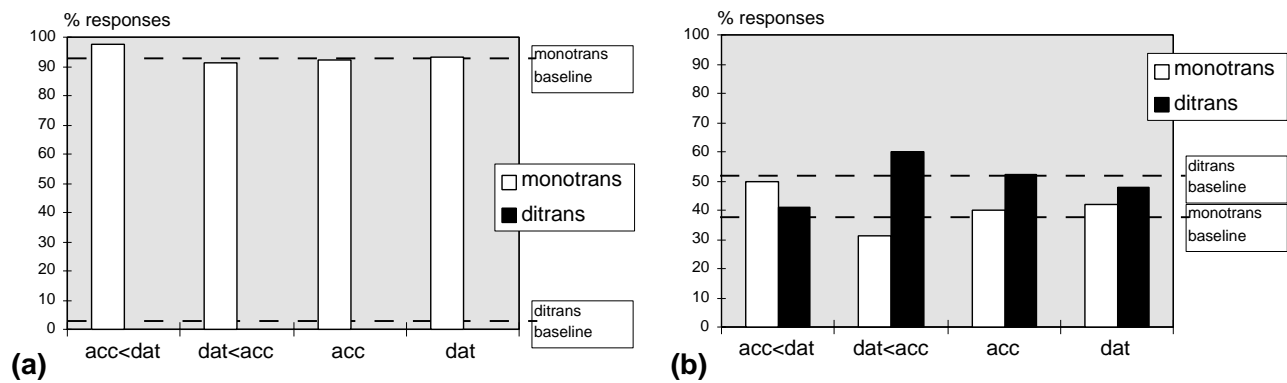


Figure 1: percentages of monotrans and ditrans completions for (a) accusative and (b) dative targets, by type of correctly completed prime.

(25.6%) *ditrans*, and 104 (7.8%) *other* responses in total. Interestingly, participants reused a verb from one of the preceding primes in less than 0.5% of the target completions.

Figure 1 shows the frequencies of *monotrans* and *ditrans* responses (in proportion to the total number of responses per condition) separately for accusative targets (Fig. 1a) and for dative targets (Fig. 1b). Dashed lines indicate baseline response rates. The remaining four prime conditions are represented by data columns.

**Baselines** In the *baseline* prime condition, the predicted biases were confirmed: for accusative targets, *monotrans* completions were clearly predominant (91%, *ditrans* = 3%, *other* = 6%); for dative targets, however, *ditrans* responses were the most frequent (52%, *monotrans* = 38%, *other* = 10%). In fact, irrespective of prime type, there was a significant overall interaction between target type and completion type ( $LR\chi^2_{(subj, item, marg)} > 500$ ;  $df = 2$ ;  $p < .001$ ) which replicates the pattern found in the baseline condition: for accusative targets, *monotrans* completions were the most frequent (93%; *ditrans* = 0.5%, *other* = 6.5%:  $LR\chi^2_{(marg)} > 600$ ;  $df = 1$ ;  $p < .001$ ); for dative targets, *ditrans* completions were the most frequent (51%; *monotrans* = 40%, *other* = 9%:  $LR\chi^2_{(marg)} > 9.950$ ;  $df = 1$ ;  $p < .002$ ).

### Priming Effects

The prime type  $\times$  target type  $\times$  completion type interaction was significant, at least by tests adjusting for subject and item variation ( $LR\chi^2_{(subj, item)} > 27.0$ ;  $df = 8$ ;  $p < .001$ ;  $LR\chi^2_{(marg)} = 15.295$ ;  $df = 8$ ;  $p = .054$ ). Unfortunately, the strong *monotrans* bias in the accusative target condition (there were virtually no *ditrans* responses) rendered any further statistical exploration in this target condition infeasible. Therefore, only the dative target condition was examined in detail. This was done by partitioning the prime type factor into ‘monotransitive’ and ‘ditransitive’ primes.<sup>6</sup>

**Dative Targets** In order to examine the effects of ‘monotransitive’ primes on the distribution of responses in the da-

tive target condition, a reduced model comprising only the *acc*, *dat*, and *baseline* conditions was generated. Testing this model revealed no significant interaction between prime type and completion type ( $LR\chi^2_{(subj, item, marg)} < 0.7$ ;  $df = 2$ ;  $p > .70$ ). Testing ‘ditransitive’ primes via a model including the *acc<dat*, the *dat<acc*, and the *baseline* condition revealed a reliable impact of prime type on completion type ( $LR\chi^2_{(subj, item, marg)} > 7.930$ ;  $df = 2$ ;  $p < .02$ ): as can be seen in Figure 1b, the tendency to produce *ditrans* completions was more pronounced after canonical *dat<acc* primes; the **reverse** tendency, i.e., to produce *monotrans* rather than *ditrans* completions, was found in the non-canonical *acc<dat* priming condition. Statistically, the proportion of *monotrans* and *ditrans* target completions clearly differed between the two ‘ditransitive’ priming conditions ( $LR\chi^2_{(subj, item, marg)} > 6.950$ ;  $df = 1$ ;  $p < .01$ ). Contrasts with the baseline condition were confirmed as statistical trends (*acc<dat* vs. *baseline*:  $LR\chi^2_{(subj)} = 1.903$ ;  $df = 1$ ;  $p < .17$ ;  $LR\chi^2_{(item)} = 4.758$ ;  $df = 1$ ;  $p < .03$ ;  $LR\chi^2_{(marg)} = 3.210$ ;  $df = 1$ ;  $p < .08$ ; *dat<acc* vs. *baseline*:  $LR\chi^2_{(subj)} = 3.638$ ;  $df = 1$ ;  $p < .06$ ;  $LR\chi^2_{(item)} = 2.058$ ;  $df = 1$ ;  $p < .16$ ;  $LR\chi^2_{(marg)} = 4.100$ ;  $df = 1$ ;  $p < .05$ ).

### Discussion

The observed data pattern (at least as established in the dative target condition) bears some interesting implications for the representation of combinatorial information in sentence production (cf. Pickering & Branigan, 1998), and may even challenge some architectural assumptions about the human language production system: it appears that subcategorisation properties of verbs *per se* (in terms of verb valence and case of arguments) are not subject to syntactic priming. This is highlighted by the fact that (a) ‘monotransitive’ primes (*acc* and *dat*) have no significant impact on the distribution of

<sup>6</sup>These analyses considered only the distributions of *monotrans* and *ditrans* completions, as the proportion of *other* responses was totally unaffected by prime type in the dative target condition ( $LR\chi^2_{(subj, item, marg)} < 1$ ;  $df = 4$ ;  $p > .95$ ).

the (dative) target completions, and (b) 'ditransitive' primes (acc<dat and dat<acc) have *facilitatory* as well as *inhibitory* effects on the relative proportions of *ditrans* to *monotrans* responses, dependent on the sequence of arguments specified in the prime. As the latter indicates, there is clear evidence for the importance of positional information in syntactic priming, comparable to recent results from Dutch (cf. Hartsuiker, in preparation; Hartsuiker & Kolk, 1998b).

With respect to representational aspects of a production model, the results could be interpreted as suggesting that combinatorial nodes in the verb lexicon encode subcategorisation information as well as information about the (canonical/non-canonical) sequencing of arguments, i.e., something similar to what is encoded in traditional context free grammar rules (cf. P&B). Unfortunately, our data remain unclear regarding the precise nature of these representations, since the accusative target condition was uninformative (due to a massive bias towards monotransitive responses in this condition): the observed ordering effects could be due to 'canonical vs. non-canonical' argument ordering (i.e., ditransitive verb-frames become more easily retrievable after canonical primes, but less easily retrievable after non-canonical primes) or to a 'match vs. mismatch' in (implied) argument order between prime and target. At this point, we leave this as a question for future research.

Our data do however greatly constrain the range of plausible architectural assumptions about sentence production. Our findings can be taken as strong evidence against a model which claims that processes at the functional level (i.e., verb retrieval and syntactic function assignment) necessarily precede, and therefore determine, positional processing, but not vice versa (e.g., Bock & Levelt, 1994). It appears that positional processing can, under certain circumstances, determine the outcome of processes at the functional level, in such a way that the ease of retrieving a ditransitive verb (in a target trial) is dependent on the argument order specified in a preceding ditransitive priming construction. Note that the retrievability of the verb (or of its corresponding subcategorisation frame) in the prime cannot account for this evidence, since strictly ditransitive prime verbs had already been presented to participants (unlike the target verbs which participants were free to choose). Thus, it must have been the linear order of the arguments that had to be produced in a correctly completed prime that affected the increased or decreased availability of a ditransitive verb frame in the target trials. This is clearly incompatible with (at least) models which claim that there is no feedback from the positional to the functional level of sentence planning (e.g., Bock & Levelt, 1994).

In general, the results of this and other experiments highlight the importance of cross-linguistic research for refining, and possibly revising, existing theories of human language processing, most of which were developed on the basis of English data. The Internet may provide the ideal medium for this kind of research.

**Note** The order of the authors is arbitrary. We wish to thank Ulf Reips, Bernad Batinic, Axel Theobald, and John Krantz for kindly providing links to our web experiment from their host pages. We are especially grateful to Frank Keller for his technical support.

## References

- Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355–387.
- Bock, J. K. (1989). Closed-class immanence in sentence production. *Cognition*, *31*, 163–186.
- Bock, J. K., & Levelt, W. J. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics*. San Diego, CA: Academic Press.
- Bock, J. K., & Loebell, H. (1990). Framing sentences. *Cognition*, *35*, 1–39.
- Branigan, H. P., Pickering, M. J., Liversedge, S. P., Stewart, A. P., & Urbach, T. P. (1995). Syntactic priming: Investigating the mental representation of language. *Journal of Psycholinguistic Research*, *24*, 489–506.
- Hartsuiker, R. J. (in preparation). Determination of word order in written and spoken sentence production.
- Hartsuiker, R. J., & Kolk, H. H. J. (1998a). Syntactic facilitation in agrammatic sentence production. *Brain and Language*, *62*, 221–254.
- Hartsuiker, R. J., & Kolk, H. H. J. (1998b). Syntactic persistence in Dutch. *Language and Speech*, *41*, 143–184.
- Howell, D. C. (1997). *Statistical methods for psychology* (4 ed.). Belmont, CA: Duxbury Press.
- Keller, F., Corley, M., Corley, S., Konieczny, L., & Todirascu, A. (1998). *WebExp: A Java toolbox for web-based psychological experiments* (Technical Report No. HCRC/TR-99). Human Communication Research Centre, University of Edinburgh.
- Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*, 633–651.
- Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42*, 107–142.
- Roelofs, A. (1993). Testing a non-decompositional theory of lemma retrieval in speaking: Retrieval of verbs. *Cognition*, *47*, 59–87.

# Hypertext Navigation and Conflicting Goal Intentions: Using Log Files to Study Distraction and Volitional Protection in Learning and Problem Solving

**Katharina Scheiter**

Graduate College for Cognitive Science  
University of Saarland  
D-66041 Saarbruecken/Germany  
katharis@cops.uni-sb.de

**Peter Gerjets**

Collaborative Research Center 378  
University of Saarland  
D-66041 Saarbruecken/Germany  
pgerjet@cops.uni-sb.de

**Elke Heise**

Department of Psychology  
Georg-August-University  
D-37073 Goettingen/Germany  
eheise@gwdg.de

## Abstract

This paper describes a theoretical analysis and experimental investigation of difficulty related distraction by conflicting goal intentions in learning and problem solving with hypertext. Log files are used to capture hypertext navigation in the face of opportunities to implement competing goal intentions. We study how differences in task difficulty influence the volitional protection of the current goal intention. First attempts to integrate volitional processes of action control into cognitive architectures are presented.

## Conflicting Goal Intentions in Hypertext Navigation

The investigation of learning and problem solving with hypertext gains increasing importance as the use of computer-based learning environments and information retrieval systems develops. The term "hypertext" refers to "computer-based texts that are read in a non-linear fashion and that are organized in multiple dimensions" (Landow, 1992, p. 166). A main feature of hypertext is that the user is not reacting to static texts, but is rather choosing according to his or her current intention when and in which order the information is to be presented (Barab, Bowdish, Young & Owen, 1996). Thus, the navigational path through a given hypertext environment depends mainly on the current intentions of the user. Accordingly, Barab et al. (1996) have shown that users' intentions in interacting with hypertext can be predicted from navigational paths captured in log files (computerized records of screens visited that are stamped with the amount of time spent on each screen). The opportunity of navigating through hypertext environments allows for great flexibility and adaptivity of learning and problem solving with hypertext, it is however also responsible for some difficulties. Users tend to be *structurally or conceptually disoriented* in complex hyperspaces and they seem to suffer from *cognitive overload*, if the navigational task consumes too much of their resources (Conklin, 1987).

In this paper we will focus on a further problem concerning navigating through hypertext environments, namely the problem of *being distracted by conflicting goal intentions*. We assume that learning and problem solving are to be analyzed as goal-directed behavior and furthermore that most learners possess numerous waiting goal intentions not related to the current problem. These waiting intentions can be activated by situational cues in the hy-

perertext environment and then compete with the current goal intention for execution. If the user is attracted by these cues, the current goal intention may be suspended in favor of activities related to the competing intention, or in favor of deliberating which of the two intentions should be further pursued. These interruptions and distractions due to conflicting goal intentions should lead to more or less severe efficiency impairments in learning and problem solving depending on the relative strength of the competing goal intentions. As a theoretical basis for analyzing these issues theories of action control are especially useful.

## Cognitive, Motivational, and Volitional Approaches to Action Control

If actions are considered as sequences of activities directed toward a common goal, the term "action control" can be used to describe automatic and controlled processes determining which activity is selected in the next step. Furthermore, action control includes processes that are predominantly *cognitive* (like the selection of a schema or production rule), predominantly *motivational* (like the deliberation of goal values in the course of intention formation), or predominantly *volitional* (like the maintenance of a goal intention in the face of competing intentions). Accordingly, theories from different fields of psychology are concerned with the analysis of action control.

**Purely cognitive approaches** Most of these approaches to action control postulate processing goals, but do not assume that differences in goal values are relevant for action control. Examples are theories of working memory and attention that postulate a supervisory attentional system responsible for intentional shifts of task sets and the control of working memory contents (Norman & Shallice, 1986). On a higher level of abstraction, theories of planning, strategy selection and metacognition are purely cognitive approaches. All of these approaches typically confine themselves to assuming mental representations and cognitive variables describing them, like activation, availability, or subjective probability. Furthermore, most cognitive approaches focus on single task situations and do not consider conflicting goal intentions.

**Approaches with motivational assumptions** Expanding on cognitive assumptions these approaches introduce variables that can be interpreted as goal values or as being dependent on goal values. Examples are theories of motiva-



tion and decision making that postulate expectancy-value-considerations as a basis for choosing between goals and action alternatives. Goal values and success probabilities are combined by calculating resulting motivational tendencies or subjective utilities that serve as a basis for decisions between goals or actions. An example from cognitive science is the cognitive architecture ACT-R (Anderson & Lebiere, 1998). The mechanisms of production rule selection in ACT-R depend on expected utilities of production rules calculated from goal values and success probabilities. A problem of many approaches to action control based on motivational assumptions is that they take differences in goal values for granted without giving further explanations for these values (Anderson & Lebiere, 1998, p. 63).

**Approaches with volitional assumptions** These theories describe control processes that help to initiate goal intentions, to maintain them in the face of difficulties and to protect them against distractions and competing goal intentions. Like motivational theories, volitional approaches are based on variables that depend on goal values (e.g. volitional strength of intentions) but, volitional approaches also describe how these variables change after a goal intention has been formed. They postulate automatic processes of goal protection like the adaptive increase of volitional strength in the face of increasing task difficulty (Gollwitzer, 1990; Heckhausen, 1991) as well as several kinds of volitional strategies to maintain goal intentions (Kuhl, 1987).

Because our paper is concerned with efficiency impairments caused by situational cues for competing goal intentions, theories of volitional action control seem to be best suited for a first analysis. As a framework for the description of volitional control processes we use a condensed and precise version of the rubicon theory of action phases (Gollwitzer, 1990; Heckhausen, 1991) called PART (Heise, Gerjets & Westermann, 1994), which comprises the **P**ivotal Assumptions of the **R**ubicon Theory.

PART allows for the derivation of specific predictions concerning efficiency impairments due to competing goal intentions under different conditions. We designed a hypertext learning environment in order to test these predictions within an experimental setting. PART can also serve as a basis for developing cognitive models of our experimental log file data, because the integration of motivational and volitional aspects of action control into cognitive architectures is easier, if a formalized model of volitional action control is available.

### PART: A Theory of Volitional Action Control

PART describes the entire course of actions from a time-sequential perspective. In addition to analyzing volitional processes, such as maintaining and protecting a goal intention, the theory also handles motivational processes, such as choosing a goal or assessing action outcomes. Within this framework, an action is typically composed of a sequence of four phases, beginning with the *predecisional* action phase and followed by the *preexecutive*, *executive*, and *postexecutive* phases (see Figure 1).

In the predecisional phase, one of several possible goals is chosen as the current goal intention to be pursued. This decision is based on the *motivational tendencies* associated

with the possible goals. In the preexecutive phase, which commences after commitment to a goal intention has been formed, intention-related activities are planned and the intention is maintained until a favorable opportunity for the initiation of these activities occurs.

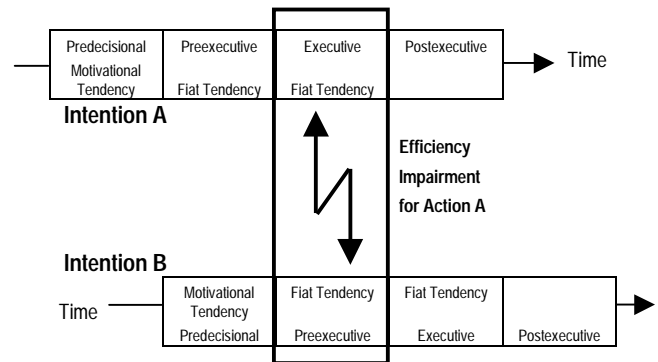


Figure 1: Action phases in PART

The initiation of intention-related activities is controlled by the so called *fiat tendency* and marks the transition to the executive phase. The fiat tendency of an intention depends partly on the suitability of a situation for the implementation of goal-directed behavior. During action execution, the fiat tendency is responsible for the maintenance of the goal intention. If difficulties occur, the fiat tendency of the goal intention increases. This variable is thus of central theoretical importance for explaining volitional action control. It can be interpreted as expressing how strongly an intention demands for implementation in a given situation. The executive phase is followed by the postexecutive phase in which the attained outcome is evaluated.

The theory of action phases is especially suitable for the analysis of action conflicts. In this paper we consider a specific type of conflict that can be described as suspended-intention conflict and is illustrated in Figure 1. It occurs when subjects are instructed to keep working on a task A for a longer period of time, while a competing intention B is waiting to be executed subsequently. In this case, one intention is supposed to remain in the executive phase, while the other is to remain in the preexecutive phase (*pre-executive-executive* conflict). Contrary to the superficially similar task-shift paradigm, where subjects are required to rapidly alternate between the execution of two intentions, no alternation is supposed to take place.

If the fiat tendency of the waiting intention B is strong, then the efficiency of action A should be impaired, because activities related to the competing intention occur or because a process of decision making is initiated in order to determine which intention should be pursued further. The distracting effect of the waiting intention B should be larger the stronger the fiat tendency of B is in comparison to the fiat tendency of the current intention A. The theory of action phases allows for the derivation of several empirically testable hypotheses, from which we chose two predictions that can be easily applied to hypertext navigation:

*Hypothesis of distraction by competing goal intentions:* The efficiency of a currently executed action A will be impaired if a favorable opportunity for the implementation of a competing intention B occurs. This prediction results

from the assumption that an opportunity to realize goal intention B leads to an increased fiat tendency of B.

*Hypothesis of difficulty-related volitional protection:* Efficiency impairments due to waiting intentions should be stronger for a low level of task difficulty than for a high level of task difficulty. This prediction results from the assumption that an increasing level of task difficulty for intention A results in an increased fiat tendency of A.

In several simple reaction time experiments using word classification tasks we were able to confirm both of these predictions (Heise, Gerjets & Westermann, 1997). In the domain of hypertext navigation we can test our predictions within a naturalistic setting, where problems of distraction due to competing intentions are of practical relevance. Furthermore, the use of log files to capture hypertext navigation enables us to investigate whether the distraction effects of a competing goal intention can be traced back to cognitive activities related to the implementation of this competing intention. Finally, we assume that research on hypertext navigation can benefit from insights in the way information processing strategies may change in the face of conflicting goal intentions.

## Experiment

### Method

**Subjects:** 134 students (84 female, 50 male) at the University of Goettingen, Germany participated in the experiment. The average age was 24,8 years.

**Procedure** The subjects' main task (the current goal intention) consisted of a hypertext-based learning and problem-solving task. Subjects had to solve three probability word problems. For each problem the correct solution principle and two correct variable values had to be marked in a multiple-choice form available in the hypertext environment. All three problems were presented at the beginning of the experiment. Subjects were instructed to solve the problems as fast and as correctly as possible using information provided in the hypertext environment. To acquire the relevant knowledge subjects could browse the hypertext environment freely. Six problem categories from the domain of probability theory were explained using worked out examples for illustration. All examples were embedded in interesting cover stories about attractiveness and mate choice (e.g., the probability of guessing the first three winners in a beauty competition between 10 people). The examples and the explanations of the problem categories were available during the whole experiment.

**Design** As independent variables two different levels of difficulty of the word problems to be solved (easy versus difficult problems) and two levels of distraction due to conflicting goal intentions (strong versus weak distraction) were introduced. Both variables of the resulting 2x2-design were varied interindividually. Two further levels of distraction were introduced as control conditions.

In accordance with preliminary studies we manipulated the level of difficulty by using smaller numbers in the easy problems and by stating them in a more familiar way than the difficult problems. The method used to increase difficulty was similar to the one used by Ross and Kilbane

(1997). The cover story and the underlying solution principle of a problem were not affected by this manipulation.

In the *condition with strong distraction* we introduced a competing intention and a favorable opportunity for its implementation. Subjects were informed that they would have to work on a second task within the same hypertext environment after having finished the problem-solving task. The second task consisted of answering three questions about attractiveness and mate choice that were presented briefly at the beginning of the experiment. Subjects were instructed to work on the problem-solving task first and to postpone thinking about the question-answering task until they finished the three word problems. They were assured to have enough time afterwards to browse the hypertext environment for information relevant to the second task. Subjects were told that all information available could be helpful in solving the word problems. As favorable opportunities to execute activities related to the waiting intention we included additional information about attractiveness and mate choice in the hypertext environment. This information was not helpful for solving the probability word problems, but it was related to the topic of the waiting intention. To make this information available during the first task, the examples used to explain the solution principles contained "hot words" linked to that information.

In the *condition with weak distraction* no competing intention was induced. Subjects were only required to solve the three word problems. In order to keep the number of hyperlinks in the learning environment constant, the same amount of irrelevant information was linked to the worked out examples as in the condition with strong distraction. In order to prevent subjects from forming an intrinsically motivated competitive intention to browse the irrelevant information, we replaced the interesting information about attractiveness and mate choice with rather uninteresting information concerning irrelevant terms in the cover story.

To control motivational effects of this replacement a *condition with intermediate distraction* was used. In this condition hyperlinks to irrelevant information about attractiveness and mate choice were inserted but no competing intention concerning that information was induced. If subjects form competing intentions based on personal interest, stronger effects of distraction than in the condition with uninteresting irrelevant information are to be expected.

A *baseline condition* with no hyperlinks to irrelevant information was implemented as a second control condition. This condition was used to estimate additional cognitive costs of navigating hypertext environments containing irrelevant information.

**Dependent variables** To test our hypotheses concerning efficiency impairments we obtained two different kinds of dependent variables. As an outcome measure the percentage of errors for the three word problems was registered. As process measures several time and frequency parameters were calculated from the log file data recorded during subjects' interaction with the hypertext system. Especially, the total amount of time spent on relevant information as well as time spent on irrelevant information was calculated. The latter measure was obtained to test whether efficiency impairments can be traced back to cognitive activities related to the competing intention.

## Results and Discussion

Comparing the conditions with strong and weak distraction yields a significant main effect of distraction on error rates (cf. Figure 2)<sup>1</sup>. In accordance with our distraction hypothesis, subjects with competing intentions and favorable opportunity to initiate corresponding activities show worse performance in the problem-solving. No differences between the condition with weak distraction and the two control conditions were found.

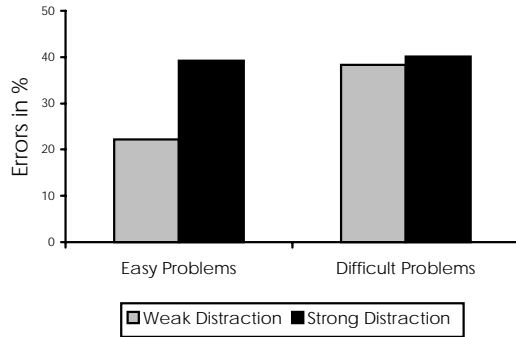


Figure 2: Error rates as a function of task difficulty and distraction (N = 68)

The manipulation of difficulty was successful. The respective main effect is significant. As predicted, the influence of the competing intention on performance depends on the difficulty of the problem-solving task. In the condition with low task difficulty, efficiency impairments due to competing intentions are larger than in the condition with high task difficulty. The respective interaction is significant.

To test whether these efficiency impairments can be traced back to cognitive activities related to the competing intention, we compared the four conditions regarding time spent on irrelevant information (see Figure 3). In the groups with competing intention and opportunity for implementation, the time spent on irrelevant information was significantly longer than in the groups with no competing intention. This was especially the case for the easy word problems. The respective interaction was marginally significant.

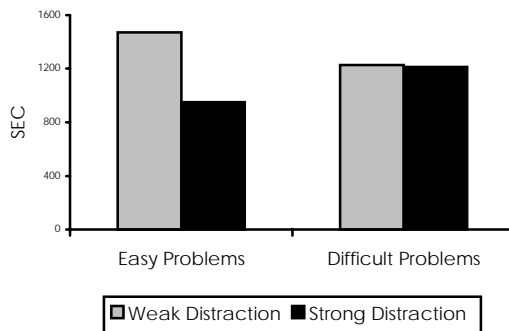


Figure 3: Time spent on irrelevant information as a function of task difficulty and distraction (N = 68)

<sup>1</sup> Our specific predictions have been tested using one-tailed *t*-tests. Concerning the general advantage of planned contrasts as opposed to unspecific ANOVA *F*-tests see Hays (1988) or Rosenthal and Rosnow (1988).

These data support the assumption that the observed efficiency impairments under the condition of low task difficulty result from cognitive activities which are relevant for the implementation of the competing intention.

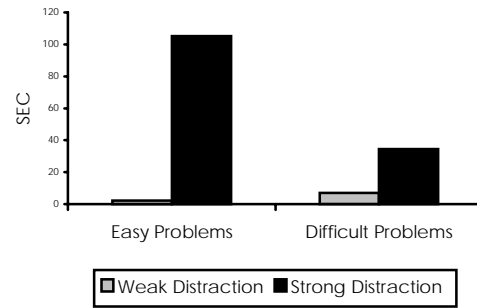


Figure 4: Time spent on relevant information as a function of task difficulty and distraction (N = 68)

In a second step we analyzed time spent on relevant information (Figure 4). As can be seen, higher difficulty of the word problems caused no increase in time on relevant information. Unexpectedly, the groups with a competing intention and opportunity for implementation spent significantly less time on relevant information than groups without such intention. This main effect is caused by differences in the low difficulty task condition. The respective interaction effect is significant.

Taken together, our data support the following conclusions: For high levels of task difficulty no distraction effects can be observed, whereas for low levels of task difficulty the presence of a competing intention leads to an increase in error rates and in time spent on irrelevant information as well as to a decrease in time spent on relevant information. This pattern of results can be interpreted as indicating strategy shifts if a strong competing intention with favorable opportunity is present (speed-accuracy trade-off).

Further analysis of the log file data yields several other strategy shifts under different levels of distraction and difficulty. For example, they concern the time spent on studying the solutions of worked out examples or the order in which the three word problems were solved.

## Summary

The aim of our study was to investigate how conflicting goal intentions can influence learning and problem solving in hypertext environments. As a theoretical background we used a theory of volitional action control (PART) that describes efficiency impairments caused by competing goal intentions under different levels of task difficulty. We used PART to derive two hypotheses about hypertext navigation in the face of conflicting goal intentions. These hypotheses could be confirmed in an experimental study. Furthermore, the experimental log file data show that there are numerous differences between the experimental conditions that cannot be completely explained by our volitional framework (e.g., different kinds of strategy shifts). To further analyze this data, it would be helpful to use a theoretical model that combines volitional assumptions about conflicting goal intentions and more detailed cognitive assumptions about

learning and problem solving behavior. We therefore began to integrate the volitional mechanisms of PART into the cognitive architecture ACT-R (Gerjets, 1997).

## Modeling Volitional Action Control

As illustrated in Figure 5, PART includes detailed assumptions about the interrelations of variables underlying volitional processes.

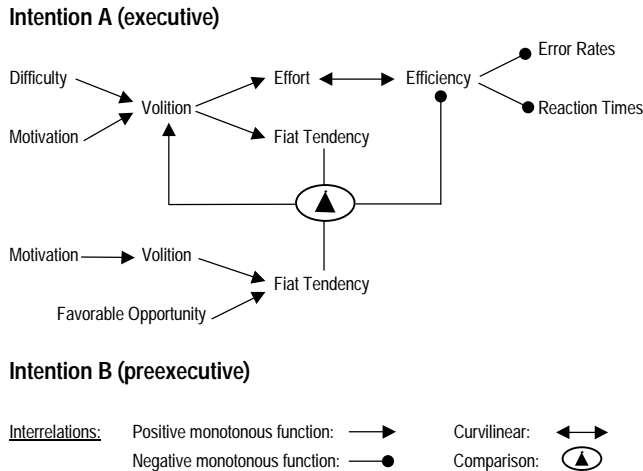


Figure 5: Interrelations between variables related to fiat tendencies in PART

Each possible goal is associated with a certain *motivational tendency* that determines which of the competing goals in the predecisional action phase will be pursued as goal intention. Each goal intention is assigned a specific degree of *volitional strength* that determines how much effort will be exerted for the implementation of that goal intention. Furthermore, each goal intention is assigned a *fiat tendency* that expresses its demand for implementation. In the pre-executive-executive conflict, the fiat tendencies of the competing goal intentions determine which of the two intentions becomes dominant. The fiat tendency of the current intention A depends essentially on the volitional strength of this intention. The volitional strength, however, not only determines the respective fiat tendency but also the level of effort and the efficiency of the implementing activities. The volitional strength of the current intention is affected by its motivational strength and the level of task difficulty. An increase in task difficulty results in an increased volitional strength of the current goal intention. This dynamic regulation of volition and effort in the face of increasing task difficulty is one of the main volitional mechanisms of the action phase theory and corresponds to the so-called *law of difficulty* (cf. Heckhausen, 1991).

The fiat tendency of the waiting intention B not only depends on its volitional strength but also on the perceived favorability of the opportunity to initiate corresponding activities. If the fiat tendency of the waiting intention becomes sufficiently large (relative to the fiat tendency of intention A), then the efficiency of the current action can be impaired as will be reflected in error rates or reaction times. It can also be assumed that the presence of a waiting intention with a strong fiat tendency will be perceived as an increased level of task difficulty that reactively results in an

increased level of volitional strength. This may at least partly compensate for efficiency impairments. Volitional action control is most adaptive when it results in a balance between shielding a current intention against competing intentions and flexibly responding to situational changes.

Based on our theoretical framework and our experimental findings, there are at least three requirements for the cognitive modeling of volitional action control. First, a cognitive model of our task will have to take into account that learners may simultaneously possess multiple conflicting goal intentions of differing strength. The model has to explain *efficiency impairments* caused by situational cues related to waiting intentions. Second, the model has to reflect the *law of difficulty*. Third, the model has to account for data indicating *strategy shifts* under different levels of distraction and difficulty that cannot be explained at PART's level of abstraction (time spent on different kinds of information, order of solving different problems and trade-offs between speed and accuracy).

As a theoretical basis for cognitive modeling we will refer to Anderson's ACT-R architecture (Anderson & Lebiere, 1998). ACT-R has been developed as a unified theory of cognition applying to domains as diverse as problem solving, learning, or memory. In ACT-R human actions are analyzed in terms of production rules and spreading activation in a network of declarative memory chunks. Production rules are matched to currently activated memory chunks and can be executed if their conditions are sufficiently satisfied. Actions are described as sequences of production rule firings. Action-guiding intentions can be represented by a specific type of declarative memory chunks (*goal chunks*). These chunks are organized by means of a last-in-first-out goal stack and act as temporary sources of activation that guide current information processing by spreading activation to other memory chunks and by thus constraining the set of executable production rules. Most productions are goal specific and can only be executed if the goal referred to in their conditions is the current goal on top of the goal stack.

A major drawback of the ACT-R architecture for our current purposes is that ACT-R is mainly designed as a single-task architecture for modeling tasks in isolation. Processing is completely controlled by the current goal on top of the goal stack. Production rules referring to other than the current goal cannot be selected for execution. Alternative cognitive architectures like EPIC (Meyer & Kieras, 1997) are explicitly designed for modeling dual task performance and multiple goal handling but are however restricted to very simple cognitive tasks lacking complex goal structures. Furthermore, they are not capable of integrating different cognitive components like learning, memory, and problem solving. For that reason, it seems easier to adapt ACT-R to handling multiple goals than to adapt architectures like EPIC for modeling complex cognition.

Our approach for modeling volitional action control in ACT-R comprises two main steps: In a first step we will try to map the concepts and assumptions of PART onto concepts and assumptions of ACT-R. These mappings can be evaluated theoretically (Gerjets, 1997) as well as empirically (Gerjets, Heise & Westermann, 1997). Because dynamic variables like motivational tendency, volitional strength and fiat tendency are of major importance in

PART, the modeling in ACT-R will focus on variables with analogous functional roles, e.g., goal values or source activation. If no satisfying mapping can be found for necessary assumptions of PART, we attempt to develop new concepts and mechanisms based on the ACT-R framework that are compatible with the main assumptions of the theory.

The aforementioned requirements for modeling volitional action control lead to three main subtasks in developing an ACT-R model for our domain.

**Efficiency impairments** Efficiency impairments due to competing intentions cannot be explained in ACT-R without additional theoretical assumptions, since the theory assumes that only the top goal on the goal stack controls performance. To model our data it will be necessary to introduce a new chunk type representing *preexecutive intentions* waiting for implementation. To allow the system to interrupt its performance for information processing related to a waiting intention, we will have to introduce goal unspecific production rules which, in the case of goal conflicts, initiate a decision about which task to pursue further. These *interrupt productions* should be executed whenever declarative memory chunks associated to waiting intentions become activated. Chunks representing waiting intentions can be equipped with additional functional characteristics like an increased base level activation to account for their superior availability in memory (Goschke & Kuhl, 1993).

**Difficulty-related effort allocation** Two interpretations of effort can be modeled in ACT-R. First, high effort can be interpreted as working more accurately. This can be modeled by high goal values, which lead to the selection of production rules with higher probability of success. These rules, however, also yield higher costs for execution. Second, high effort can be interpreted as working faster. This can be modeled by high source activation of goal chunks, which leads to a high amount of activation that spreads to associated chunks in declarative memory. Matching a production rule to chunks in declarative memory will be the faster the higher these chunks are activated. To model the law of difficulty, the concept of task difficulty has to be mapped onto analogous concepts in ACT-R. In our experimental context task difficulty can be best interpreted as missing declarative or procedural knowledge that results in problem-solving impasses. These impasses can be connected to goal values and source activation without violating the ACT-R theory.

**Strategy shifts** As our log file data indicate, level of distraction and difficulty influence the trade-off between speed and accuracy, the order in which test problems are solved, and the time spent on studying the solutions of worked out examples. As aforementioned, speed-accuracy trade-offs can be modeled by goal values and source activation. Time spent on studying worked out examples can be interpreted as reflecting a time demanding strategy with high success probability. Deviations from a given order of test problems can be explained by interrupt productions that react to activated chunks representing waiting intentions. To test whether these ideas suffice to model the influences of difficulty and distraction on strategy selection it will be necessary to develop a detailed ACT-R model of our experimental task.

## Acknowledgements

We would like to thank Peter Breuer for his support.

## References

- Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Erlbaum.
- Barab, S. A., Bowdish, B. E., Young, M. F. & Owen, S. V. (1996). Understanding kiosk navigation: Using log files to capture hypermedia searches. *Instructional Science*, 24, 377-395.
- Conklin, J. (1987). Hypertext: An Introduction and Survey. *Computer*, 20, 17-41.
- Gerjets, P. (1997). Volitionale Handlungssteuerung und kognitive Mechanismen: Reduktionsmöglichkeiten auf der Basis intertheoretischer Bänder. *Psychologische Beiträge*, 39, 441-470
- Gerjets, P., Heise, E. & Westermann, R. (1997). Motivational strength of goals and cognitive strength of goal representations. In M.G. Shafto & P. Langley (Eds.), Proceedings of the 19th Annual Conference of the Cognitive Science Society (S. 928). Mahwah, NJ: Erlbaum.
- Gollwitzer, P. M. (1990). Action phases and mind sets. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Vol II* (pp. 53-92). New York: Guilford Press.
- Goschke, T. & Kuhl, J. (1993). Representation of intentions: Persisting activation in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 1211-1226.
- Hays, W. L. (1988). *Statistics (4th ed.)*. Fort Worth, TX: Holt, Rinehart and Winston.
- Heckhausen, H. (1991). *Motivation and action*. New York: Springer.
- Heise, E., Gerjets, P. & Westermann, R. (1994). Idealized action phases: A concise Rubicon-theory. In M. Kuokkanen (Ed.), *Structuralism, idealization and approximation* (pp. 141-158). Amsterdam: Rodopi.
- Heise, E., Gerjets, P. & Westermann, R. (1997). The influence of competing intentions on action performance: Efficiency impairment and volitional protection in tasks of varying difficulty. *Acta Psychologica*, 97, 167-182.
- Kuhl, J. (1987). Action control: The maintenance of motivational states. In F. Halisch & J. Kuhl (Eds.), *Motivation, intention, and volition* (pp. 279-291). Berlin: Springer.
- Landow, G.P. (1992). *Hypertext. The convergence of contemporary literary theory and technology*. Baltimore: John Hopkins University Press.
- Meyer, D. E. & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review*, 104, 3-65.
- Norman, D. & Shallice, T. (1986). Attention to action. In R. J. Davidson (Ed.), *Consciousness and self-regulation* (pp. 1-18). New York: Plenum Press.
- Rosenthal, R. & Rosnow, R. L. (1988). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge: Cambridge University Press.
- Ross, B. H. & Kilbane, C. (1997). Effects of principle explanation and superficial similarity on analogical mapping in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 427-440.

# Clarifying Word Meanings in Computer-Administered Survey Interviews

**Michael F. Schober** ([schober@newschool.edu](mailto:schober@newschool.edu))

New School for Social Research; Department of Psychology, AL-340; 65 Fifth Avenue  
New York, NY 10003 USA

**Frederick G. Conrad** ([conrad\\_f@bls.gov](mailto:conrad_f@bls.gov))

Room 4915; Bureau of Labor Statistics; 2 Massachusetts Ave. NE  
Washington, DC 20212 USA

**Jonathan E. Bloom** ([jonathan\\_bloom@dragonsys.com](mailto:jonathan_bloom@dragonsys.com))

Dragon Systems, Inc.; 320 Nevada St.; Newton, MA 02160 USA

## Abstract

We investigated the extent to which a collaborative view of human conversation transfers to interaction with non-human agents. In two experiments we contrasted user-initiated and mixed-initiative clarification in computer-administered surveys. In the first study, users who could clarify the interpretations of questions by clicking on highlighted text comprehended questions more accurately (in ways that more closely fit the survey designers' intentions) than users who couldn't, and thus they provided more accurate responses. They were far more likely to obtain help when they had been instructed that clarification would be essential than when they were merely told it was available. In the second study, users interacting with a simulated speech interface responded more accurately, and asked more questions, when they received unsolicited clarification about question meaning from the system in response to their linguistic cues of uncertainty. The results suggest that clarification in collaborative systems will only be successful if users recognize that their own conceptions may differ from the system's, and if they are willing to take extra turns to improve their understanding.

## Introduction

Saying something doesn't guarantee it will be understood. People engage in dialog to make sure that what the speaker intended has been understood—to ground their understanding (e.g., Clark & Brennan, 1991; Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989). People ground their understanding to a criterion sufficient for their current purposes; in casual conversations (e.g., at a cocktail party), people may not need to understand precise details to satisfy their conversational goals, but in other settings (e.g., air traffic control tower conversations, calls to a technical help desk when your computer crashes, or conversations with your ex-spouse about child visitation) the stakes are higher.

This collaborative view of human conversation differs from traditional accounts of language use (what Akmajian et al., 1990 called the "message model" of communication), where listeners interpret utterances directly. The traditional view is that the meaning of an utterance is

contained within the words themselves, and that the process of comprehension involves looking up those meanings in the mental dictionary and combining them appropriately; a collaborative view argues that accurate comprehension also requires dialog so that people can clarify what is meant (see Clark, 1996).

In the studies reported here we investigate the extent to which this collaborative view of human conversation transfers to interaction with non-human agents, and we examine whether a collaborative view can improve user interface design. Examining collaboration in human-computer interaction forces us to specify details of the collaborative view that can test its limits and refine our theories of human communication.

We contrast two approaches to designing collaborative systems that support the clarification of word meanings. Under one approach, clarification is user-initiated—that is, if the user explicitly requests clarification, the system provides it. This requires users to recognize that they need clarification and to be willing to ask for it. Under the other approach, clarification is mixed-initiative—that is the system also provides (or offers to provide) clarification when it diagnoses misunderstanding, based on user behavior. For example, in a desktop or speech interface a system could provide clarification when the user takes too long to act; in a speech interface a system could provide clarification when the user's speech is hesitant or disfluent (containing *ums* and *uhs*, restarts, etc.).

We examine these issues in the context of survey interviewing systems, where systems present questions and users answer them. To our knowledge, current dialog systems for surveys (see Couper et al., 1998 on "computerized self-administered questionnaires") do not allow either user-initiated or mixed-initiative clarification of meaning. Rather, they embody strict principles of standardization developed for human-human interviews, where the interpretation of questions should be left entirely up to respondents (e.g., Fowler & Mangione, 1990). The argument for standardization is that if interviewers help respondents to interpret questions, they might influence responses, but if interviewers read scripted questions and provide only "neutral" feedback, responses are less

likely to be biased. We have demonstrated that in human-human interviews even supposedly nonbiasing feedback by interviewers can affect responses (Schober & Conrad, 1997, in press). More importantly, strict standardization can actually harm data quality because it prevents respondents from grounding their understanding of the questions. This is a problem because people's interpretations of seemingly straightforward questions like "How many bedrooms are there in your house?" can vary enormously; without grounding their understanding of questions, respondents may conceive of questions in unintended ways, and the resulting data may not fulfill the survey designers' purposes (Clark & Schober, 1991). We have shown that responses in strictly standardized interviews can be less accurate than responses in more interactive interviews where respondents can ground their understanding of questions with the interviewers (Conrad & Schober, 2000; Schober & Conrad, 1997).

The task of responding to a computerized survey differs from many human-computer interaction situations. First, in survey systems users provide information to the system rather than retrieving information from the system, as with a database query system or a web search interface. Second, survey system users' need for precise understanding may be lower than when they interact with other systems. Users may care less about precisely understanding the words in survey questions when providing opinions to market researchers (misunderstanding has few consequences for the user) than understanding the words in an on-line job application or an on-line health claims form (where misunderstandings can be costly).

## Experimental Methods

In our studies we assess whether systems that enable users to clarify the survey concepts do actually lead to improved comprehension of questions (and thus improved response accuracy), as a collaborative theory would predict. We examine the effects of clarification on task duration—clarification probably takes more time, and this may offset any benefits. We also examine the effects of clarification on user satisfaction; even if clarification improves comprehension, it could be annoying.

Our first study (Conrad & Schober, 1999) uses a desktop interface, in which the computer displays questions on a screen. The user enters responses and asks for clarification with the keyboard and mouse. Our second study (Bloom, 1999) uses an interface, in which questions are presented in a synthesized voice through a headset. The user answers questions and asks for clarification by speaking into the headset microphone.

In both studies, all users were asked the same survey questions, which had been used in earlier studies of human-human survey interviews (e.g. Schober & Conrad, 1997). We adapted 12 questions from three ongoing U.S. government surveys. Four questions were about employment, from the Current Population Survey (e.g., "Last week, did you do any work for pay?"); four questions

were about housing, from the Consumer Price Index Housing survey (e.g., "How many people live in this house?"); four questions were about purchases, from the Current Point of Purchase Survey (e.g., "During the past year, have you purchased or had expenses for household furniture?"). For each question, the survey designers had developed official definitions for the key concepts, which clarified whether, for example, a floor lamp should be considered a piece of household furniture, or whether a student away at college should be considered to be living at home.

Users answered these questions on the basis of fictional scenarios, so that we could measure response accuracy—that is, the fit between users' answers and the survey designers' official definitions. For each question there were two alternate scenarios, one typical and one atypical. With the typical scenario, the survey question was designed to be easy for users to interpret—to map onto the user's (fictional) circumstances in a straightforward way. For example, for the question "Has Kelley purchased or had expenses for household furniture?", the typical scenario was a receipt for an end table, which is clearly a piece of furniture. With the atypical scenario, it was less clear how the survey question should be answered. For example, for the household furniture question the atypical scenario was a receipt for a floor lamp, which is harder to classify without knowing the official definition of "household furniture."

For each user, half the scenarios described typical situations and half atypical situations.

## Study 1: Desktop interface

In this study, we varied the way the survey system provided clarification. When clarification was user-initiated, users could request the official definition for a survey concept by clicking the mouse on highlighted text in the question. When clarification was mixed-initiative, the system would also offer a definition when users were "slow" to respond. This was defined as taking longer than the median response time for atypical scenarios when no clarification was available. This offer was presented as a Windows dialog box; users could reject the offer by clicking "no" if they didn't think clarification was needed.

We also varied instructions to the users about how precisely they would need to understand the system's questions—that is, we varied the grounding criterion. Some users were told that clarification was essential; they were encouraged to obtain definitions from the computer because their everyday definitions might differ from the survey's. Other users were told merely that clarification was available, that definitions would be available if users wanted them. The five experimental conditions are displayed in Table 1.

54 users, recruited from an advertisement in the Washington Post, were paid to participate. Most (44) reported using a computer every day.

Table 1: Experimental conditions, Study 1.

Type of clarification	User instructed that...
1 no clarification	
2 at user's request	Clarification essential
3 at user's request	Clarification available
4 when user is slow or at user's request	Clarification essential
5 when user is slow or at user's request	Clarification available

## Results

Users' responses were almost perfectly accurate (their responses fit the official definitions) when they answered about typical scenarios. For atypical scenarios, users were more accurate when they could get clarification than when they couldn't (see Figure 1). Response accuracy mainly depended on the instructions to the user about the grounding criterion. When users had been told that definitions were merely available, their accuracy was as poor as when they couldn't get clarification. When they had been told that definitions were essential, response accuracy was much better, whether users had to request clarification,  $F(1,49) = 9.82, p < .01$ , or whether the system also offered it,  $F(1,49) = 14.38, p < .01$ .

As Figure 2 shows, response accuracy was strongly related to how often users received clarification. When users had been told that definitions were essential, they requested clarification most of the time; in fact, they frequently requested it for typical scenarios, when presumably it wasn't necessary. They also requested clarification quickly, which meant that when the system could also provide clarification (conditions 4 and 5) it rarely did. In contrast, users who had been told that clarification was merely available rarely asked for it, and they responded to the questions so quickly that system-initiated clarification was rarely triggered. Apparently, it didn't occur to these users that their interpretation of ordinary terms like "bedroom" and "job" might differ from the system's, and so they answered confidently, quickly, and inaccurately.

As Figure 3 shows, clarification took time. Response times were much longer in cases where users received clarification. As we anticipated, improved accuracy from clarification can be costly.

Users' ratings of their satisfaction with the system suggested two things. First, users who could not get clarification reported that they would have asked for clarification if they could. This suggests that interacting with dialog survey systems that don't allow clarification may be relatively unsatisfying. Second, users' grounding criteria affected their perceptions of the system. System-initiated clarification was rated on a 7 point scale as useful (6.0)

and not annoying (1.0) by "clarification essential" users, and less useful (3.9) and more annoying (4.25) by "clarification available" users. Presumably users who had been told that clarification was available found it jarring for the system to offer unsolicited help for seemingly straightforward questions.

Overall, these results suggest that the success of human-machine collaboration may depend both on users' grounding criteria—how important they believe it is to understand accurately—and also on whether users recognize that system concepts may differ from theirs.

## Study 2: Speech interface

This study used a Wizard-of-Oz technique to simulate a speech interface. Users believed they were interacting with a computer, when actually a hidden experimenter presented the questions and scripted clarification. To enhance believability, we used an artificial-sounding computer voice (Apple's "Agnes" voice).

This study used exactly the same questions and scenarios as Study 1. Users participated in one of four experimental conditions. In the first condition, the system never provided clarification. In the second condition, clarification was user-initiated—the system would provide clarification if users asked for it explicitly. In the third condition, the initiative was mixed—the system would "automatically" provide full definitions when users displayed specific uncertainty markers that had been shown to be more prevalent in atypical situations in human-human interviews collected with these materials (Bloom & Schorer, 1999). These included *ums*, *uhs*, pauses, repairs, and talk other than an answer. In the fourth condition, the system always provided clarification; no matter what the user did, the system would present the full official definition for every question.

40 users recruited from an advertisement in the *Village Voice* were paid to participate.

## Results

As in Study 1, users' responses were almost perfectly accurate when they answered about typical scenarios. For atypical scenarios, users were substantially more accurate when they were always given clarification (80%) than when they were never given clarification (33%),  $F(1,36) = 10, p < .005$ . When users initiated clarification, their response accuracy was no better (29%) than when they were never given clarification, because they almost never asked for it. As in Study 1, it seems likely that it didn't occur to users that clarification was necessary. Response accuracy was better when the initiative for clarification was mixed (59%),  $F(1,36) = 10.11, p < .005$ , although it was not as good as when clarification was given always.



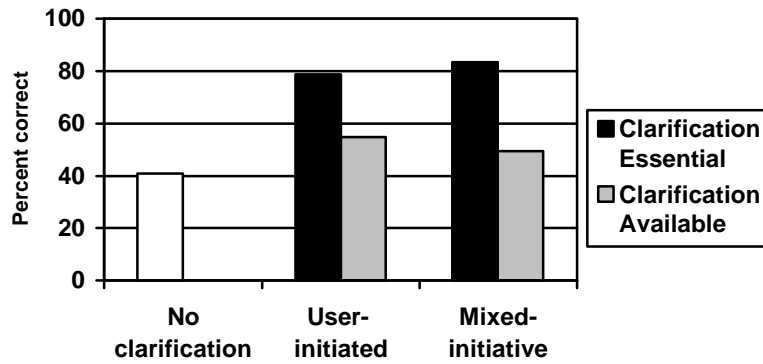


Figure 1: Response accuracy for atypical scenarios, Study 1

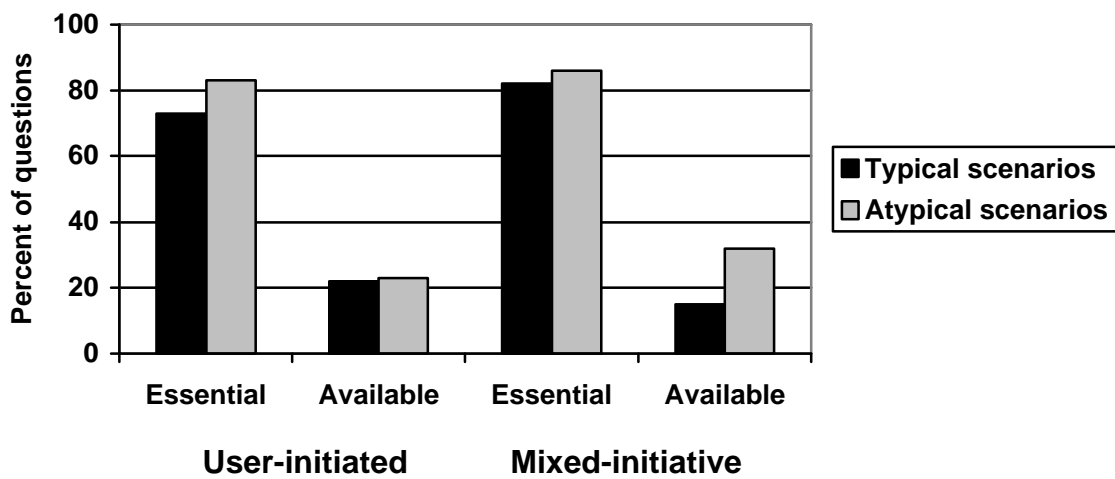


Figure 2: How often users received clarification, Study 1

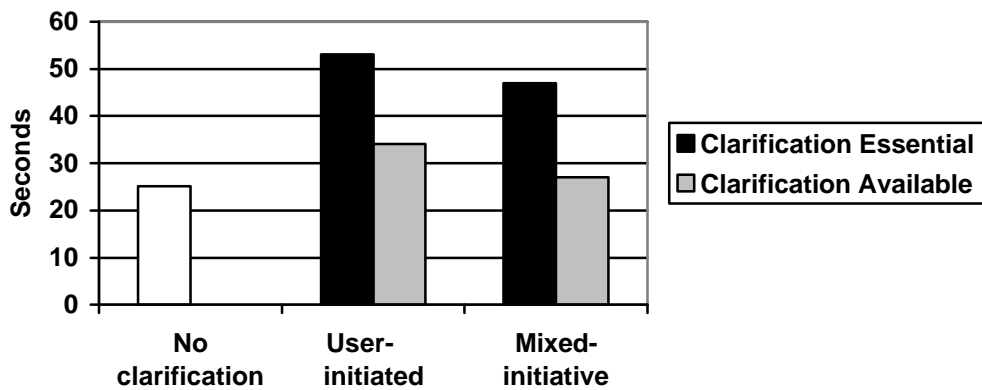


Figure 3: Response time per question, Study 1

System-initiated clarification increased the amount of user-initiated clarification: users were more likely to ask questions in the mixed-initiated condition, presumably because they were more likely to recognize that clarifica-

tion might be useful. These users also spoke less fluently, producing more *ums* and *uhs*. We speculate that this was because these users at some level recognized that the system was sensitive to their cues of uncertainty.

Overall, the users in this study requested clarification far less often than the users in Study 1. This might result from any or all of the differences between our desktop and speech interfaces. In the speech interface, clarification was harder to request; requests had to be formulated into explicit questions rather than being accomplished by simple mouse clicks. Also, in the speech interface the definition unfolded over time (sometimes a substantial amount of time, up to 108 seconds), rather than appearing all at once, and in our application it was impossible to shut off; in the desktop interface, the definition appeared all at once and could be closed with a simple mouse click. Also, unlike in the desktop study, users couldn't reject system-initiated offers of clarification; here the system immediately provided clarification when triggered, without giving the option of rejecting the help.

As in Study 1, clarification took time. The more clarification a user received, the more time the interviews took. Sessions where clarification was always provided took more than twice as long as sessions with no clarification or when it was (rarely) user-initiated (12.8 versus 5.2 and 4.9 seconds per question, respectively); mixed-initiative clarification took an intermediate amount of time (9.6 seconds per question).

Also as in Study 1, users rated the system more positively when it was responsive (user- or mixed-initiative conditions). When the system was not responsive (no clarification or clarification always), users wanted more control and felt that interacting with the system was unnatural. Users didn't report finding system-initiated clarification particularly more annoying than user-initiated clarification—which they almost never used.

Overall, these results suggest that enhancing the collaborative repertoire of a speech system can improve comprehension accuracy without harming user satisfaction, as long as the system provides help only when it is necessary. But these improvements come at the cost of increased task duration, which could make such systems impractical in real-world survey situations.

## Conclusions

Our findings demonstrate that a collaborative view can indeed transfer to interaction with non-human agents. Increased system clarification abilities can improve users' comprehension (and thus their response accuracy), while increasing (or not reducing) user satisfaction. But this comes at the cost of increased task duration, which could lower survey completion rates in the real world.

Our findings also demonstrate that extended clarification sequences are likely to be rare or unnecessary when users' conceptions are likely to be the same as the system's, as in our typical scenarios. The need for building survey systems with enhanced collaborative abilities may depend on the likelihood of potential misunderstandings; if this likelihood is high or unknown, enhanced collaborative abilities may be worth implementing.

The benefits we have shown for collaboratively enhanced survey systems come even with our rudimentary

implementations, which are based on the most generic of user models (see Kay, 1995). A stronger test of collaborative approaches requires more customized interfaces, in which, for example, the system would reason about which parts of definitions would be appropriate to present at any given moment, what particular users are likely to misunderstand, etc. (see Moore, 1995).

Our findings demonstrate that computer implementations of surveys seem to run into exactly the same problems as human-human survey and instructional situations, where people don't always recognize they need help or aren't willing or able to ask for help (e.g., Graesser & McMahan, 1993; Schober & Conrad, 1997).

But our findings also show that in some situations (our desktop interface, when users were told that clarification was essential), users are indeed willing to ask for clarification more often than they are with human interviewers (Schober & Conrad, 1997). This is consistent with findings in other domains that interaction with a computer can lead to better task outcomes than interaction with a person. For example, people may be more willing to accept correction from an intelligent computer tutor than from a human tutor (Schofield, 1995), and people are more willing to admit to undesirable behaviors when asked about them on self-administered computer surveys than in human-administered surveys (Tourangeau & Smith, 1996).

We propose that some of these improvements from interacting with computers don't arise simply from the fact that the computer isn't a person. They arise in part from the fact that the costs and constraints of grounding vary in different media, as Clark and Brennan (1991) argued. Most tutoring and survey systems to date have been direct manipulation or simple (textual) character entry systems like our desktop interface; in such interfaces the user's costs of requesting information from the system can be low. The human interactions to which such systems are often compared are speech interactions, where people have to formulate clarification requests explicitly and clarification takes significant amounts of time. Any differences in task performance may just as likely result from the differences between direct manipulation and speech as from the differences between computers and humans.

We believe our findings also require us to refine a theory of human-human collaboration by explicitly introducing the notion of initiative. Our findings that comprehension success can vary depending on whether the user or system takes the initiative should be extended to the human realm; a collaborative theory should include who takes the responsibility for clarifying meaning. In many cases speakers are responsible for what they mean, and listeners assume that what speakers say is readily interpretable to them in the current context (the "interpretability presumption," in Clark and Schober's [1991] terms). But in situations where the speaker is less competent than the addressee, the addressee may take responsibility for the meaning, and may initiate clarification (Schober, 1998). Who should be responsible under what circumstances, and what determines how speakers decide whose effort should be minimized, are important questions for a theory of collaboration.

Altogether, our results suggest that user-initiated clarification will work only if users recognize that clarification will help, recognize that the system's concepts may differ from theirs, are motivated to understand precisely, and are willing to take the extra turns to ground understanding. Explicit instructions to users can help make this happen—help set a high grounding criterion—but it's unclear whether such instruction is feasible in real-world situations. Our results suggest that system-initiated clarification will work only if users give reliable evidence of misunderstanding and if they are willing to accept offers of clarification. It won't work if users are confident in their misinterpretations.

In general, the opportunity for clarification dialog won't help if users don't recognize it's needed.

### Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No SBR9730140 and by the Bureau of Labor Statistics. The opinions expressed here are those of the authors and not of the Bureau of Labor Statistics. We thank Susan Brennan and Albert Corbett for help with an earlier version of this paper.

### References

- Akmajian, A., Demers, R.A., Farmer, A.K., & Harnish, R.M. (1990). *Linguistics: An introduction to language and communication, 3rd ed.* Cambridge, MA: MIT.
- Bloom, J.E. (1999). *Linguistic markers of respondent uncertainty during computer-administered survey interviews.* Doctoral dissertation, Department of Psychology, New School for Social Research, New York.
- Bloom, J.E., & Schober, M.F. (1999). Respondent cues that survey questions are in danger of being misunderstood. In *Proc. of the ASA, Section on Survey Research Methods* (pp. 992-997). Alexandria, VA: American Statistical Association.
- Clark, H.H. (1996). *Using language.* Cambridge, UK: Cambridge University Press.
- Clark, H.H., & Brennan, S.E. (1991). Grounding in communication. In L.B. Resnick, J.M. Levine, & S.D. Teasley, (Eds.), *Perspectives on socially shared cognition.* Washington, DC: APA.
- Clark, H.H., & Schaefer, E.F. (1989). Contributing to discourse. *Cognitive Science, 13*, 259-294.
- Clark, H.H., & Schober, M.F. (1991). Asking questions and influencing answers. In J.M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys.* New York: Russell Sage Foundation.
- Clark, H.H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-39.
- Conrad, F.G., & Schober, M.F. (1999). A conversational approach to text-based computer-administered questionnaires. *Proc. of the 3rd ASC Int'l Conference* (pp. 91-102). Chesham, UK: Association for Survey Computing.
- Conrad, F.G., & Schober, M.F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly, 64*, 1-28.
- Couper, M.P., et al. (Eds.) (1998). *Computer assisted survey information collection.* New York: Wiley.
- Fowler, F.J., & Mangione, T.W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error.* Newbury Park, CA: SAGE Publications, Inc.
- Graesser, A.C., & McMahan, C.L. (1993). Anomalous information triggers questions when adults solve quantitative problems and comprehend stories. *Journal of Educational Psychology, 85*, 136-151.
- Kay, J. (1995). Vive la difference! Individualized interaction with users. In C.S. Mellish, (Ed.), *Proc. of the 14th International Joint Conference on Artificial Intelligence*, 978-984. San Mateo, CA: Morgan Kaufmann.
- Moore, J.D. (1995). *Participating in explanatory dialogues: Interpreting and responding to questions in context.* Cambridge, MA: MIT.
- Schober, M.F. (1998). Different kinds of conversational perspective-taking. In S.R. Fussell & R.J. Kreuz (Eds.), *Social and cognitive psychological approaches to interpersonal communication.* Mahwah, NJ: Erlbaum.
- Schober, M.F., & Clark, H.H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*, 211-232.
- Schober, M.F., & Conrad, F.G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly, 61*, 576-602.
- Schober, M. F., and Conrad, F. G. (In press). A collaborative view of standardized survey interviews. In D. Maynard, H. Houtkoop, N. C. Schaeffer, & J van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview*, New York: Wiley.
- Schofield, J.W. (1995). *Computer and Classroom Culture.* Cambridge, UK: Cambridge University Press.
- Tourangeau, R., & Smith, T. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly, 60*, 275-304.

# Seeking coherent explanations — a fusion of structured connectionism, temporal synchrony, and evidential reasoning

Lokendra Shastri and Carter Wendelken  
International Computer Science Institute  
1947 Center Street, Suite 600  
Berkeley, CA 94704

## Abstract

A connectionist model capable of performing rapid inferences to establish explanatory and referential coherence is described. The model's ability to perform such inferences arises from (i) its structure, (ii) its use of mutual inhibition among "sibling" types, entities, and rules, (iii) the use of temporal synchrony for representing dynamic bindings, and (iv) its ability to rapidly modify weights in response to convergent activity.

## Introduction

Consider the following simple narrative: "John fell in the hallway. Tom had cleaned it. He got hurt." Upon hearing the above narrative most of us would infer that Tom had cleaned the hallway, John fell because he slipped on the wet hallway floor, and John got hurt because of the fall. These inferences allow us to establish causal and referential coherence among the events and entities involved in the narrative. They help us explain John's fall by making plausible inferences that the hallway floor was wet as a result of the cleaning and John fell because he slipped on the wet floor. They help us causally link John's hurt to his fall. They help us determine that "it" in the second sentence refers to the hallway, and "He" in the third sentence refers to John, and not to Tom. Empirical data strongly suggests that inferences required to establish referential and causal coherence occur automatically during language understanding (see e.g., Just & Carpenter 1977; Keenan, Baillet, and Brown 1984; Kintsch 1988; McKoon & Ratcliff 1980, 1992; Potts, Keenan, & Golding, 1988).

Any system that attempts to explain our ability to establish causal coherence during language understanding must possess a number of properties: First, such a system must be representationally adequate. It must be capable of encoding specific facts and events and expressing general regularities (aka rules) that capture the causal structure of the environment. In particular, the system should be capable of encoding context-dependent and evidential cause-effect relationships. Second, the system should be inferentially adequate, that is, it should be capable of drawing a range of explanatory inferences by combining evidence and arriving at *coherent* interpretations that are quasi-optimal with reference to a cost-function (Hobbs et. al, 1993). Third, the system should be capable of establishing referential coherence. In particular, it should be able to unify entities and events by recognizing that multiple designations might refer to the same entity or event. Fourth, the system should be capable of learning and fine-tuning its causal model based on experience, instruction, and exploration. Finally, the system should be scalable and computationally effective. The causal model underlying human language understanding would be extremely large. Yet we understand language at the rate of several hundred words per minute (Just & Carpenter 1977). Hence, a system for establishing causal coherence should also be capable of encoding

a large causal model and rapidly performing the requisite inferences within fractions of a second.

This paper describes several key extensions to the connectionist model SHRUTI that enable it to draw the sorts of inferences described above. SHRUTI is a neurally plausible system capable of expressing causal knowledge involving n-place relations, limited quantification, and type restrictions. It encodes specific events as well as context-sensitive priors over events. It expresses dynamic bindings via the synchronous firing of appropriate node clusters and performs inferences via the propagation of rhythmic activity over node clusters. This propagation amounts to a parallel breadth first activation of the underlying causal graph, and hence, the reasoning in SHRUTI is extremely fast. The use of weighted links and activation combination functions at nodes allow SHRUTI to encode soft rules and perform evidential inference. SHRUTI supports supervised learning which allows it to fine-tune its causal model in a data-driven manner (Shastri & Ajjanagadde, 1993; Shastri & Grannes, 1996; Shastri, 1999; Shastri & Wendelken, 1999; Wendelken & Shastri, 2000).

In order to carry out inferences for establishing referential and causal coherence, however, SHRUTI's core functionality had to be extended in a number of ways. These include the ability to (i) unify entities and relational instances (events) (ii) posit the existence of entities that are left implicit in the utterance, and (iii) favor interpretations that are more plausible and more likely over others that are less so. These functional extensions were realized in part by introducing mutual-exclusion clusters in the encoding of types and entities and by modifying the behavior of node-types. But more importantly, SHRUTI's inferential behavior was modified by (i) introducing inhibitory interactions among rules sharing a common consequent (effect) and (ii) modeling *short-term-potiation*, a biological phenomena whereby synaptic strengths (link weights) undergo rapid but short-lived changes in response to convergent activity. Both these changes play a critical role in favoring coherent and more-likely interpretations over less coherent and less likely ones.

The rest of the paper is organized as follows. The next section presents SHRUTI's basic representational machinery. This is followed by an elaboration of evidential reasoning in SHRUTI. Next we discuss mechanisms particularly aimed at the problem of establishing coherence and illustrate the functioning of the model with the help of an example.

## SHRUTI's representational machinery

Figure 1 illustrates the encoding of the following fragment of knowledge (expressed in SHRUTI's input syntax):

- (1)  $\forall x:\text{Agent}, y:\text{Location} [\text{slip}(x,y) \Rightarrow \text{fall}(x,y) (600,900)];$
- (2)  $\forall x:\text{Agent}, y:\text{Location} [\text{trip}(x,y) \Rightarrow \text{fall}(x,y) (800,900)];$
- (3) \*TF: trip(Person, Location) 100;
- (4) \*TF: slip(Person, Location) 50;

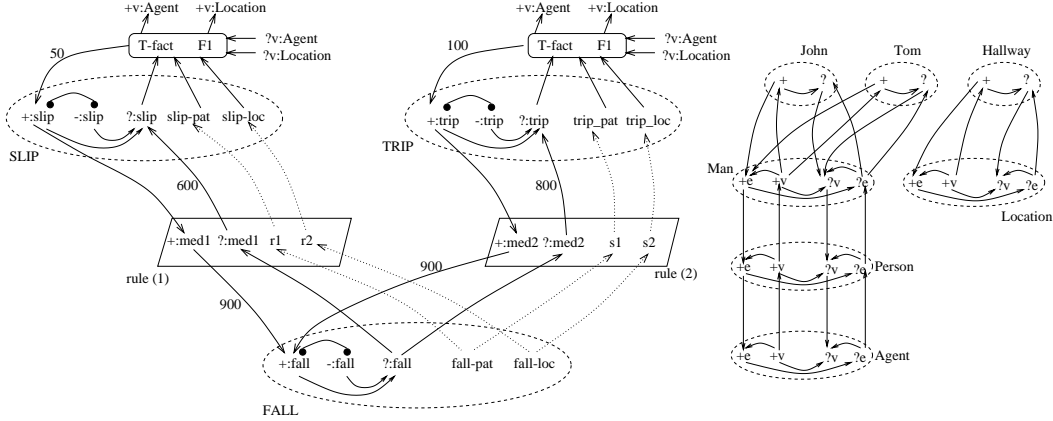


Figure 1: An example SHRUTI network encoding two rules (i)  $\forall x:\text{Agent}, y:\text{Location} [ \text{slips}(x,y) \Rightarrow \text{falls}(x,y) (600,900)]$ ; and (ii)  $\forall x:\text{Agent}, y:\text{Location} [ \text{trips}(x,y) \Rightarrow \text{falls}(x,y) (800,900)]$ ; two T-facts, F1 and F2; and a type hierarchy fragment. Links between mediator and type structures, and inhibitory links between sibling rules, entities, and types, have been omitted.

- (5) is-a( John, Man ); (6) is-a( Tom, Man );
- (7) is-a( Man, Person ); (8) is-a( Person, Agent );
- (2) is-a( Hallway, Location );

Items (1–2) are rules, items (3–4) are taxon-facts (T-facts), and item (5–9) are assertions about types. The first rule states that when an entity of type *Agent* slips at a location, then the latter may fall at that location. The weights (a,b) associated with a rule have an evidential interpretation and we discuss this in the section on evidential reasoning. The weight associated with a T-fact is indicative of the prior probability of the specified event type. All weights lie in the interval [0,1000].

### Encoding Relations, Entities, and Types

Each relation is represented by a focal cluster depicted by a dotted ellipse in Figure 1. Consider the focal cluster for *slip*. This cluster includes an enabler node labeled *?:slip*, two collector nodes labeled *+:slip* and  *-:slip*, and two role nodes labeled *slip-pat* and *slip-loc* for its two roles *patient* and *location*. In general, the cluster for an *n*-place relation contains *n* role nodes. The positive and negative collectors are mutually inhibitory (inhibitory links are depicted by filled circles).

Assume that the roles of *slip* have been dynamically bound to some fillers and thereby represent an active instance of *slip* (we will see how, shortly). The activation level of *?:slip* indicates the strength with which the system is seeking an explanation for the currently active instance of *slip*. The activation levels of *+:slip* and  *-:slip* encode graded beliefs about currently active instance of *slip* ranging continuously from *no* on the one extreme (only  *-:slip* is active), to *yes* on the other (only *+:slip* is active), and *don't know* in between (neither collector is very active). If both the collectors receive comparable and strong activation then both collectors can be active, despite mutual inhibition. This signals a contradiction.

The collector nodes of each relation are connected to the enabler node of the relation. For example, *+:fall* and  *-:fall* are connected to *?:fall*. These links cause *?:fall* to become active whenever *+:fall* or  *-:fall* become active. In effect, these links cause any active assertion about a relation to lead to a query about the assertion. Thus the system continually seeks an explanation for active assertions. The weight on the link from *+:fall* ( *-:fall*) to *?:fall* is inversely proportional to the probability of occurrence (non-occurrence) of an instance of *fall* — the less likely an event, the stronger the search for an

explanation of the event.

The encoding of types and instances is illustrated at the right of Figure 1. The focal cluster of each entity, *A* consists of a *?:A* and a *+:A* node. In contrast, the focal cluster of each type, *T* consists of a pair of *?* (*?:e:T* and *?:v:T*) and a pair of *+* nodes (*+:e:T* and *+:v:T*). While the nodes *+:v:T* and *?:v:T* participate in expression of knowledge (facts and attributes) involving the whole type *T*, the nodes *+:e:T* and *?:e:T* participate in the encoding of knowledge involving particular instances of type *T*. Thus the pair of *v* nodes and the pair of *e* nodes signify universal and existential quantification, respectively. The *levels* of activation of *?:A*, *?:v:T*, and *?:e:T* nodes signify the strength with which information about entity *A*, type *T*, and an instance of type *T*, respectively, is being sought. Similarly, the *levels* of activation of *+:A*, *+:v:T*, and *+:e:T* signify the degree of belief that the entity *A*, the type *T*, and an instance of type *T*, respectively, play appropriate roles in the current situation.

Nodes are computational abstractions and correspond to *small ensembles of cells*, and a connection between nodes corresponds to several connections from cells in one ensemble to cells in the other. Phasic nodes, of which role nodes are an example, produce output spikes in synchrony with their inputs. Temporal-and nodes, such as the enablers and collectors, integrate activity over a broader time window and produce wider output pulses (such a pulse may be identified with recurring high-frequency bursts of spikes).

### Dynamic bindings

The *dynamic* encoding of a relational instance corresponds to a *rhythmic* pattern of activity wherein bindings between roles and entities are represented by the *synchronous* firing of appropriate role and entity nodes (von der Malsburg 1981; Shastri & Ajjanagadde 1993; Hummel & Holyoak 1997). With reference to Figure 1, the dynamic representation of the relational instance (*fall*:  $\langle \text{fall-pat}=\text{John} \rangle, \langle \text{fall-loc}=\text{Hallway} \rangle$ ) (i.e., “John fell in the Hallway”) will involve the synchronous firing of *+:John* and *fall-pat*, and the synchronous firing of *+:Hallway* and *fall-loc*. The entities *+:John* and *+:Hallway* will fire in distinct phases.

### Encoding E-facts and T-facts

SHRUTI encodes two types of facts in its long-term memory: episodic facts (E-Facts) and taxon facts (T-facts). These facts

provide closure between the enabler node and the collector nodes. While an E-fact corresponds to a specific instance of a relation, a T-fact corresponds to a distillation or statistical summary of various instances of a relation and can be viewed as coding *prior probabilities*. T-facts are conditioned on the type of role-fillers. Typically, T-facts involving salient role-filler combinations such as  $[buy(a-Parent, a-Minivan) w1]$  (i.e., the prior probability that a parent buys a minivan is  $w1$ ) as well as more generic T-facts such as  $[buy(a-Person, a-Car) w2]$  would be learned. The priors for role-filler combinations not explicitly encoded would be inherited from generic T-facts.

### Encoding rules

A rule is encoded via a mediator focal cluster (shown as a parallelogram) that mediates the flow of activity between the antecedent and the consequent clusters.<sup>1</sup> The mediator consists of a collector and an enabler node and as many role-instantiation nodes as there are distinct variables in the rule. The enablers of the consequent relations are connected to the enablers of the antecedent relations via the enabler of the mediator. The (+/-) collectors of the antecedent relations are linked to the appropriate (+/-) collectors of the consequent relations via the collector of the mediator. Each of these enabler and collector links for a rule has a weight. The roles of the consequent relations are linked to the roles of the antecedent relations via the corresponding role-instantiation nodes in the mediator. This linking reflects the correspondence between antecedent and consequent roles specified by the rule.

If a role-instantiation node receives activation from the mediator enabler and a consequent role node, it simply propagates the activity onward to connected antecedent role nodes. If the role-instantiation node receives activity *only* from the mediator enabler it sends activity *only* to the node  $?e:T$ , where  $T$  is the type specified in the rule as the role type. This causes node  $?e:T$  to become active in an unoccupied phase. Node  $?e:T$  now conveys this activity to the role-instantiation node which in turn propagates this activity to connected antecedent role nodes. This interaction between the mediator and the type hierarchy, in effect, creates activity corresponding to “Does there exist some role filler of the specified type?” This is the mechanism by which new entities are posited and new phases emerge during the course of inference.

### Evidential Reasoning

The interpretation of link weights and activation values is intentionally underspecified in the core SHRUTI model. The goal has been to provide a flexible and expressive representational structure which can be fine-tuned according to specific modeling and task requirements. The following describes a specific interpretation of link weights in terms of probabilities that leads to satisfactory explanatory inferences.

#### A probabilistic interpretation of weights

Refer to the simplified SHRUTI network shown in Figure 2. The weight of the link from the enabler (?) of a relation to its collector (+) equals the (prior) probability of the occurrence of an instance of the relation. This weight corresponds to the weight of a T-fact associated with the relation. The weight of the link from the collector (+) of a relation to the enabler (?) of the relation is *inversely* proportional to the prior probability of the occurrence of an instance of the relation.

<sup>1</sup>The inclusion of a mediator was motivated, in part, by discussions the author had with Jerry Hobbs.

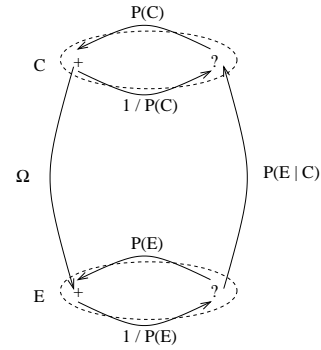


Figure 2: A simplified depiction of SHRUTI’s encoding of a rule and T-facts. The rule is  $C \rightarrow E$  and the T-facts are the prior probabilities of  $C$  and  $E$ . The negative collector and all roles nodes have been suppressed.

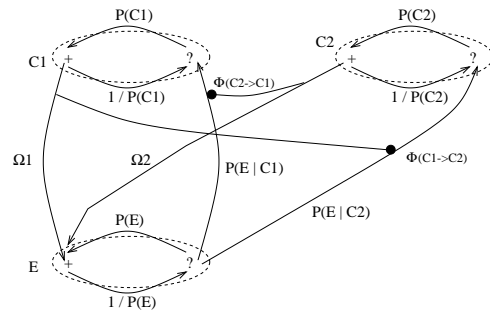


Figure 3: Inhibitory interaction between rules sharing a common consequent.

Now consider the encoding of the rule  $C \rightarrow E$ . The link weight from  $?E$  to  $?C$  equal  $P(E|C)$ , the probability of  $E$  given  $C$ . The weight,  $\Omega$ , of the link from  $+C$  to  $+E$  can be interpreted in several ways, as elaborated below. The simplest of these interpretations is  $P(E|only C)$ , the causal strength of  $C$  for  $E$  (this is essentially the independent component of a noisy-or). Another is  $P(C|E)$ .

When  $E$  is observed to be true, and hence,  $+E$ ’s activity level is clamped to 1.0, the activation of  $?E$  will equal  $1 * (1/P(E))$ , the activation of  $?C$  will equal  $(1/P(E)) * P(E|C)$ , and that of  $+C$  will equal  $(1/P(E)) * P(E|C) * P(C)$ . A direct application of Bayes Rule shows that the activation of  $+C$  reduces to  $P(C|E)$  — the desired degree of belief in  $C$  under a probabilistic interpretation. If there are multiple causes of  $E$ , say  $C1$  and  $C2$ , then subsequent to the clamping of  $+E$ ,  $C1$  and  $C2$  will become active at levels  $P(C1|E)$  and  $P(C2|E)$ , respectively, which is again as desired under a probabilistic interpretation (see Figure 3).

### Evidence combination

Where there are multiple sources of evidence for some predicate, then we must have some way to combine them. Since each source must communicate independently, along a single weighted link, the approach taken follows that of a belief-net noisy-or (Pearl 1988). However, to allow for more flexible evidence combination within this framework than what a single function can provide, a set of evidence combination functions was developed, based on notions of sufficiency or necessity of factors, and also on degrees of correlation.

Interestingly, these functions suggest several different interpretations of the link weights. At one end of this range is the familiar *noisy-or* function  $1 - \prod_i (1 - x_i * w_i)$ , where each weight  $w_i$  is essentially a measure of the sufficiency of each (independent) potential cause for bringing about the effect. At the other end of the spectrum, a sort of *noisy-and* function  $\prod_i (1 - (1 - x_i) * w_i)$  is used where the weight is interpreted as a degree of necessity, the probability that the consequent is false given that the particular antecedent is false (but all other necessary antecedents are true). In between these are *soft-or* (wherein positive correlation is allowed), a set of power averages  $((\sum_i X_i^k W_i) / (\sum_i W_i))^{1/k}$  ranging from *max* down to *min* depending on the parameter  $k$ , and a *soft-and* analogous to the *soft-or* (see Shastri & Wendelken, 1999).

## Mechanisms to support coherence

Several mechanisms have been developed which support the establishment of referential and causal coherence. These include inhibitory connections in the causal model, short-term potentiation, and the ability to create and collapse phases.

### Role of inhibitory connections

The encoding of a rule  $C \rightarrow E$  in SHRUTI involves inhibitory connections from  $+C$  to all the  $? \rightarrow ?$  links that originate from  $? \rightarrow E$  (see Figure 3) and reduce activity at their targets to a degree proportional to the activation of  $+C$ . These inhibitory links serve two purposes. First, they provide a mechanism for *contrast enhancement* since they allow stronger explanations to dominate over weaker explanations. Second, they serve the purpose of *explaining away*.<sup>2</sup> It is well known that combining explanatory and predictive inference can lead to problems in an inference system. For example, a system that can infer “John fell” from “John slipped”, and “John tripped” from “John fell” can also have the unfortunate tendency to infer “John tripped” based on “John slipped”. The inhibitory links prevent such unwarranted proliferation of evidence.<sup>3</sup> The precise impact of inhibition depends on the evidence combination function deployed at the site where the inhibitory links converge.

### Short-term Potentiation

If  $+fall$  receives activity from one of its T-facts it means that  $?fall$  is active, and hence, *fall* is being sought as a possible explanation of some event (say, hurt). If at the same time,  $+fall$  receives concurrent activity from  $+med1$  it means that *fall* is also being predicted as a possible consequence of a *slip* event. In these circumstances, it is *highly likely* that the fall event actually occurred and is both an effect of the slip event and an explanation of the fall event. SHRUTI expresses this increased likelihood via the biologically plausible mechanism of short-term potentiation (STP) (Bliss and Collingridge, 1993). Whenever a collector  $+P$  receives activity from one of its T- or E-fact and concurrent activity from a mediator collector node, then the weights of the links from the mediator collector to  $+P$  and from the active T-facts to  $+P$  increase for a short-duration. Analogous short-term weight changes occur due to convergence of top-down and bottom-up activity at links incident on  $-P$ : and at  $?P$ .

<sup>2</sup>This use of inhibitory connections is motivated in part by Ajanagadde (1991).

<sup>3</sup>The weights of these inhibitory links can be given a probabilistic interpretation. For example, the weight  $\phi(C2 \rightarrow C1)$  in Figure 3 can be viewed as  $[P(E) * P(E|C1, C2) * P(C1|C2)] / [P(E|C1) * P(E|C2) * P(C1)]$ .

With reference to Figure 3, consider a domain where  $A$  is a possible cause of  $C1$ , and hence we have the rule  $A \rightarrow C1$ . Now consider a situation where there is independent evidence for  $A$  and  $E$  and one is interested in determining the probability of  $C1$ ,  $P(C1|A, E)$ . This probability cannot be exactly computed using only information available locally at node  $C1$ . Simply combining the evidence arriving from  $E$  (i.e.,  $P(C|E)$ ) and  $A$  (i.e.,  $P(C1|onlyA)$ ) using an evidence combination function such as *noisy-or* would typically lead to an underestimation of the correct value. However, the short-term potentiation (STP) of links allows SHRUTI to partially offset this underestimation of the probability of an intermediate relation when both the cause and the effect of a relation are observed. At the same time, the unpotentiated weights continue to propagate the correct probability values when only the cause or only the effect is observed.

At a more global level, STP also has the effect of *priming* the whole subnetwork of nodes and links that constitute a coherent interpretation and creating a strong feedback loop of reverberant activity in a subnetwork of causal knowledge corresponding to a coherent interpretation.

Taken together, the short-term associative increase in weights and the inhibitory interactions leading to the explaining away phenomena, provide a powerful and neurally plausible mechanism that enable SHRUTI to prefer coherent explanations over non-coherent ones.

### Mutual exclusion and collapsing of phases

Entities in the type hierarchy can be part of a *phase-level* mutual exclusion cluster ( $\rho$ -mex cluster). Consequently, only the most active entity within a  $\rho$ -mex cluster can remain active in any given phase. A similar  $\rho$ -mex cluster can be formed by mutually exclusive types. Mutual exclusion also occurs in the type hierarchy as a result of inhibitory connections from the  $+$  nodes of a type (or an entity) to the  $?$  nodes of all its siblings. This inhibition leads to another sort of “explaining away” phenomenon. If for example, the type query “Is it a Person?” (i.e., activation of  $?e:Person$ ) leads to the queries “Is it a Man?” and “Is it a Woman?”, then strong support received by  $+e:Woman$  reduces the strength of the query  $?e:Man$ . In essence, the query “Is it a Man?” is no longer considered important by the system since it was seeking a person and it has already found a woman.

SHRUTI allows separate phases to coalesce into a single phase, or new phases to emerge, as a result of inference. The latter is realized by the allocation of new phases resulting from the interaction between role-instantiation nodes in mediators and the type hierarchy, as described above. The unification of phases is realized in the current implementation by the collapsing of phases based on activity within an entity cluster or within a focal cluster. In the first case, phase collapsing occurs whenever a single entity dominates multiple phases (for example if the same entity comes to be the answer to multiple queries). In the second case, phase collapse occurs if two unifiable instantiations of a relation arise within a focal cluster. For example, the active assertion  $+fall(John, Hallway)$  alongside the query  $\exists x:Man ?fall(x, Hallway)$  (Did a man fall in the Hallway?) will result in the merging of the two phases for “a man” and “John” via the inferred assertion  $\exists x:Man +fall(x, Hallway)$ . The same assertion alongside the query  $\exists x:Woman ?fall(x, Hallway)$  would not lead to a similar phase merge because the types Man and Woman are mutually exclusive, and hence, would mutually inhibit one another.

SHRUTI’s ability to readily and flexibly instantiate entities and collapse them into a single entity during inference is due to its use of temporal synchrony to represent dynamic bindings.

## Simulation Result

The activation trace resulting from the processing of the "John fell" story by a SHRUTI network encoding the rules, T-facts, and type hierarchy described in Section is shown in Figures 4 and 5. Figure 4 shows the actual activation levels of the  $+:slip$  and  $+:trip$  nodes as the story is processed by SHRUTI. Figure 5 depicts the activation trace of a larger *subset* of nodes. The depiction in this figure, however, has been simplified to highlight key aspects of the network behavior. In particular, several nodes have been omitted, some intermediate cycles have been omitted and the activation levels of collector and enabler nodes have been discretized to four levels. Please note that due to simplifications made to Figure 5, the time scales along the x-axis in Figures 4 and 5 are not the same. To minimize confusion, we will refer to the times in Figure 4 as cycles and in Figure 5 as steps. The reader may also wish to refer to Figure 1 to ground some of the following description.

Each sentence in the narrative is conveyed to SHRUTI by activating the  $+$  node of the appropriate relation and establishing role-entity bindings by the synchronous activation of the appropriate role and entity nodes. The sentences are presented in sequence and after each sentence presentation, the network is allowed to propagate activity for a fixed number of cycles. For example, the first sentence (S1) is communicated to SHRUTI in step 1 (cycle 0) by activating the node  $+:fall$ , the nodes  $fall-pat$  and  $+:John$  in synchrony, and the nodes  $fall-loc$  and  $+:Hallway$  in synchrony. The firing of nodes  $John$  and  $+:Hallway$  occupy distinct phases —  $\rho_1$  and  $\rho_2$ , respectively.

Activation from the focal cluster for  $fall$  reaches the mediator structure of rules (1) and (2). Consequently, nodes  $r1$  and  $r2$  in the mediator for rule (1) become active in phases  $\rho_1$  and  $\rho_2$ , respectively. Similarly, nodes  $s1$  and  $s2$  in the mediator of rule (2) become active in phases  $\rho_1$  and  $\rho_2$ , respectively. At the same time, the activation from  $+:fall$  activates  $?:fall$  which in turn activates the enablers  $?:med1$  and  $?:med2$  (the activity of mediator nodes, and role nodes of  $slip$  and  $trip$  is not depicted in Figure 5). The activation from nodes  $r1$  and  $r2$  reaches the roles  $slip-pat$  and  $slip-loc$  in the  $slip$  focal cluster, respectively. Similarly, activation from nodes  $s1$  and  $s2$  reach the roles  $trip-pat$  and  $trip-loc$  in the  $trip$  focal cluster, respectively. In essence, the system has created new bindings for the  $slip$  and  $trip$  relations. These bindings together with the activation of the nodes  $?:slip$  and  $?:trip$  encode two queries: "Did John slip in the hallway?", and "Did John trip in the hallway?". At the same time, activation travels in the type hierarchy and activates the nodes  $?v:Man$ , then  $?v:Person$ , and then  $?v:Agent$  in phase  $\rho_1$ , and the  $?v:Location$  node in phase  $\rho_2$ . The coincident activity of  $slip-pat$  and  $?v:person$  node, and the coincident activity of the  $slip-loc$  and  $?v:Location$  nodes leads to the firing of the T-fact F1 associated with slip. The activation of F1 causes activation from  $?:slip$  to flow to  $+:slip$ . The T-fact F2 associated with trip also becomes active in an analogous manner and conveys activation from  $?:trip$  to  $+:trip$ . The level of these activations is a measure of the probability that a person may slip and fall, respectively. At this time, "John tripped" is believed to be a more likely explanation of "John fell" than "John slipped."

While the activation spreads "backwards" from the  $fall$  focal cluster in the manner described above, activation also travels "forwards" to the  $hurt$  focal-cluster (not shown in Figure 1) as a result of the encoding of rule (iii) (also not shown) and leads to the weak prediction that John got hurt.

The introduction of sentence S2 in step 6 (Figure 5) (cycle 40 Figure 4) results in the instantiation of  $clean$  with the bindings ( $\{clean-agt=+:Tom\}$ , and  $\{clean-loc=+:e:Location\}$ ). As a result, Tom gets active in phase  $\rho_3$  and  $+:e:Location$  in phase

$\rho_4$ . Note that now we have two instantiations of a location. The second instantiation gets merged with the first (Hallway) as a result of phase merging. This happens in step 8 (see activity of  $+:e:location$  in Figure 5). The pressure for this merging comes from the strong compatibility, and hence, the strong coherence between the activity of hallway and the new location. Note that in the ongoing activity, hallway and the new location (say, Loc1) are active in parallel assertions such as: "John fell on the hallway floor", "The hallway floor might have been wet", "The hallway floor might have been cleaned" and "The Loc1 floor was cleaned" "The Loc1 floor might be wet", "John might have fallen in the hallway floor." At this time,  $+:wetFloor$  also becomes active as a result of activity arriving from  $+:clean$  via the mediator of rule (4) (cleaning leads to a wet floor).

By step 10 (Figure 5)  $+:slip$  becomes more active as a result of the high activation of  $+:wetFloor$ . The effect of "explaining away" kicks in and causes the activation of  $+:trip$  to go down by step 12. The strength of  $+:slip$  increases even further due to (i) the potentiation of links from the mediator for rule (4) (walking on a wet floor may cause slipping), (ii) the potentiation of the link from  $?:med1$  to  $?:slip$ , and (iii) the effect of explaining away. The effect of these changes on the activation levels of  $+:slip$  and  $+:trip$  may be seen more vividly in the detailed trace shown in Figure 4.<sup>4</sup>

S3 is introduced in step 14 (cycle 80) with the binding ( $\{hurt-pat=+:e:Man\}$ ). This leads to  $+:e:Man$  becoming active in phase  $\rho_4$  and a second dynamic instantiation of  $hurt$  (in addition to the earlier instantiation resulting from the inference  $hurt(John)$ ). These two instantiations get merged immediately, and phase  $\rho_4$  gets merged with  $\rho_1$  (John), in step 15 as a result of the phase merging described in the previous Section.

## Conclusion

SHRUTI shows how explanatory and referential coherence can arise within a neurally plausible system as a result of spontaneous activity in a network. The network's structure reflects the causal model of the environment and when the nodes in the network are activated to reflect a given state of affairs, the network spontaneously seeks coherent explanations. The time taken to perform an inference is simply proportional to the depth of the causal derivation and is otherwise independent of the size of the causal model. The state of coherence is reflected as reverberatory activity around *closed loops*. The system also makes predictive (forward) inferences, but only those predictions that become part of a coherent explanation gain strength and persist. Coherence arises in SHRUTI as a result of (i) inhibitory interactions among sibling entities, types and rules, (ii) short-term increase in link weights resulting from short-term potentiation, and (iii) the dynamic merging and instantiation of entities.

## Acknowledgments

This work was partially funded by NSF grants No. 9720398 and N0. 9970890 and subcontracts from Cognitive Technologies Inc. related to ARI contract DASW01-97-C-0038. Thanks to Jerry Feldman, Jerry Hobbs, Marvin Cohen and Bryan Thompson.

<sup>4</sup>If sentence S2 were delayed, the activity in  $slip$  would lead to the instantiation of an instance of  $clean$  with an entity of type  $agent$  being instantiated as a potential filler of the role  $clean-agt$ . This entity, however, gets unified with  $Tom$  upon the introduction of S2.



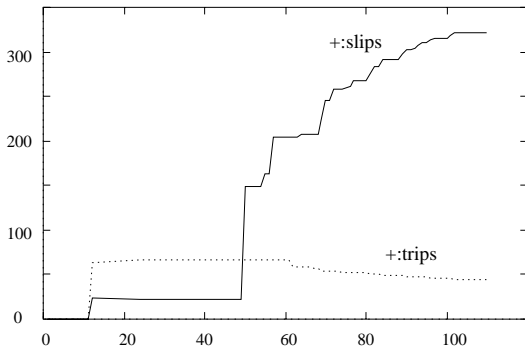


Figure 4: The activation trace of collector nodes  $+:slip$  and  $+:trip$  during the processing of the “John fell” story. X-axis is time. The activity of these collectors around cycle 12 is due to associated T-facts. Since tripping is more likely than slipping (100 versus 50),  $+:trip$  has a higher activation. Activity from the *clean* predicate arrives (via *wetFloor*) at the *slip* collector at cycle 50 due to the introduction of S2 at cycle 40, giving  $+:slip$  a significant boost. >From here onwards the associative weight changes along highly active pathways into  $+:slip$  result in a large increase in values at around cycle 55. The potentiation of the path from  $?:fall$  to  $?:slip$  also contributes to this increase. At the same time, the “explaining away” phenomena leads to the decrease in the activation of  $+:trip$ . The activity stabilizes around cycle 100. Note that each cycle in SHRUTI roughly corresponds to twice the period of  $\gamma$  band activity, i.e., about 40-50 msec. (see Shastri & Ajjanagadde 1993).

## References

- Ajjanagadde, V. (1991) Abductive reasoning in connectionist networks. TR WSI 91-6, Wilhelm-Schickard Institute, University of Tübingen, Tübingen, Germany.
- Bliss, T.V.P. and Collingridge, G.L. (1993) A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* 361, 31–39.
- Just, M.A. & Carpenter, P.A. Eds. (1977) *Cognitive processes in comprehension*. Erlbaum.
- Hobbs, J., Stickel, M., Appelt, D., & Martin, P. (1993) Interpretation as abduction. *Artificial Intelligence*, 63, 69–142.
- Hummel, J. E., & Holyoak, K.J. (1997) Distributed representations of structure: a theory of analogical access and mapping. *Psychological Review*, 104, 427-466.
- Keenan, J. M., Baillet, S. D., & Brown, P. (1984) The Effects of Causal Cohesion on Comprehension and Memory. *Journal of Verbal Learning and Verbal Behavior*, 23, 115-126.
- Kintsch, W. (1988) The Role of Knowledge Discourse Comprehension: A Construction-Integration Model. *Psychological Review*, Vol. 95, 163-182.
- McKoon, G., & Ratcliff, R. (1980) The Comprehension Processes and Memory Structures Involved in Anaphoric Reference. *Jrnl. of Verbal Learning and Verbal Behavior*, 19, 668-682.
- McKoon, G., & Ratcliff, R. (1992) Inference During Reading. *Psychological Review*, 99, 440-466.
- Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.
- Potts, G. R., Keenan, J. M., & Golding, J. M. (1988) Assessing the Occurrence of Elaborative Inferences: Lexical Decision versus Naming. *Journal of Memory and Language*, 27, 399-415.

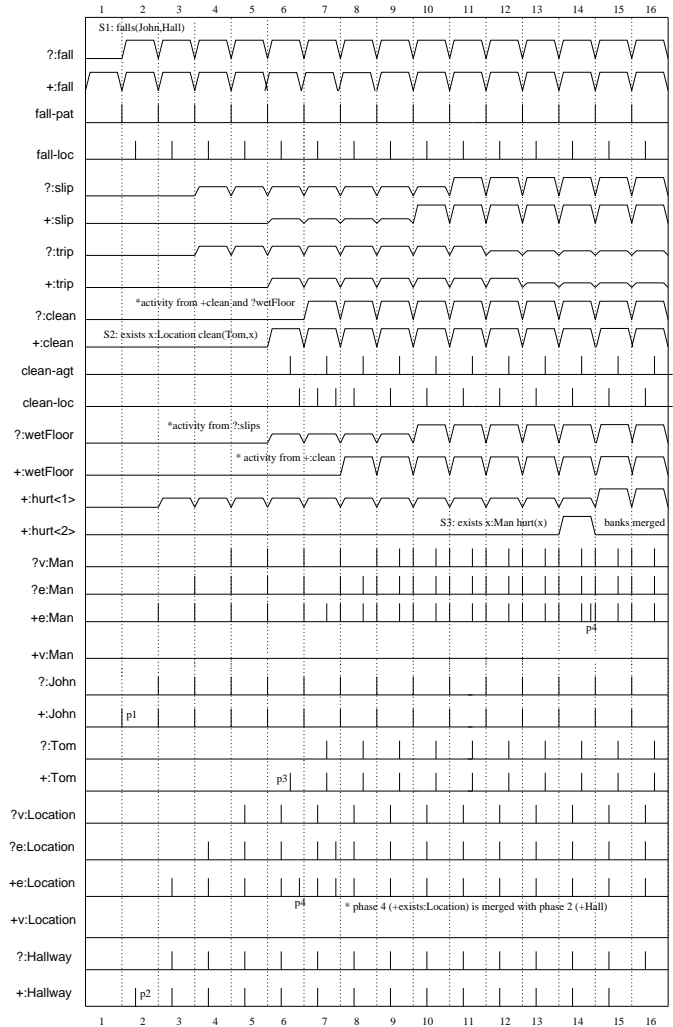


Figure 5: A schematized activation trace of selected nodes.

- Shastri, L. (1999) Advances in SHRUTI — A neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony, *Applied Intelligence*, 11, 79–108.
- Shastri, L. & Ajjanagadde V. (1993) From simple associations to systematic reasoning. *Behavioral and Brain Sciences*, 16:3 p. 417-494.
- Shastri, L. & Grannes, D. (1996) A connectionist treatment of negation and inconsistency, *Proc. Eighteenth Conference of the Cognitive Science Society*, San Diego, CA. 1996.
- Shastri, L. & Wendelken, C. (1999) Soft Computing in SHRUTI. In *Proc. the Third International Symposium on Soft Computing*, Genova, Italy. June, 1999, pp. 741–747.
- von der Malsburg, C. (1981) The correlation theory of brain function. Internal Report 81-2. Department of Neurobiology, Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany.
- Wendelken, C & Shastri, L. (2000) Probabilistic Inference and Learning in a Connectionist Causal Network. In *Proc. Neural Computation 2000*, Berlin 2000. To appear.

# Infant Familiarization to Artificial Sentences: Rule-like Behavior Without Explicit Rules and Variables

**Thomas R. Shultz** (shultz@psych.mcgill.ca)  
Department of Psychology; McGill University  
Montreal, QC H3A 1B1 Canada

**Alan C. Bale** (alan\_bale@sympatico.ca)  
Department of Linguistics; McGill University  
Montreal, QC H3A 1G5 Canada

## Abstract

A recent study of infant familiarization to artificial sentences claimed to produce data that could only be explained by symbolic rule learning and not by unstructured neural networks. Here we present successful unstructured neural network simulations showing that these data do not uniquely support a rule-based account. In contrast to other neural network simulations, our simulations cover more aspects of the data with fewer assumptions using a more realistic coding scheme based on sonority of phonemes. Our networks show exponential decreases in attention to a repeated sentence pattern, more recovery to novel inconsistent sentences than to novel consistent sentences, some preference reversals, and extrapolation.

One of the most simulated phenomena in developmental psychology is a data set that was claimed to be immune from simulation by unstructured neural networks (Marcus, Vijayan, Bandi Rao, & Vishton, 1999). Although the authors maintained that their results could only be explained by explicit rules and variables, there are now at least eight connectionist simulations of the data, most of which do not use explicit variable binding and none of which use explicit rules. Here we present additional neural simulations of these data, arguing that our model may provide the currently most satisfying account. The paper reviews the relevant infant data, presents various interpretations and models, and then focuses on our current model.

## The Infant Data

The relevant experiments familiarized 7-month-old infants to three-word artificial sentences and then tested them on novel sentences that were either consistent or inconsistent with the familiar pattern. The design of these experiments is shown in Table 1. In Experiment 1, infants were familiarized to sentences with either an ABA pattern (e.g., *ni la ni*) or an ABB pattern (e.g., *ta gi gi*). There were 16 of these sentences, constructed by combining four A-category words (*ga*, *li*, *ni*, and *ta*) with four B-category words (*ti*, *na*, *gi*, and *la*). After infants became familiar with a sentence pattern, they were tested with two sentences having novel words

that were either consistent or inconsistent with the familiar pattern.

Table 1: Marcus et al. (1999) experiments.

Pattern	Experiments 1 & 2		Experiment 3	
	Cond. 1	Cond. 2	Cond. 1	Cond. 2
Familiarize	ABA	ABB	ABB	AAB
Consistent	ABA	ABB	ABB	AAB
Inconsistent	ABB	ABA	AAB	ABB

When an infant looked at a flashing light to the left or right, a test sentence was played from a speaker situated next to the light. Each test sentence was played until the infant either looked away or 15 s elapsed. Infants attended more to inconsistent novel sentences than to consistent novel sentences, showing that they distinguished the two sentence types.

Experiment 2 was the same except that the words were chosen more carefully so that phoneme sequences were different in the familiarization and test patterns. Experiment 3 used the same words as Experiment 2, but in contrastive syntactic patterns that each duplicated a consecutive word: AAB vs. ABB. The idea was to rule out the possibility that infants might have used the presence or absence of consecutively duplicated words to distinguish sentence types.

In all three experiments, infants attended more to inconsistent than to consistent novel sentences. Our concern is with the best theoretical account of these data. Is the infant cognition based on rules and variables or on connections?

## A Rule and Variable Interpretation

Marcus et al. (1999) argued that these grammars could not be learned by the statistical methods common to standard neural networks. They also tried some unsuccessful neural network simulations using Simple Recurrent Networks (SRN). The authors proposed that only a rule-based model could cover their data. "We propose that a system that could account for our results is one in which infants extract algebra-like rules that represent relationships between placeholders (variables) such as 'the first item X is the same as the third

item Y' (p. 79)." They allowed that their data might also be accounted for by structured neural networks that implement explicit rules and variables in a neural style: "The problem is not with neural networks per se but with the kinds of neural networks that are currently popular. These networks eschew explicit representations of variables and relations between variables; in contrast, some less widely discussed neural networks with a very different architecture do incorporate such machinery and thus might form the basis for learning mechanisms that could account for our data (pp. 79-80)."

### **Psychology of Familiarization**

A leading psychological analysis of familiarization assumes that infants build categories for stimuli (Cohen, 1973; Sokolov, 1963). Subsequently, they ignore stimuli that correspond to their categories, and concentrate on stimuli that are relatively novel. These processes are often discussed in terms of recognition memory. If there is substantial recovery to a novel test stimulus, then it is considered novel. But if there is little or no recovery, then the stimulus is considered to be recognized as a member of a familiar category. During familiarization there is typically an exponential decrease in attention.

### **Familiarization in Neural Networks**

Encoder networks that learn to reproduce their inputs on their output units can simulate familiarization and novelty effects in infants (Mareschal & French, 1997). Relations among stimulus features are encoded in hidden unit representations, and accuracy is tested by decoding these hidden unit representations onto output units. Discrepancy between output and input representations is network error. Familiar stimuli produce less error than novel stimuli, which presumably deserve further learning. Such hidden unit representations enable prototypes, generalization, and pattern completion (Hertz, Krogh, & Palmer, 1991).

### **Other Neural Network Models**

There are at least eight alternative computational models of the Marcus et al. (1999) data, all of them connectionist models, presumably attracted by the challenge that ordinary connectionist models would not be able to simulate the data. Most of these models are ordinary unstructured connectionist models without explicit rules and variables. All eight of these models cover the basic finding of the Marcus et al. (1999) experiments, namely noticing the difference between consistent and inconsistent sentences. It is beyond the scope of this brief paper to thoroughly review all of these models, many of which are as yet only sketchily reported. However, we can briefly characterize each model and identify what we believe to be its best virtue and most significant limitation.

Four of the unstructured models use the SRN architecture, construing the network's task to be prediction of the next word in a sentence. Negishi (1999a, b) used an SRN without

hidden units, coding each word in analog fashion with place of consonant articulation and vowel height. This is a simple network requiring no unusual hand-wired assumptions or pre-experimental experience. However, Marcus (1999a) claimed that it essentially implemented variables by using continuous values on the input units that are transmitted directly to the outputs, thus arguably disqualifying the model from meeting the challenge that variable binding is required.

Following an argument that Marcus et al.'s (1999) SRNs failed because they lacked normal phonemic experience (Seidenberg & Elman, 1999), Elman (1999) pre-trained an SRN to distinguish whether each word differed or not from the previous word. Each word was coded on 12 binary phonetic features. Although 7-month-olds obviously know something about phonemes and it may be reasonable to include such knowledge in models, it is unlikely that infants receive any target signals about phonemic sameness and difference. More seriously, the network's task in both the pre-training and habituation phases of the simulation was discrimination rather than habituation as it was for the infants.

Christiansen and Curtin (1999) pre-trained an SRN on word segmentation. The network learned to predict the identity and stress of the next phoneme in sentences from information on 11 binary phonological features and the stress and utterance boundaries of individual phonemes. Presented with the Marcus et al. test sentences, the network then showed slightly better prediction of words occurring in inconsistent than those occurring in consistent sentences. Again, the use of prior knowledge seems reasonable. However, it is unclear why the network would perform better on inconsistent sentences, with which it is less familiar, than on consistent sentences whose pattern it has just learned.

Altmann and Dienes (1999) used SRNs with an extra encoding layer between the input and hidden layers. Unlike some models, this one does not require any questionable pre-training and is performing the habituation task. On the negative side, Marcus (1999b) reports that only when somewhat unconventional correlation and distance measures are used can the network discriminate between consistent and inconsistent sentences. It would be more typical to measure error or relative output activation for such networks.

Gasser and Colunga (1999) used a specially-designed network with micro-relation units whose activations correlated with inputs from two different syntactic categories. Hard-wired connections caused similar syllables to be synchronized, producing low activations on the micro-relation units, and dissimilar syllables to be desynchronized, producing high activations on the micro-relation units. No pre-training was necessary, but the hardwiring of connection weights is of questionable psychological validity.

Shastri and Chang (1999; Shastri, 1999) designed a structured connectionist model with explicit variable binding, implemented by temporal synchrony of activations on units representing sequential position and other units representing arbitrary binary word features. The network learned to represent an ABA pattern by firing the first position unit synchronously with the third position unit. This network would seem to generalize well to any novel sentences of

three words, regardless of the particular features of the words used. But the network is extensively hand-built, and the critically important feedback signals about the position of words in a sentence are psychologically implausible.

None of the foregoing reports of models include evidence on the course of habituation or provide predictions that could be tested with infants.

Shultz (1999) used an encoder version of the cascade-correlation algorithm with arbitrary analog coding of syllables. With an encoder network, the task is construed as word and sentence recognition. Besides covering the consistency effect, these networks learned the training patterns with an exponential decrease in error and showed occasional reversals of preference that were found with the infants. Because the coding was arbitrary, however, it was not possible to simulate the detailed phonetic differences between Marcus et al.'s (1999) Experiments 1 and 2.

### Our Model

Here we present a simulation like that of Shultz (1999), but with phonetically realistic encoding of the input sentences using a continuous sonority scale. A successful result would suggest that such coding could be used by infants in their sentence processing. Sonority is the quality of vowel likeness, and can be defined by perceptual salience (Price, 1980) or by openness of the vocal tract (Selkirk, 1984). The coding scheme is shown in Table 2. The specific numbers are somewhat arbitrary, but their ordering is based on phonological work (Selkirk, 1984; Vroomen, van den Bosch, & de Gelder, 1998).

Table 2: Sonority scale with examples in IPA.

Phoneme category	Examples	Sonority
low vowels	/a/ /æ/	6
mid vowels	/e/ /e/ /o/	5
high vowels	/i/ /i/ /u/ /u/	4
semi-vowels and laterals	/w/ /y/ /l/	-1
nasals	/n/ /m/	-2
voiced fricatives	/z/ /v/	-3
voiceless fricatives	/s/ /f/	-4
voiced stops	/b/ /d/ /g/	-5
voiceless stops	/p/ /t/ /k/	-6

Sonorities range from -6 to 6 in steps of 1, with a gap and change of sign between the consonants and vowels. Each word was coded on two units for the sonority of its consonant and that of its vowel. This is similar to Negishi's (1999b) coding, except that we place consonants and vowels on a single scale, rather than on separate scales. We coded each sentence in the artificial language with six units, two for each one-syllable word. For example, the sentence *ni la ni* was coded as (-2 4 -1 6 -2 4).

Our learning algorithm, cascade-correlation, grows during learning by recruiting new hidden units into the network as required to reduce error (Fahlman & Lebiere, 1990). Recruited

hidden units are installed each on a separate layer, receiving input from the inputs and from existing hidden units. The candidate hidden unit that gets recruited is the one whose activations correlate best with current error. After recruiting a hidden unit, the network returns to the phase in which weights feeding the output units are adjusted to reduce error. An encoder option to cascade-correlation (Shultz, 1999) freezes direct input-output connections at 0 to prevent trivial solutions in which weights of about 1 are learned between each input unit and its corresponding output unit.

The cascade-correlation algorithm has been used to simulate many other aspects of cognitive development, including the balance scale (Shultz, Mareschal, & Schmidt, 1994), conservation (Shultz, 1998), seriation (Mareschal & Shultz, 1999), discrimination shift learning (Sirois & Shultz, 1998), pronoun semantics (Oshima-Takane, Takane, & Shultz, 1999), and integration of velocity, time, and distance cues (Buckingham & Shultz, in press).

In these models, network behavior becomes rule-like with learning, but knowledge is clearly not represented in rules and cognitive processing is definitely not accomplished by explicit variable binding and rule firing. Instead, rules are viewed as abstract, epi-phenomenal characterizations of processes occurring at the sub-symbolic level of unit activations and connection weights (Smolensky, 1988).

There are several advantages of implementing rule-like behavior with neural processes, including the acquisition of psychologically realistic non-normative rules, integration of perceptual and cognitive phenomena, natural variation across problems and individuals, and achievement of the right degree of crispness in knowledge representations. In many cases, universally quantified rules are too crisp to model knowledge representations in children.

Neurological justification for generative networks such as cascade-correlation is provided by recent findings on learning-driven neurogenesis and synaptogenesis throughout the lifespan (Quartz & Sejnowski, 1997). Although neurogenesis and neural migration may be too slow to account for learning within the time frame of the typical infant familiarization experiment, there is evidence that synaptogenesis can occur within seconds (Bolshakov, Golan, Kandel, & Siegelbaum, 1997).

Like most models of higher cognition, cascade-correlation is not a model of detailed neural circuits. Instead, it is an abstracted and simplified model that is partly inspired by neural principles. Individual units in cascade-correlation networks may correspond roughly to groups of biological neurons, and connection weights may correspond roughly to neural pathways.

### Results

Mean network error on test patterns for the three experiments is shown in Table 3. Main effects of consistency were significant at  $p < .0001$ . The results show more network error to inconsistent test patterns than to consistent test patterns

for each experiment. On the assumption that error represents a need for further cognitive processing, these results capture the infant data.

Table 3: Mean error on test patterns.

Expt.	Patterns	Consistent	Inconsistent
1	ABA v. ABB	8.2	14.5
2	ABA v. ABB	13.1	15.8
3	AAB v. ABB	12.9	15.3

The proportion of networks showing a reversal of the consistency effect was .0667, which is close to the .0625 obtained with infants.

A plot of mean error over epochs for a representative network from the ABB condition of Experiment 1 is shown in Figure 1. The first few epochs are omitted for clarity because error started so high, at around 350. Such plots reveal exponential decreases in error on the training patterns over time, similar to the shape of declining attention in infant familiarization. The epochs at which hidden units are installed are shown with diamond shapes just above the training error. As in most cascade-correlation simulations, error decreases sharply after a hidden unit is recruited. After training, error is higher on inconsistent test patterns than on consistent test patterns.

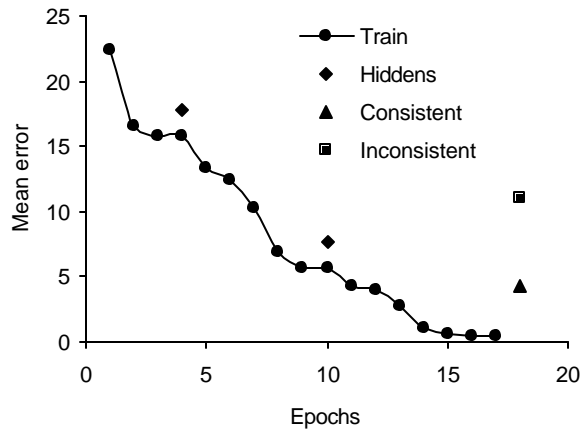


Figure 1: Error reduction in one network.

Generalization tests show that the consistency effect actually grows larger with increasing distance from the training set, a prediction quite different than universally quantified rules would make.

Network analysis revealed that hidden units used sonority sums of consonant and vowel to represent sonority variation first in the duplicated-word category and second in the single-word category. Networks decoded this hidden unit representation with virtually duplicate weights to outputs representing the duplicate-word category.

## Discussion

Like other neural models, our model easily captures the consistency effect. In contrast to alternate models of these data, ours has several features to recommend it. Our model does not require extensive pre-experiment experience (Christiansen & Curtin, 1999; Elman, 1999), extensive hand-wiring of networks (Gasser & Colunga, 1999; Shastri & Chang, 1999), external feedback signals not available in the stimuli (Elman, 1999; Shastri & Chang, 1999), unusual interpretation of outputs (Altmann & Dienes, 1999), or explicit variable binding (Shastri & Chang, 1999). On grounds of theoretical parsimony, the more unsupported assumptions that a model requires the less plausible it becomes.

Unlike some alternate models (Shastri & Chang, 1999; Shultz, 1999), our model uses a realistic coding of the stimuli. Like Negishi (1999b), we used an analog coding of inputs based on the manner in which the phonemes are produced. But our representation scheme is a bit more compact and uniform because we use a single sonority scale for both consonants and vowels, whereas he used two separate scales, one for place of consonant articulation and another for vowel height. Moreover, our use of hidden units with non-linear transfer functions ensures that any possible variable binding at the input level is lost as activation is propagated forward through the hidden layers.

Our model is the only one so far to capture the other feature of the Marcus et al. (1999) infant data, the occasional reversal of preference for novel patterns. It is unclear how easily other models might be able to capture these reversals, but there are hints that it might be difficult for some models. Elman's (1999) model, for example, had such a strong consistency effect that reversals of preference would be unlikely: mean activation to ABB sentences was 123 times higher than to ABA sentences. Likewise, the Shastri and Chang (1999) model learns a very strong representation of serial position. The correlation between weights to position nodes were .9993 for positions 1 and 3 in networks habituated to ABA sentences, and .9998 for positions 2 and 3 in networks habituated to ABB sentences. This rather crisp representation produced 3.4 times more error to inconsistent than to consistent sentences in the ABA condition of Experiment 1, which would seem to preclude reversals.

Although it is not known why infants show occasional reversals, our simulations show that they can be a natural part of learning. With limited exposure, as in both the psychological experiments and our simulations, exceptions naturally occur. This is a parsimonious explanation of reversals because it does not require assumptions of any extraneous processes.

In summary, our model might be currently preferred because it covers more of the infant data, with less pre-experimental experience, less network design, and more realistic stimulus coding than alternate models. It also uses a general learning algorithm that has been applied successfully to several other phenomena in cognitive development.

With so many successful neural models of the consistency effect, there is no question that ordinary, unstructured neural networks can cover these data. The modeling shows

that some of the functionality of symbolic rules and variable binding can be constructed from sub-symbolic processes without having to be explicitly built in. The time is now ripe to generate and test predictions from these alternate models.

### Acknowledgments

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. We thank David Buckingham, Jacques Katz, Yuriko Oshima-Takane, Sylvain Sirois, and Yoshio Takane for comments.

### References

- Altmann, G. T. M., & Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. *Science, 284*, 875.
- Bolshakov, V. Y., Golan, H., Kandel, E. R., & Siegelbaum, S. A. (1997). Recruitment of new sites of synaptic transmission during the cAMP-dependent late phase of LTP at CA3-CA1 synapses in the hippocampus. *Neuron, 19*, 635-651.
- Buckingham, D., & Shultz, T. R. (in press). The developmental course of distance, time, and velocity concepts: A generative connectionist model. *Journal of Cognition and Development*.
- Christiansen, M. H., & Curtin, S. L. (1999). The power of statistical learning: No need for algebraic rules. *Proceedings of the Twenty-first Annual conference of the Cognitive Science Society* (pp. 114-119). Mahwah, NJ: Erlbaum.
- Cohen, L. B. (1973). A two-process model of infant visual attention. *Merrill-Palmer Quarterly, 19*, 157-180.
- Elman, J. L. (1999). Generalization, rules, and neural networks: A simulation of Marcus et al. [www.crl.ucsd.edu/~elman/Papers/MVRVsim.html](http://www.crl.ucsd.edu/~elman/Papers/MVRVsim.html)
- Fahlman, S. E., & Lebiere, C. (1990). The Cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems 2* (pp. 524-532). Los Altos, CA: Morgan Kaufmann.
- Gasser, M., & Colunga, E. (1999). Babies, variables, and connectionist networks. *Proceedings of the Twenty-first Annual conference of the Cognitive Science Society* (p. 794). Mahwah, NJ: Erlbaum.
- Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Reading, MA: Addison Wesley.
- Marcus, G. F. (1999a). Do infants learn grammar with algebra or statistics? *Science, 284*, 433.
- Marcus, G. F. (1999b). Response: Rule learning by seven-month-old infants and neural networks. *Science, 284*, 875.
- Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283*, 77-80.
- Mareschal, D. & French, R. M. (1997). A connectionist account of interference effects in early infant memory and categorization. *Proceedings of the 19th annual conference of the Cognitive Science Society* (pp. 484-489). Mahwah, NJ: LEA.
- Mareschal, D., & Shultz, T. R. (1999). Development of children's seriation: A connectionist approach. *Connection Science, 11*, 149-186.
- Negishi, M. (1999a). Do infants learn grammar with algebra or statistics? *Science, 284*, 433.
- Negishi, M. (1999b). Rule learning by seven-month-old infants and by a simple-recurrent-network. [www.cns-web.bu.edu/pub/mnx/sci.html](http://www.cns-web.bu.edu/pub/mnx/sci.html)
- Oshima-Takane, Y., Takane, Y., & Shultz, T. R. (1999). The learning of first and second pronouns in English: Network models and analysis. *Journal of Child Language, 26*, 545-575.
- Price, P.J. (1980). Sonority and syllabicity: Acoustic correlates of perception. *Phonetica, 37*, 327-343.
- Quartz, S. R., & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioural and Brain Sciences, 20*, 537-596.
- Seidenberg, M. S., & Elman, J. L. (1999). Do infants learn grammar with algebra or statistics? *Science, 284*, 433.
- Selkirk, E.O. (1984). On the major class features and syllable theory. In M. Aronoff & R.T. Oehrle (Eds). *Language sound structure* (pp. 107-136). Cambridge MA: MIT Press.
- Shastri, L. (1999). Infants learning algebraic rules. *Science, 285*, 1673.
- Shastri, L., & Chang, S. (1999). A spatiotemporal connectionist model of algebraic rule-learning. TR-99-011. International Computer Science Institute, Berkeley, CA. [www.icsi.berkeley.edu/~shastri/babytalk](http://www.icsi.berkeley.edu/~shastri/babytalk)
- Shultz, T. R. (1998). A computational analysis of conservation. *Developmental Science, 1*, 103-126.
- Shultz, T. R. (1999). Rule learning by habituation can be simulated in neural networks. *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society* (pp. 665-670). Mahwah, NJ: Erlbaum.
- Shultz, T. R., Mareschal, D., & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning, 16*, 57-86.
- Shultz, T. R., Oshima-Takane, Y., & Takane, Y. (1995). Analysis of unstandardized contributions in cross connected networks. In D. Touretzky, G. Tesauro, & T. K. Leen, (Eds). *Advances in Neural Information Processing Systems 7* (pp. 601-608). Cambridge, MA: MIT Press.
- Sirois, S., & Shultz, T. R. (1998). Neural network modeling of developmental effects in discrimination shifts. *Journal of Experimental Child Psychology, 71*, 235-274.
- Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*, 1-74.
- Sokolov, E. N. (1963). *Perception and the conditioned reflex*. Hillsdale, NJ: Erlbaum.
- Vroomen, J., van den Bosch, A., & de Gelder, B. (1998). A connectionist model for bootstrap learning of syllabic structure. *Language and Cognitive Processes, 13*, 193-220.

# Simulation of Self-affirmation Phenomena in Cognitive Dissonance

**Thomas R. Shultz** (shultz@psych.mcgill.ca)

Department of Psychology; McGill University  
Montreal, QC H3C 1B1 Canada

**Mark R. Lepper** (lepper@psych.stanford.edu)

Department of Psychology; Stanford University  
Stanford, CA 94305-2130 USA

## Abstract

The consonance constraint-satisfaction model, which has simulated the major paradigms of classical cognitive dissonance theory, is here extended to deal with more contemporary findings concerning self-affirmation phenomena in dissonance reduction. The key addition to the model, which has also figured in recent simulations of arousal phenomena, is to lessen activity level within the neural network model in self-affirmation conditions. These and other simulations continue to show that dissonance phenomena can be explained in terms of constraint satisfaction.

## Introduction

One of the fundamentally important theories in social psychology is cognitive dissonance theory, which has generated a literature of more than 1000 studies over the past 40 years (Festinger, 1957; Thibodeau & Aronson, 1992). We have recently modeled a number of the central dissonance phenomena using constraint-satisfaction neural networks (Shultz & Lepper, 1996, 1998a&b, 1999a&b). Our so-called consonance model covered insufficient justification, free choice, arousal, and some self-concept phenomena. The model also predicted new free-choice effects that were subsequently confirmed by further psychological experimentation (Shultz, Léveillé, & Lepper, 1999). In this paper, we report on an extension of the model to deal with a prominent self-concept effect in dissonance called self-affirmation.

Dissonance is hypothesized to occur when behavior is inconsistent with self-concept (Steele, 1988; Thibodeau & Aronson, 1992). Because most people have a positive self-concept, behaviors such as lying or trying to persuade others of a position that one does not agree with arouse dissonance and lead to attitude change that reduces the dissonance. However, if important aspects of the self-concept have been recently affirmed, even aspects irrelevant to an experimentally induced inconsistency, there may be no need to reduce dissonance via attitude change. Steele (1988) presented experiments in which fairly subtle self-affirmation manipulations eliminated dissonance effects. Some of these experiments concern insufficient justification via forced compliance, and others deal with free choice. We return to these experiments after reviewing the consonance model used in the simulations.

## The Consonance Model

The consonance model holds that dissonance reduction is a constraint satisfaction problem. The motivation to reduce dissonance stems from the various soft constraints on the beliefs and attitudes that an individual holds. A consonance network corresponds to a person's representation of the situation created in the conditions of a dissonance experiment. Activations of network units represent the direction and strength of a person's cognitions. Weights between cognitions represent psychological implications. These unit activations and weights may vary across the different conditions of a single experiment.

Consonance is the degree to which similarly evaluated units are linked by excitatory weights and oppositely valued units are linked by inhibitory weights. More formally, consonance in a network is defined by

$$\text{consonance} = \sum_i \sum_j w_{ij} a_i a_j$$

where  $w_{ij}$  is the weight between units  $i$  and  $j$ ,  $a_i$  is the activation of the receiving unit  $i$ , and  $a_j$  is the activation of the sending unit  $j$ .

Activation spreads over time cycles by two update rules:

$$a_i(t+1) = a_i(t) + \text{net}_i(\text{ceiling} - a_i(t)) \text{ when } \text{net}_i \geq 0$$

$$a_i(t+1) = a_i(t) + \text{net}_i(a_i(t) - \text{floor}) \text{ when } \text{net}_i < 0$$

where  $a_i(t+1)$  is the activation of unit  $i$  at time  $t+1$ ,  $a_i(t)$  is the activation of unit  $i$  at time  $t$ ,  $\text{ceiling}$  is the maximum activation,  $\text{floor}$  is the minimum activation, and  $\text{net}_i$  is the net input to unit  $i$ , defined as:

$$\text{net}_i = \text{resist}_i \sum_j w_{ij} a_j$$

where  $\text{resist}_i$  refers to the resistance of receiving unit  $i$  to having its activation changed.

At each time cycle,  $n$  units (normally the number of units in the network) are randomly selected and updated. The update rules ensure that consonance increases or stays the same across cycles. Consonance increases because positive net inputs drive unit activations toward the ceiling and negative net inputs drive them toward the floor. Consonance increases until units reach extreme values or net inputs fall to 0. When consonance reaches asymptote, updating stops.

Consonance networks are hand-built to implement particular dissonance experiments using a set of five principles that map dissonance theory to the consonance model:

1. A cognition is implemented by the net activation of a pair of negatively connected units, one of which represents the positive aspect and the other the negative aspect of the cognition.
2. Cognitions are connected to each other based on their causal implications.
3. Dissonance is the negative of consonance divided by the number of nonzero inter-cognition relations.
4. Networks settle into more stable, less dissonant states as unit activations are updated.
5. Unit activations, but not connection weights, are allowed to change, and some cognitions are more resistant to change than others. In particular, beliefs, behaviors, and justifications are more resistant to change than are evaluations and attitudes.

Additional details about the consonance model and its assumptions are available in our previous papers (Shultz & Lepper, 1996, 1998a).

### Forced Compliance

Forced compliance is the most popular dissonance technique within the most prominent dissonance paradigm of insufficient justification. Insufficient justification concerns cases in which a person does something inconsistent with his or her attitudes without much justification. The less the justification, the more cognitive dissonance is created.

In a forced-compliance experiment (Steele, 1988, p. 272), college students were selected for their strong opposition to an increase in tuition fees. They were then persuaded to write essays supporting a substantial tuition increase. In one condition, they were given a choice about whether to write the essay; in another condition, they were given very little choice about whether to write the essay. When a person freely agrees to argue against personal beliefs, this creates dissonance, which can be reduced by changing attitudes in the direction of the argument. There should be little or no dissonance when one is pressured to make such arguments.

Before measuring post-experimental attitudes, some participants were first asked to complete the political sub-scale of the Allport-Vernon Study of Values. One-half of them had been previously assessed as having a strong economic-political value orientation, whereas the others did not have this value orientation. Completing the political value scale was supposed to affirm a valued self-concept only for those students with a strong economic-political value orientation.

As shown by the solid line in Figure 1, there was the familiar dissonance effect of more attitude change under high choice than under low choice. Moreover, as predicted, self-affirmation eliminated attitude change, even under high choice conditions. Two other experiments with minor variations yielded similar results (Steele, 1988).

### Method

Network specifications for the three conditions are shown in Table 1. There are two relevant cognitions, attitude and es-

say, and relations between them. As in our previous simulations, each cognition is implemented with a pair of negatively related units, one to represent the positive aspect of the cognition and the other to represent the negative aspect. Net activation for a cognition is computed as activation on the positive unit minus activation on the negative unit. Positive relations between cognitions are implemented by positive weights between their positive units and between their negative units, and negative weights between the positive unit of one cognition and the negative unit of the other cognition. All weights are bi-directional.

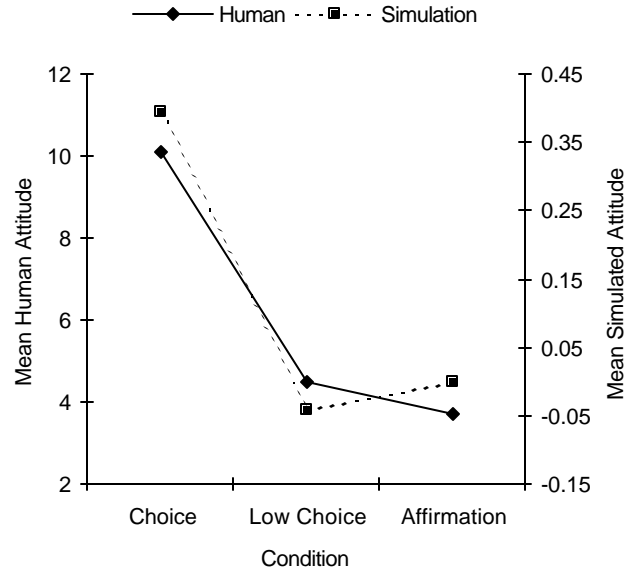


Figure 1: Mean attitude following forced compliance.

All weights and initial unit activations are assigned either high (0.5) or low (0.1) values, according to the five mapping principles described earlier and the descriptions of the experiments being modeled. The floor parameter is 0; the ceiling parameter for positive units is set to 1, and that for negative units is set to 0.5. A *cap* parameter is set to -0.5. This corresponds to the value of the weight between each unit and itself and it prevents activations from growing to ceiling. The *resist* parameter is set to 0.5 for low resistance, and 0.01 for high resistance. These parameter settings are standard across all our dissonance simulations, and some justification for them is provided in our longer papers, (Shultz & Lepper, 1996, 1998a, 1999a).

Table 1: Network specifications for forced compliance.

Condition	Attitude	Essay	Relation
Choice	-0.5	0.5	0.5
Low Choice	-0.5	0.5	0.1
Affirmation	-0.25	0.25	0.25

In this experiment, there is a positive relation between attitude and essay because the more positive one's attitude toward tuition increases, the more likely one would be to



agree to write an essay in favor of tuition increases. This relation is high in the choice condition and low in the low-choice condition. Initially, attitude is given a high negative value to reflect students' initial attitudes; and essay is given a high positive value because the essay was indeed written by all students. An activity-level scalar of 0.5 (the same value used in our other simulations of arousal and self concept) reduces initial activations and weights in the self-affirmation condition, relative to the no-affirmation conditions. The theoretical justification for using a scalar in this way is that self-affirmation is hypothesized to reduce the importance of a dissonant situation (Steele, 1988, p. 292).

All initial unit activations and weights are randomized for each network by adding or subtracting a random proportion of their initial amounts. The three proportion ranges in which additions or subtractions are randomly selected under a uniform distribution are .1, .5, and 1. This increases psychological realism because not everyone can be expected to share the same parameter values. It also allows a test of robustness of the model. Twenty networks were run in each condition at these three different levels of parameter randomization. Networks were run for 30 cycles, which was sufficient to approach asymptotic activation levels.

## Results

Mean attitude toward the view espoused in the essay is presented, in the dashed line in Figure 1, for networks at the .5 level of parameter randomization. As with Steele's (1988) subjects, attitudes are more positive under choice than under the other two conditions. An ANOVA with condition as the single factor revealed significant main effects of condition,  $F(2, 57) = 67, p < .001$ . A contrast  $F$  with weights of +2 for choice, -1 for low choice, and -1 for self-affirmation is significant  $F(1, 57) = 135, p < .001$ , with no significant residual,  $F(1, 57) < 1$ . Proportion of total variance accounted for by this  $F$  is .99.

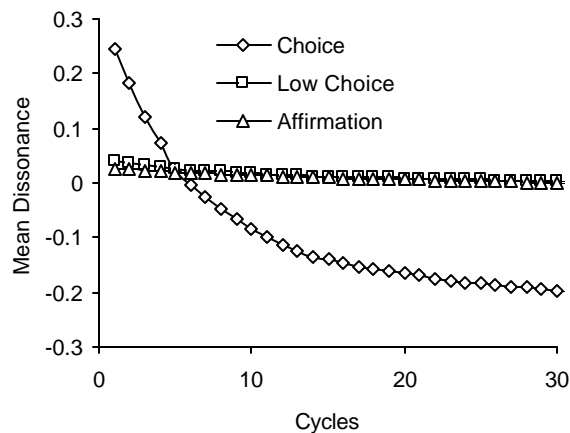


Figure 2: Mean dissonance following forced-compliance.

Mean dissonance scores over time cycles, for networks run at .5 parameter randomization for the three conditions, are shown in Figure 2. Dissonance starts high in the choice

condition and is greatly reduced over time. In contrast, there is minimal dissonance in the other two conditions and very little dissonance reduction. Similar results were obtained at parameter randomization levels of .1 and 1.

## Discussion

The consonance networks provide a good fit to the attitude change data reported by Steele (1988). There is considerable attitude change in the choice condition, but very little in the low-choice and self-affirmation conditions. There is also a close correspondence between amount of attitude change and plots of dissonance reduction in that the condition with sharp dissonance reduction is also the one with the most attitude change. Examination of dissonance plots is a bonus of computer simulations -- there is no known way to measure dissonance directly in humans. Such plots of simulated dissonance can help to understand the more indirect attitude-change effects that occur as a way of reducing dissonance.

### Free Choice

Steele (1988, p. 276) also presents a free-choice experiment that shows self-affirmation effects. Participants rated and ranked 10 music albums and were then given a choice to keep either their fifth- or sixth-ranked album. Choosing between qualitatively distinct objects creates dissonance because the chosen object is less than perfect and the rejected object has some desirable features that are forgone when an irreversible choice is made. The dissonance arising from a free choice is typically reduced by increasing evaluation of the chosen object and decreasing evaluation of the rejected object (Brehm, 1956; Shultz et al., 1999).

In Steele's experiment, one-half of the participants had been previously selected for having a strong scientific-value orientation and for indicating that a lab coat symbolized these values. The others did not share these values. One-half of the participants in each of these groups were asked to wear a lab coat for the rest of the experiment, during which they rated the albums again, after making their choices.

Post-decisional spread of alternatives was measured by adding the increase in the value of the chosen item and the decrease in the value of the rejected item. There were three control conditions, one with participants not having a science orientation and not wearing a lab coat, another with participants not having a science orientation but wearing a lab coat, and a third with participants having a science orientation but not wearing a lab coat. There were identical dissonance effects in all three control conditions, but not for the self-affirmed, scientifically-oriented students wearing a lab coat. Mean spread of alternatives was higher in the control conditions than in the self-affirmation condition, as shown by the solid line in Figure 3. Once again, apparently irrelevant self-affirmation precluded dissonance reduction.

## Method

Network specifications for these two groups of conditions are shown in Tables 2 and 3. There are three cognitions: a decision and evaluations of the chosen and the rejected objects. Because the decision is public and irreversible, it has

high resistance and high initial activation; the two evaluations have low resistance. Initial evaluation of the chosen object is somewhat higher than that for the rejected object because people generally choose items that they rate higher.

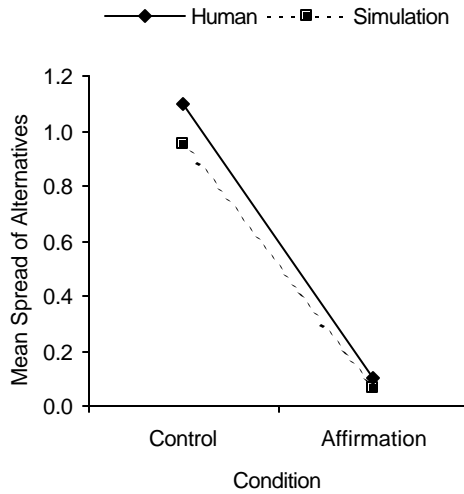


Figure 3: Mean spread of alternatives following free choice.

The relation between the decision and the chosen object is positive because the better-liked object is chosen. The two objects are negatively related because they compete for an exclusive choice. Both relations have high values in the control condition. To implement self-affirmation, initial activations and weights are scaled by .5. Networks in each condition were run for 40 cycles, which was sufficient for saturation. As is customary in our simulations, all weights and initial unit activations were randomized at up to .1, .5, or 1 of the values shown in Tables 2 and 3. Other parameter settings are also the same as in our other dissonance simulations.

Table 2: Initial net activations for free choice.

Cognition	Condition	
	Control	Affirmation
Chosen	.30	.15
Rejected	.20	.10
Decision	.50	.25

## Results

Spread between evaluations of the two choices was computed as in Steele (1988). Change in evaluation of each object is the difference between initial evaluation and evaluation after 40 cycles. Spreading of alternatives is the sum of the increase in evaluation of the chosen alternative and the decrease in evaluation of the rejected alternative. Mean spreading of the alternatives is plotted, on the dashed line in Figure 3, at the .5 level of parameter randomization. There is a larger spread of the alternatives in the control than in the self-affirmation condition,  $F(1, 38) = 76, p < .001$ .

Mean dissonance scores across time cycles in networks at .5 parameter randomization are shown in Figure 4 for the two conditions. Although dissonance starts low in both condi-

tions, it drops only in the control condition. Similar results were found at parameter randomizations of .1 and 1.

Table 3: Relations between cognitions for free choice.

Relation of chosen to Rejected	Condition	
	Control	Affirmation
Decision	.50	.25
Rejected	-.50	-.25

## Discussion

Consonance networks yield greater separation of alternatives in the control than in the self-affirmation condition, as found with human participants (Steele, 1988). Dissonance reduction is also greater in the control than in the self-affirmation condition, consistent with the idea that attitude change is driven by dissonance reduction.

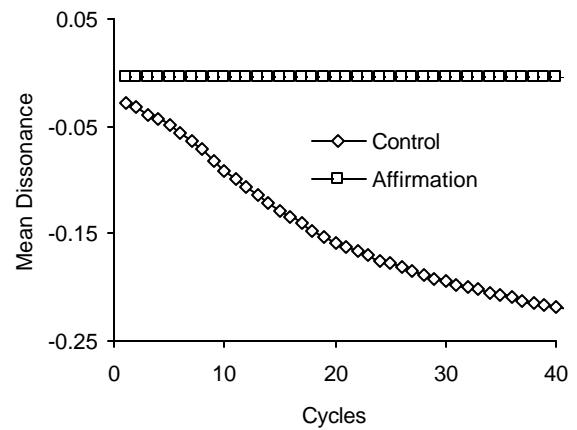


Figure 4: Mean dissonance following free choice.

## General Discussion

These simulations extend the consonance model to rather subtle aspects of dissonance reduction involving the self-concept, using the same conventions, mapping principles, and default parameter values as in previous simulations. In all of these cases, dissonance arises when constraints between simultaneously held cognitions are unsatisfied. Dissonance is reduced as the constraints are satisfied, typically by changing evaluations of entities in the situation defined by the dissonance experiment. The self-affirmation phenomena considered here had not previously been simulated and were not generally seen as being closely related to other contemporary dissonance phenomena on emotional arousal. As in earlier simulations, the consonance model is here shown to be robust against parameter variation, as revealed by the fact that even a high degree of parameter randomization does not affect the pattern of overall results.

A key, unifying concept in simulating contemporary dissonance phenomena in self-concept and arousal is that of activity level. An activity scalar adjusts the overall level of activation in networks that represent dissonant situations. In the present simulations, the activity-level scalar operates

much like a tranquilizing drug in arousal simulations (Shultz & Lepper, 1999b), by decreasing activation of the representation of the dissonant situation.

Self-affirmation manipulations are thus hypothesized to decrease the relative importance of being in a dissonant situation. When you feel good about yourself, being in a dissonant situation is not nearly so bothersome, and you become immune to the effects of dissonance reduction. This reveals a somewhat unexpected theoretical communality between arousal and self-concept effects.

This analysis is consistent with recent results on *trivialization* as a mode of dissonance reduction (Simon, Greenberg, & Brehm, 1995). Merely making salient to participants asked to write counter-attitudinal essays the contrast between issues they believe to be of great consequence and the less important topic of their own essays reduces attitude change in the direction of the position advocated.

At the level of the brain or an artificial neural network, the key theoretical notion is that of activity level. Dissonance effects are enhanced by increases in activity level and dampened by decreases in activity level. There are a variety of ways to modulate activity level, including general manipulations such as drugs (Cooper, Zanna, & Taves, 1978) and specific manipulations such as attention to particular cognitions (Read & Miller, 1998a). Consequently, activity level has the potential to unify theoretical understanding of several apparently different dissonance phenomena.

The general success of the consonance model enables a theoretical reinterpretation of dissonance that stresses commonalities with other psychological phenomena that result from constraint satisfaction. Phenomena such as analogical reasoning, person perception, schema completion, attitude change, and dissonance reduction can all be understood in terms of the dynamics of constraint satisfaction (Holyoak & Thagard, 1989; Read & Miller, 1998a, b; Rumelhart, Smolensky, McClelland, & Hinton, 1986; Spellman & Holyoak, 1992; Spellman, Ullman, & Holyoak, 1993; Thagard, 1989).

### Acknowledgments

This research was supported by a grant to the first author from the Social Sciences and Humanities Research Council of Canada and by grant MH-44321 to the second author from the U.S. National Institute of Mental Health.

### References

- Brehm, J. W. (1956). Post-decision changes in the desirability of choice alternatives. *Journal of Abnormal and Social Psychology, 52*, 384-389.
- Cooper, J., Zanna, M. P., & Taves, P. A. (1978). Arousal as a necessary condition for attitude change following forced compliance. *Journal of Personality and Social Psychology, 36*, 1101-1106.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson.
- Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13*, 295-355.
- Read, S. J., & Miller, L. C. (1998a). On the dynamic construction of meaning: An interactive activation and competition model of social perception. In S. J. Read & L. C. Miller (Eds.), *Connectionist models of social reasoning and social behavior* (pp. 27-68). Hillsdale, NJ: Erlbaum.
- Read, S. J., & Miller, L. C. (Eds.). (1998b). *Connectionist models of social reasoning and social behavior*. Hillsdale, NJ: Erlbaum.
- Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. (1986). Schemata and sequential thought processes in PDP models. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 7-57). Cambridge, MA: MIT Press.
- Shultz, T. R., & Lepper, M. R. (1996). Constraint satisfaction modeling of cognitive dissonance phenomena. *Psychological Review, 103*, 219-240.
- Shultz, T. R., & Lepper, M. R. (1998a). The consonance model of dissonance reduction. In S. J. Read & L. C. Miller (Eds.), *Connectionist models of social reasoning and social behavior* (pp. 211-244). Hillsdale, NJ: Erlbaum.
- Shultz, T. R., & Lepper, M. R. (1998b). A constraint-satisfaction model of Machiavellianism effects in cognitive dissonance. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 957-962). Hillsdale, NJ: Erlbaum.
- Shultz, T. R., & Lepper, M. R. (1999a). Computer simulation of cognitive dissonance reduction. In E. Harmon-Jones & Mills, J. (Eds.), *Cognitive dissonance: Progress on a pivotal theory in social psychology* (pp. 235-265). Washington, DC: American Psychological Association.
- Shultz, T. R., & Lepper, M. R. (1999b). Consonance network simulations of arousal phenomena in cognitive dissonance. *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society* (pp. 659-664). Hillsdale, NJ: Erlbaum.
- Shultz, T. R., Léveillé, E., & Lepper, M. R. (1999). Free choice and cognitive dissonance revisited: Choosing "lesser evils" vs. "greater goods." *Personality and Social Psychology Bulletin, 25*, 40-48.
- Simon, L., Greenberg, J., & Brehm, J. (1995). Trivialization: The forgotten mode of dissonance reduction. *Journal of Personality and Social Psychology, 68*, 247-260.
- Spellman, B. A., & Holyoak, K. J. (1992). If Saddam is Hitler, then who is George Bush? Analogical mapping between systems of social roles. *Journal of Personality and Social Psychology, 62*, 913-933.
- Spellman, B. A., Ullman, J. B., & Holyoak, K. J. (1993). A coherence model of cognitive consistency: Dynamics of attitude change during the Persian Gulf war. *Journal of Social Issues, 49*, 147-165.
- Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 261-302). New York: Academic Press.
- Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435-502.
- Thibodeau, R., & Aronson, E. (1992). Taking a closer look: Reasserting the role of the self-concept in dissonance theory. *Personality and Social Psychology Bulletin, 18*, 591-602.

# Linguistic Labels and the Development of Inductive Inference

Vladimir M. Sloutsky (sloutsky.1@osu.edu)

Center for Cognitive Science & School of Teaching & Learning, The Ohio State University  
21 Page Hall, 1810 College Road, Columbus, OH 43210, USA

Ya-Fen Lo (lo.37@osu.edu)

Center for Cognitive Science & School of Teaching & Learning, The Ohio State University  
21 Page Hall, 1810 College Road, Columbus, OH 43210, USA

## Abstract

The paper presents a model suggesting that inductive generalizations in young children could be a function of similarity among compared stimuli. Predictions derived from the model were tested in two experiments where young children and preadolescents were presented with triads of schematic faces (a Target and two Test stimuli) that varied in perceptual similarity, with one of the Test stimuli sharing a linguistic label with the Target. Participants were taught a biological property about the Target and asked to generalize the property to one of the Test stimuli. Results from both experiments support predictions, indicating that for young children, proportions of label-based generalizations varied with featural overlap among the compared stimuli. There were also developmental differences found in effects of labels: while for young children these effects varied with featural overlap, preadolescents relied solely on linguistic labels when performing inductive generalizations.

## Introduction

Inductive generalization is an important component of human thought. Furthermore, some believe that it the most important component because "inductive inference is the only process... by which new knowledge comes into the world" (Fisher, 1935). Therefore, understanding of the development of induction is an important step in understanding of human thought.

One theoretical proposal suggests that induction starts out as a category-based process (see Gelman & Coley, 1991; Gelman, Coley, & Gottfried, 1994, for reviews and discussions). In this case, independent of similarity, generalization within a theoretically-defined category is more likely than generalization across categories. For example, a person is more likely to generalize a property (e.g., the ability to drink) from a bird to another dissimilar looking bird than from a bird to a similar looking airplane (Mandler & McDonough, 1998; Gelman & Markman, 1986; Gelman & Coley, 1991).

The alternative, similarity-based approach, suggests that induction starts out as a special case of the "universal law of generalization" (Shepard, 1987). The law states that the probability of generalizing a response (e.g., fear) from one stimulus to another stimulus varies with featural similarity between the stimuli.

Although the similarity-based approach seems to be more

appealing on the basis of parsimony, it has been often criticized for the failure to constrain the notion of similarity (e.g., Goodman, 1992/1972). Indeed, with the increase of the complexity of predicate structure, it becomes unclear which of these predicates will be used in computing similarity.

Recently, we proposed a model suggesting that for young children, linguistic labels might be an important constraining factor (Sloutsky & Lo, 1999). In a series of experiments, we demonstrated that linguistic labels have larger weights in similarity judgment of young children than other perceptual attributes. We argued that similarity between stimuli patterns decreases as a function of exponential decay (cf. Estes, 1994; Medin, 1975). That is similarity between two labeled stimuli patterns could be calculated using Equation 1:

$$Sim(i, j) = S_{Label}^{1-L} S_{Vis.attr}^{N-k}$$

where  $N$  denotes the total number of visual attributes,  $k$  denotes the number of matches,  $S_{vis.attr}$  denotes values (weights) of a mismatch on a visual attribute,  $S_{Label}$  denotes values of label mismatches, and  $L$  denotes a label match. When there is a label match,  $L = 1$ , and  $S_{Label} = 1$ ; when there is a label mismatch,  $L = 0$ , and  $S_{Label} < 1$ . Because  $S$  varies between 0 and 1, similarity equals to one when there are no mismatches, otherwise it is smaller than 1.

We also suggested that when a child is presented with a Target feature pattern (T) and Test feature patterns (A and B) and asked which of the Test patterns is more similar to the Target, the probability of choosing B could be predicted using Equation 2:

$$P(B) = \frac{Sim(T, B)}{Sim(T, B) + Sim(T, A)}$$

In this paper, we present evidence that the model can account not only for similarity judgement, but for inductive inference of young children as well.

Of course, it could be argued that reliance of young children on linguistic labels when performing induction is an indicator that they perform induction in a category-based manner, because they use linguistic labels as category markers. There is an important caveat, however. If they rely on linguistic labels as category markers, labels should affect induction in a qualitative "all-or-none" manner (i.e., presence or absence of the shared label should be a critical factor in

induction). We predict an alternative course: labels affect induction in a quantitative manner, in accordance with equations 1 and 2. In other words, proportions of label-based generalizations in young children should vary, with the number of visual attributes shared between the compared entities.

## Experiment 1

### Method

**Participants** A group of 87 children aged 4 to 12 years participated in the study. The participants represented three age groups: (1) 32 four-to-five year-olds ( $M = 4.5$  years,  $SD = 0.56$  years; 14 boys and 18 girls), 30 seven-to-eight year-olds ( $M = 8.1$  years,  $SD = 0.5$  years; 15 boys and 15 girls), and 25 eleven-to-twelve year-olds ( $M = 11.8$  years,  $SD = 0.5$  years; 15 boys and 10 girls). The participants were recruited from daycare centers, elementary and middle schools located in middle class suburbs of Columbus, Ohio.

**Materials** The materials included triads of 2" by 2" schematic faces, two of which were Test stimuli and one of which was a Target. Each schematic face had three distinct attributes (shape of head, shape of ears, and shape of nose), and each attribute had three values (e.g., "curve-lined" nose, "straight-lined" nose, and "angled" nose). These materials were identical to those used in Part 1 experiments (Sloutsky & Lo, 1999). Materials also included 36 artificial bi-syllable labels (e.g., *Bala*, *Gula*, and so forth) and a set of unobservable biological properties of the Target. These properties were as follows:

1. Has pink bones
2. Has green brain
3. Has white heart
4. Has orange stomach
5. Has blue fat
6. Has yellow blood

Participants were asked which of the Test stimuli was more likely to share a biological property with the Target.

**Design and Procedure** The experiment had a mixed design with age and labeling condition (label vs. no-label) as between-subject factors and a stimulus pattern condition as a within-subject variable. For both levels of the labeling condition, participants were presented with the same triads of schematic faces, two of which were Test stimuli and one of which was a Target. The only difference was that in the label condition all stimuli were labeled, whereas in the no-label condition these stimuli were not labeled. The stimulus pattern condition included six levels, T-00, T-11, T-22, T-01, T-12, and T-02. Note that T refers to the Target, the first digit refers to the number of attributes shared by Test B with the Target, and the second digit refers to the number of attributes shared by Test A with the Target. In the label condition, the Target always shared labels with Test B and always had labels different from Test A. A female researcher interviewed children in a quiet room in their schools. Before the experimental task, children were introduced to some warm-up questions and were given feedback. In the warm-up

tasks, children were presented with Test and Target stimuli and were asked to choose the Test stimulus that shared a biological property with the Target.

**Warm up Trials.** In the first warm-up trial, participants were presented with a Target (a shark) and two Test stimuli (a bear and a tree branch). In the second warm-up trial, they were presented with a rabbit as a Target, and an apple and a dog as Test stimuli. In the third warm-up trial, children were presented with a fish as a Target, and a turtle and a spider as Test stimuli. In all these warm-up trials, children were first told that the Target stimuli either had bones, blood, or skeleton inside the body. Children then were asked to determine which of the two Test stimuli has the same thing inside the body as the Target. If a child failed to answer induction questions, the researcher explained how each of the Test stimuli could have the same thing as the Target.

**Experimental Trials.** If a child was capable of giving correct answers in two out of three warm-up trials, the researcher proceeded to the main experiment. No child was eliminated from the study since all participants provided satisfactory responses in at least two out of three warm-up trials. In the Label condition, children were first introduced to the labels for the Target and Test pictures and asked to repeat them. All labels used were the same artificial names (e.g. *Bala*, *Guga*) as in Part 1 experiments. After each stimulus was labeled, children were asked to repeat these labels. No labels were introduced in the no-label condition. Children were then introduced to an unobservable biological property that belonged to the Target stimuli and were asked which of the Test stimuli was likely to have this property. Positions of the two Test pictures were counterbalanced across the experimental trials. After children answered the questions, they were asked to provide their justification for their choices. In both conditions, participants of the two older groups had 24 experimental trials (6 within-subject stimulus patterns with 4 trials each), while participants of the youngest group had 18 experimental trials (6 within-subject stimulus patterns with 3 trials each). This reduction in the number of trials was important to avoid fatigue that could lead to random responding in young children. The order of presenting of stimulus patterns was randomized within participants.

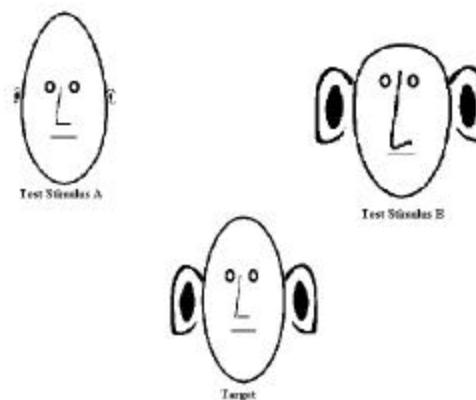


Figure 1: Example of stimuli presented in one trial in the T-1-2 condition: The Target shares the overall shape and the nose with Test A and the size of the ears with Test B.

## Results and Discussion

Results indicate that in the no-label condition, participants of all age groups based their inductive inferences on available perceptual information. Proportions of Test B choices broken down by stimulus pattern condition and age groups are presented in Figure 2. Recall that B-choices refer to the selection of the Test stimulus that in the label condition shares the label with the Target. As predicted (see Table 1), for all indeterminate stimulus pattern conditions, when the Target shared equal numbers of attributes with each Test stimulus (i.e., T-00, T-11, and T-22), the proportions of B-choices for all age groups were at chance (one-sample t-tests, all  $ps > .25$ ). Also as predicted, in all determinate stimulus pattern conditions where Test B shared fewer attributes with the Target than Test A (i.e., T-01, T-12, and T-02), the proportions of B-choices for all age groups were below chance (one-sample t-tests, all  $ps < .01$ ).

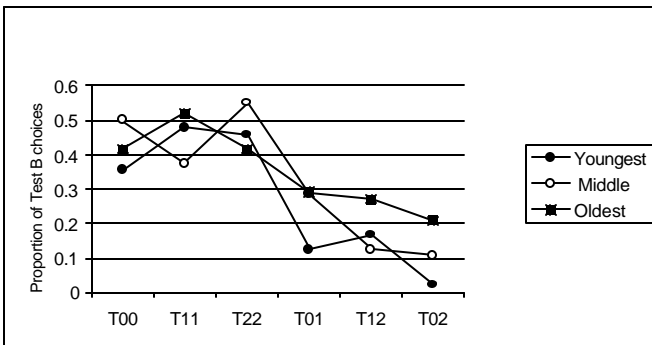


Figure 2: Induction in the No-Label condition broken down by stimulus pattern condition and age group.

To examine the direction of differences among the stimulus pattern conditions in the no-label condition, proportions of B-choices were subjected to a two-way ANOVA with age as a factor and stimulus pattern condition as a repeated measure. Because proportions of B-choices across the T-00, T-11, and T-22 conditions were very similar (all  $ts < 1$ ), proportions of B-choices were averaged across these conditions into a new aggregated variable T-Equal ("T" stands for the Target and "Equal" indicates that each of the Test stimuli shared equal number of features with the Target). While there were no significant differences in proportions of B-choices among the age groups,  $F(2, 35) = .8, p = .45$ , there was a significant main effect due to the stimulus pattern condition,  $F(3, 105) = 13.5, MSE = 0.1, p < .0001$ . Planned comparisons indicated that T-Equal exhibited the largest proportion of B-choices (46%), whereas T-02 exhibited the smallest proportion of B-choices (11%), all  $ts > 2.2, ps < .05$ . At the same time, T-12 and T-01 did not differ significantly,  $t < 1$ . These results indicate that when only perceptual information was available, participants based their inductive inference on this information: in all age groups inductive inference was a function of the number of attributes shared by the Target with Test stimuli.

Introduction of labels, however, dramatically changed the proportions of B-choices that had been observed in the

no-label condition. Recall that in the label condition, Test B always shared the label with the Target. Proportions of B-choices (i.e., label-based generalizations) in the label condition broken down by age group and stimulus pattern condition are presented in Figure 3. In the oldest group, all participants on all trials, with the exception of one participant on one trial, used labels as the only basis of their induction. At the same time, effects of labels in the two younger groups varied across stimulus pattern condition. Because participants in the older group exhibited no variability in their responses (311 out of 312 responses were label-based generalization), while participants in the two younger groups exhibited variability, the former were not included in the analysis of label-based generalizations across stimulus pattern condition.

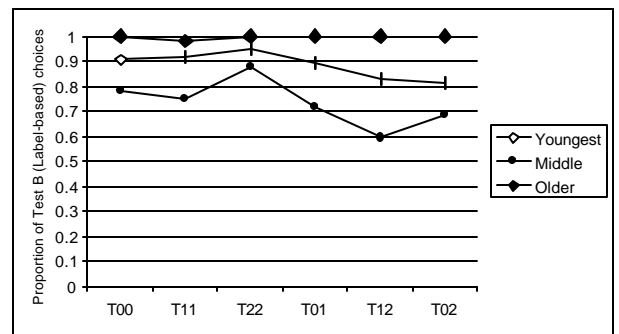


Figure 3: Induction in the Label condition broken down by stimulus pattern condition and age group.

Proportions of label-based choices across stimulus pattern condition in the two younger groups were subjected to a two-way (age by stimulus pattern condition) ANOVA with stimulus pattern as a repeated measure. Because proportions of label-based generalizations across the T-00, T-11, and T-22 conditions were statistically equivalent ( $t < 1$ ), these proportions were averaged across these conditions into a new aggregated variable T-Equal. The analysis indicates a significant main effect due to stimulus pattern condition,  $F(3, 102) = 2.8, MSE = 0.06, p < .05$ , whereas neither main effects of age group, nor the interaction of the two factors were significant ( $p = .15$  and  $p = .8$ , respectively). Planned comparisons pointed to significant differences between the T-Equal condition and the T-12 and T-02 conditions, all  $ts(35) > 2.1, ps < .05$ . In short, as predicted, in the oldest group the proportion of selecting Test B did not vary across the stimulus pattern conditions, whereas this proportion did vary as a function of stimulus pattern condition in the two younger groups. Although differences among the stimulus pattern conditions may appear relatively small (in particular, differences between T-02, on the one hand, and T-01 and T-12, on the other hand, fell short of statistical significance), the direction of these differences, except for the T-02 condition in the middle group, closely match predictions derived from Equations 1 and 2.

Overall fit between predicted probabilities and observed frequencies is presented in Figure 4. Each data point in Figure 4 represents responses of to each stimulus pattern

averaged within age groups. Note that Figure 4 depicts performance of children in the two younger groups in the label and no label conditions. In addition to a high correlation between the predicted and observed probabilities ( $r = .96$ ), the proposed theoretical model accounts for approximately 92% of the observed variance ( $R^2 = .919$ ). These findings support the notion that for younger children labels contribute to specific induction in a quantitative manner and that this contribution varies with the number of attributes shared by Test A and Test B with the Target. Note that Figure 4 does not include performance of the oldest group, because their induction was not derived from Equations 1 and 2. On the contrary, their induction was predicted to be category-based, as opposed to similarity-based induction of younger children.

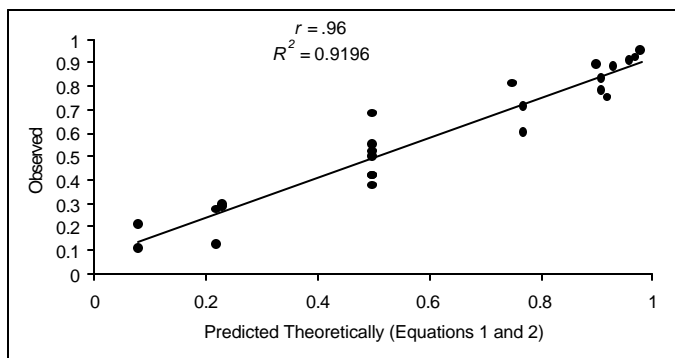


Figure 4: Theoretical probabilities (computed from the model using Equations 1 and 2) and observed probabilities of generalization of biological properties to Test B. Note: parameter  $S$  was estimated from our previous research.

Findings of this experiment support our predictions regarding quantitative contribution of labels to inductive inferences of young children. In the absence of labels, participants of all age groups based their inductive inference on perceptual information. However, when labels were introduced, the pattern of choices changed dramatically: preadolescents based their induction solely on labels, whereas younger children based their induction on a combination of labels and the number of overlapping attributes. Therefore, it seems reasonable to infer that preadolescents performed induction in a category-based manner, whereas younger children performed induction in a similarity-based manner.

However, it could be argued that these findings strongly support predictions for preadolescents' induction, while providing only tentative support of predictions for young children's induction. The support is tentative because some differences among the stimulus pattern conditions, while all in predicted directions, failed to reach significance. Because of this, we deemed it necessary to conduct a second experiment, replicating the current experiment for young children while simplifying the task, and increasing the sample sizes and the number of trials.

## Experiment 2

### Method

**Participants** A group of 30 four-year-old children ( $M = 4.3$  years,  $SD = 0.5$  years; 19 boys and 11 girls) participated in the experiment. The participants were recruited from daycare centers located in middle class suburbs of Columbus, Ohio.

**Materials** The materials included triads of schematic faces identical to those used in Experiment 1.

**Design and procedure** The design and procedure were identical to that in Experiment 1 with three exceptions. First, the current experiment included only a label condition. Second, the "why" questions that accompanied children's choices in Experiment 1 were dropped. Finally, the number of trials within each stimulus pattern condition was increased from three to four.

### Results and Discussion

Proportions of label-based generalizations across the stimulus pattern conditions are presented in Figure 5. Because proportions of label-based generalizations in T-00, T-11, T-22 conditions were statistically equivalent (87%, 85%, and 87% respectively,  $ts < 0.5$ ), participants' responses were averaged across these conditions into a new variable T-Equal. Proportions of label-based generalizations in T-Equal, T-01, T-12, and T-02 conditions were subjected to a one-way repeated measures ANOVA. The analysis points to significant differences among the stimulus pattern conditions,  $F(3, 87) = 16.744$ ,  $MSE = 1.15$ ,  $p < 0.0001$ . Planned comparisons pointed to the following order among the conditions in the proportion of label-based generalizations: T-Equal (86%) > T-12 (63%) = T-01 (56%) > T-02 (39%). All indicated differences were significant, all  $ts > 3.5$ , Bonferroni adjusted  $ps < .01$ , while the difference between T-12 and T-01 was not significant,  $t < 1$ . These results clearly indicate that the proportion of label-based generalizations varied as a function of the number of features shared by the Target with each of the test stimuli, thus further supporting the notion of similarity-based specific induction in young children.

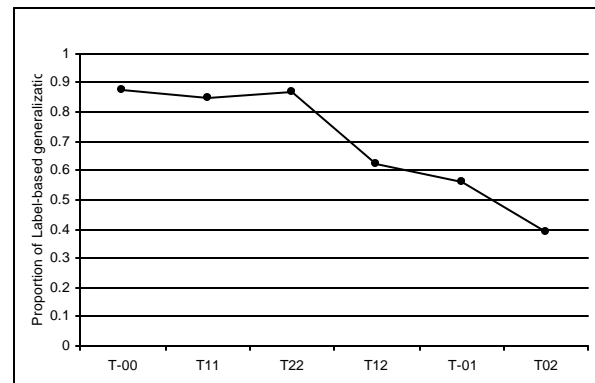


Figure 5: Proportions of label-based generalizations across the stimulus pattern conditions.

## General Discussion

Results of the two reported experiments are as follows. In Experiment 1, when labels were not provided, 4-5 year-olds, 7-8 year-olds, and 11-12 year-olds relied on perceptual similarity when making specific induction with novel entities. At the same time, when labels were introduced, preadolescents made inductive inferences based solely on the basis of the provided label, whereas specific induction of younger children varied with the number of attributes (which includes labels) shared by the Target and Test stimuli. Results of Experiment 2 further indicated that proportions of label-based generalizations in young children varied as a function of visual attributes shared by the Target with each of the Test stimuli. These results fit predictions, indicating that the model proposed by Sloutsky & Lo (1999) can account for specific induction of younger children and that specific induction of the younger children is similarity-based.

The results of the no-label condition indicate that for all three age groups, similarity-based specific induction is the default mechanism (cf. Keil, 1989). When no other information was available, participants of all age groups used perceptual similarity to generalize biological properties from the Target to Test stimuli. Therefore, if similarity-based induction is the default mechanism, it seems likely that it might developmentally precede category-based induction. This contention was supported by results of the label condition.

The results of the label condition supported the notion of different mechanisms underlying specific induction in young children and preadolescents, thus allowing the resolving of an apparent paradox of specific induction. The paradox is as follows. On the one hand, if specific induction is category-based, it should be dependent on general induction and the ability to perform induction-deduction coordination. On the other hand, even three year-olds are capable of performing specific induction (Gelman & Markman, 1987). The reported results suggest that specific induction does not have to be category-based -- it may start out as similarity-based and develop into category-based later. The reported experiments support this notion, suggesting that this shift may occur sometime between nine and eleven years of age. Indeed, while specific induction of 7-8 year-olds appeared to conform to the proposed model and to vary with a number of perceptual attributes shared by the Target and Test stimuli, specific induction of 11-12 year-olds appeared to be independent of shared attributes and to be a function of labels.

It is also important that for younger children, labels exert similar effects on similarity judgment and specific induction. At the same time, in preadolescents these effects are fundamentally different. While labels had no effect on similarity judgment of preadolescents (Sloutsky & Lo, 1999), in specific induction preadolescents relied solely on labels. These findings further support the possibility of a developmental shift from similarity-based to category-based induction occurring between 9 and 11 years of age.

This developmental shift may be a function of the

development of a categorical structure: when two objects share a label they are more likely to be considered members of the same category than to be considered members of different categories. When a categorical structure is in place, the probability that two remotely similar entities that have the same label would be considered members of different categories could be estimated by the base rate of homonyms (and homophones), and therefore is negligibly small. In fact, we drew a random sample of 200 most frequently used English nouns from Francis and Kucera (1982) and asked three native English speakers to mark those that have homonyms and homophones. While the overall rate of homonyms and homophones appeared to be relatively high (ranging from 20% to 30%), many of these homonyms and homophones were adjectives and verbs (e.g., horse/hoarse or board/bored). At the same time, the rate of noun-noun homonyms (e.g., case/case) was around 5%. Furthermore, the rate of a noun having a homonym within the same ontological class (e.g., living creature having a homonym that indicates a completely different living creature) was practically nonexistent. Hence, remotely similar entities that share the label should be interpreted as members of the same category and, therefore, to share unobservable properties as well. In short, the label-as-attribute model proposed by Sloutsky & Lo (1999) can account not only for similarity judgment of younger children, but also for their specific induction.

While the model provides a reasonable account of specific induction with artificial stimuli that are relatively similar on the overall scale, it remains unclear whether or not the model is capable of handling more naturalistic and diverse set of stimuli. Our most immediate concern is to test the model with these kinds of stimuli. Because our stimuli were quite similar overall (all pictures represent human-like faces) it is possible that results might have been different had the stimuli been more different. It is also possible that results might have been different if stimuli were not human-like: infants and young children have been shown to develop different types of representations of humans and animals (Quinn & Eimas, 1998). While the former could be represented as individual exemplars, the latter may have summary (i.e., category-based yet perceptual) representations. However, we believe that introduction of more different and more diverse stimuli would make differences between younger and older children even more apparent, because stimuli would increase perceptual-similarity-based variance in the younger groups without increasing this variance in the older group.

If induction in young children is based on overall similarity among compared entities then introduction of new attributes (both perceptual and non-perceptual) that contribute to overall similarity, should also contribute to inductive generalizations. It would be important to test this prediction and to estimate weights of different classes of attributes. It would be also important to trace changes in these weights with development and learning. Finally, it would be necessary to test the model on younger participants and have more dense developmental observations.



Because the proposed model is capable of formulating specific predictions, these predictions can be tested in future research. For example, we contend that specific induction in young children is similarity-based, whereas preadolescents it is category-based. If this is true, then for younger children specific induction should be easier than general induction, while for older children it should be more difficult (because category-based specific induction requires more mental steps than general induction). However, if specific induction in younger children is also category-based, then in both younger and older children specific induction should be more difficult than general induction.

Recall that the label-as-attribute model also affords the computation of specific probabilities of inductive generalizations across stimuli that vary in overall similarity. In future research, we plan to test these predictions of the model with respect to naturalistic stimuli. Because it is impossible to individuate features and to precisely calculate featural overlap with complex naturalistic stimuli patterns, we will manipulate similarity by "morphing" naturalistic pictures into each other in a fixed number of steps.

### Acknowledgements

This research has been supported by a grant from the James S. McDonnell Foundation to the first author.

### References

- Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.
- Fisher, R. (1935). *The design of experiments*. Oliver & Boyd, London: UK.
- Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston, MA: Houghton Mifflin.
- Gelman, S. A., & Coley, J. (1991). Language and categorization: The acquisition of natural kind terms. In S. A. Gelman, S. & J. P. Byrnes (Eds.), *Perspectives on language and thought: Interrelations in development* (146-196). New York, NY: Cambridge University Press.
- Gelman, S. A., & Markman, E. (1986). Categories and induction in young children. *Cognition*, 23, 183-209.
- Gelman, S. A., Coley, J. D. & Gottfried, G. M. (1994). Essentialist beliefs in children: The acquisition of concepts and theories. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 234-254). New York, NY: Cambridge University Press.
- Goodman, N. (1992/1972). Seven strictures on similarity. In M. Douglas & D. Hull (Eds.), *How classification works* (pp. 13-23). Edinburgh, Scotland: Edinburgh University Press.
- Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.
- Mandler, J. M., & McDonough, L. (1998). Studies in inductive inference in infancy. *Cognitive Psychology*, 37, 60-96.
- Medin, D. (1975). A theory of context in discrimination learning. In G. Bower (Ed.), *The psychology of learning and motivation* (pp. 263-314), Vol. 9. New York: Academic Press.
- Quinn, P., & Eimas, P. (1998). Evidence for global categorical representation of humans by young infants. *Journal of Experimental Child Psychology*, 69, 151-174.
- Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, 237, 1317-1323.
- Sloutsky, V. M., & Lo, Y.-F. (1999). How much does a shared name make things similar? Part 1: Linguistic labels and the development of similarity judgement. *Developmental Psychology*, 6, 1478-1492.

# Problem Representation in Experts and Novices: Part 2. Underlying Processing Mechanisms

**Vladimir M. Sloutsky (sloutsky.1@osu.edu)**

School of Teaching & Learning and Center for Cognitive Science, The Ohio State University  
21 Page Hall, 1810 College Road, Columbus, OH 43210 USA

**Aaron S. Yarlas (yarlas.1@osu.edu)**

Center for Cognitive Science and School of Teaching & Learning, The Ohio State University  
21 Page Hall, 1810 College Road, Columbus, OH 43210 USA

## Abstract

It has been well established that experts and novices focus on different aspects of problems, with novices focusing more on surface features rather than on deep principled features of a problem. What is less clear are the mechanisms that underlie these differences in construal of problem representation. The current study, which uses an 'old/new' recognition procedure, examines expert and novice representation of arithmetic equations in which the deep relational properties (i.e., principles of commutativity and associativity) were well known to both groups. Results indicate that both novices and experts encode both surface and principled features in the same serial manner, with surface features preceding principled features for both. At the same time, only for novices and not for experts, surface features compete with deep features, thus requiring additional resources to inhibit this attentional competition.

## Introduction

Mental representation is a central component of several fundamental cognitive processes, including categorization, reasoning, decision making, and problem solving. For example, the way an entity is categorized depends on the content of an organism's mental representation regarding this entity and the similarity of this representation to a composite representation stored in memory (Estes, 1994; Nosofsky, 1988). In addition, the way people reason from propositions and what they infer from these propositions depends on the manner in which these propositions are mentally represented (Byrne, 1989; Johnson-Laird & Byrne, 1991; Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999). Finally, the content of a mental representation determines the approaches and strategies people use when they attempt to solve problems (Kaplan & Simon, 1990; Larkin & Simon, 1987; Newell & Simon, 1972). Of course, the content of mental representation may depend on knowledge of the conceptual and relational structure of the domain, and transformational procedures and algorithms (Anderson, 1982; 1990; Case & Okamoto, 1996; Gelman & Meck, 1986, 1992; Hebert & Lefevre, 1986; Rittle-Johnson & Alibali, 1999). For example, the problem "Bill has eight marbles and Jill has six times more" would be represented as " $8 \times 6 = ?$ ", only if

the person has knowledge of and can abstract basic multiplication algorithms.

As noted above, there is a distinction between the content of a mental representation (or what is represented) and the process of construing this content (or what is attended to, encoded, and stored). The process of construing mental representations remains largely unknown, and is the focus of this paper. However, there are several important regularities that have been established with respect to the content of mental representation that are important for the study of the process of construing of mental representation.

In Part 1 of this paper (Yarlas & Sloutsky, 2000) and elsewhere (Yarlas & Sloutsky, 1999), we describe a large body of literature indicating that in problem solving, reasoning, learning and transfer, and problem categorization, novices and experts construe representations that differ in their content. In particular, novices tend to focus on surface features of the problem, whereas experts tend to focus on deep relational features (e.g., Chase & Simon, 1973; Chi, Feltovich, & Glaser, 1981; Gentner & Toupin, 1986; Kotovsky & Gentner, 1996; Larkin, 1983; Simon & Simon, 1978; Yarlas & Sloutsky, 1999). These effects have been demonstrated across a variety of knowledge-rich and knowledge-lean domains.

However, in spite of these well-established expert-novice differences, it remains unclear what accounts for these differences. Do differences occur because experts have knowledge of deep relational properties and novices do not? Do they occur because novices are less intelligent or younger than experts are, and they cannot grasp deep relational properties? Do experts and novices differ in processes underlying the construal of a problem representation? Or do differences stem from a combination of these factors?

In Part 1 of this paper (Yarlas & Sloutsky, 2000), we focused on expert-novice differences in the content of mental representations. It was demonstrated that when tasks are sufficiently simple and deep relational properties are well known, neither differences in knowledge, intelligence, nor development can fully account for the observed differences between novices and experts. In a series of experiments designed to distinguish among these possibilities, tasks were constructed that included principles of arithmetic familiar to novices, and surface features that were completely superfluous with respect to deep relational features. In particular,

they asked participants varying in age and degree of expertise to sort mathematical equations that could have common surface elements (e.g., commonality of numbers or the same number of constituent addends in the equation) or common deep mathematical principles (e.g., commutativity or associativity). Results indicated that only mathematics experts consistently focused on principles, whereas novices, regardless of age and intelligence, focused mostly on surface features. However, elimination of surface features led to substantial increase in focusing on principles. Interestingly, the reintroduction of surface features reduced participants' focus on principles to their original low levels. These and other manipulations allowed us to argue that differences between novices and experts stem from differences in processes underlying the construal of a problem representation. However, if novices have knowledge of the principles in question yet still fail to represent them, then several questions arise about processes underlying problem representations in novices and experts. Do novices initially encode both deep and surface features, but later discard the deep relational properties, or do they simply fail to encode the deep relational properties? And what are the processing mechanisms underlying problem representations in experts: do experts encode and discard surface features, or do they ignore these features from the very beginning?

To answer these questions, we used an 'old/new' recognition paradigm in the current experiment. This paradigm affords the creation of a set of foils, such that patterns of hits and false alarms point to which aspects of problems have been encoded and committed to memory and which aspects have been left out. In the study phase, participants were presented with a set of arithmetic equations. These equations all utilized a principled property, either associativity or commutativity. The former states that for addition, subtraction, and multiplication, constituent parts can be decomposed and recombined in different ways (e.g.,  $a + b = [a - c + c] + b$ ). The latter states that the order of elements is irrelevant for addition and multiplication (e.g.,  $a + b + c = b + c + a$ ). In addition, these equations all used consistent levels of two surface elements: all equations used numbers ranging between 1 and 9, and all used either 5 or 6 numbers in the equation. In the recognition phase of the experiment, in addition to 'old' items, four combinations of 'new' equations were presented as foils. Half of these foils, which we refer to as 'feature +' foils, maintained the same levels of surface features as used in the learning phase (i.e., numbers ranging between 1 and 9, and either 5 or 6 numbers in the equation), while the other half of the foils, which we refer to as 'feature -' foils, violated these categories (i.e., numbers greater than 9, and either 4 or 7 numbers in the equation). Also, half of the foils, which we refer to as 'principle +' foils, maintained the use of one of the two principled properties, while the other half, which we refer to as 'principle -' foils, did not use any principled properties in the equation. The two levels of the two kinds of properties (feature being either + or -, and principles being either + or -) were fully-crossed, thus creating four combinations of foils: feature + /principle + (F+/P+), feature + /principle - (F+/P-), feature -/principle + (F-/P+), and feature -/principle - (F-/P-). For example, for the equation  $5 + 3 + 6 = 3 + 6 + 5$  in the study phase, the

following foils were presented in the recognition phase: (1)  $5 + 3 + 6 = 3 + 6 + 5$  (Old), (2)  $7 + 4 + 2 = 4 + 2 + 7$  (F+/P+), (3)  $5 + 3 + 6 = 3 + 4 + 7$  (F+/P-), (4)  $11 + 9 = 9 + 11$  (F-/P+), and (5)  $14 + 7 = 9 + 12$  (F-/P-).

The goal of this paper is to elucidate processes underlying problem representations in novices and experts. In this article, we consider and test a number of possible processing models for both novices and experts, which are summarized in Table 1.

Table 1: Summary of considered processing models

Novice Model 1	Encode only surface features with no encoding of deep structural features
Novice Model 2	Encode both surface and deep structural features; attentional competition between surface and structural features, with surface features winning
Expert Model 1	Encode only deep structural features with no encoding of surface features
Expert Model 2	Encode both deep structural and surface features; attentional competition between structural and surface features, with structural features winning
Expert Model 3	Encode both deep structural and surface features; no attentional competition

For this task, if novices encode only surface features and not relational features, they should rapidly respond "Old" when surface features are present and they should rapidly respond "New" when surface features are absent (Novice model 1). Similarly, if experts encode only principles and not surface features, they should rapidly respond "Old" when principles are present and they should rapidly respond "New" when principles are absent. If either group encodes both principles and features, they should exhibit more complex patterns of responses (Expert model 1).

There is preliminary evidence (Yarlas & Sloutsky, 1999) that novices do encode both surface and deep features, but discard the latter in the course of attentional competition (Novice model 2). However, while processing mechanisms underlying problem representations in novices require further clarifications, these mechanisms in experts remain unclear. One possibility is that experts start construing problem representations from deep rather than from surface (Expert model 2). An alternative possibility is that experts construe representations in a manner similar to that of novices, except that there is no attentional competition in experts (Expert model 3). Of course, it is also possible that experts construe representations in a parallel manner, in which case their response latencies should exhibit small or no differences across the foils.

The alternative response patterns derived from the models summarized in Table 1 are presented in Table 2. These predictions are based on the following two assumptions: (1) both experts and novices process properties of problems in a serial manner and (2) each additional step in processing leads to increase in latencies. Both assumptions were previously corroborated using this task with novices (Yarlas &

Table 2: Patterns of responses and latencies predicted by alternative models for novices and experts

Models of responses	Foil Types and Patterns of Responses				
	Old targets	F+/P+	F+/P-	F-/P+	F-/P-
Novices Model 1 (Response type)	OLD	OLD	OLD	NEW	NEW
Novices Model 1 (Latency)	Fast	Fast	Fast	Fast	Fast
Novices Model 2 (Response type)	OLD	OLD	NEW	NEW	NEW
Novices Model 2 (Latency)	Slow	Slow	Very Slow	Fast	Fast
Experts Model 1 (Response type)	OLD	OLD	NEW	OLD	NEW
Experts Model 1 (Latency)	Fast	Fast	Fast	Fast	Fast
Experts Model 2 (Response type)	OLD	OLD	NEW	NEW	NEW
Experts Model 2 (Latency)	Slow	Slow	Fast	Slow	Fast
Experts Model 3 (Response type)	OLD	OLD	NEW	NEW	NEW
Experts Model 3 (Latency)	Slow	Slow	Slow	Fast	Fast

Sloutsky, 1999). Because of these assumptions, the parallel processing model is absent from Table 1; however we do not discount the possibility of parallel processing in experts. Note that predictions presented in Table 2 are qualitative, in that they do not specify accuracy or latency across the conditions, but rather point to (a) patterns of recognition responses and (b) directions of differences in latencies.

Note that the tables have two critical components. First, in novices, responses to F+/P- foils afford either corroboration or elimination of Model 1 for novices (see Table 1), whereas in experts, responses to F-/P+ foils afford corroboration or elimination of Model 1 for experts (see Table 1). Second, within experts and novices, patterns of differences in latencies afford the selection of the more plausible model as well as the description of specific processing components. Specifically, latencies in experts' responses to F+/P- items will allow for discriminating between Model 2 and Model 3 for experts. In short, patterns presented in the table should allow us to distinguish between processing models in novices and experts presented in Table 1.

## Method

### Participants

Two samples, representing novices and experts, were used in this study. The novice group included twenty-three undergraduates in an introductory psychology course at the Ohio State University who participated for partial course credit. This sample had an average age of 19.2 years ( $SD = 0.9$  years), with 12 women and 11 men. The expert group included twelve graduate students in the Mathematics Department at the same university who participated for a payment of twenty dollars. This sample had an average age of 27.6 years ( $SD = 5.8$  years), with 3 women and 9 men.

### Materials and Procedure

The materials and procedures used in this study were identical for participants in both the novice and expert samples. All participants were run individually with stimuli presented by a personal computer using SuperLab software (Cedrus Corporation, 1999).

The experiment consisted of three phases: the study phase, the distraction phase, and the recognition phase. In the study

phase, participants were presented with thirty arithmetic equations, which they had been instructed to memorize. All thirty equations used addition, used numbers ranging from 1 to 9, contained either 5 or 6 numbers, and used either the associative or commutative principle (half for each). Each equation was centered and presented in dark type on a white screen for ten seconds, with a two-second interval between each, during which only the white background was seen. The order of equations was randomized across participants.

A distraction phase followed the study phase for the purpose of clearing participants' short-term memory. For the distraction task, participants were presented with ninety letters, for which they had been instructed to indicate whether the letter was a vowel or a consonant. This phase took approximately three minutes.

Following the distraction phase was the recognition phase. Participants were told that they would be presented with a number of arithmetic equations, some of which had been presented to them earlier and some of which had not been presented earlier, and that they were to decide whether each equation was 'old' or 'new'. There were a total of sixty equations presented in the recognition phase. The order of equations presented in this phase was randomized across participants. There were five categories of foils, with twelve exemplars for each category. Recall that these foils included: (1) Old targets that had been presented earlier in the learning phase, (2) F+/P+ equations, which used similar surface features and used either the commutativity or associativity principle as in the original equations, (3) F+/P- equations, which used similar surface features as the original equations but did not use either the commutativity or associativity principle, (4) F-/P+ equations, which used surface features different from those used in the original equations but used either the commutativity or associativity principle, and (4) F-/P- equations, which used surface features different from those used in the original equations and did not use either the commutativity or associativity principle.

## Results and Discussion

In this section, we will first discuss the accuracy of recognition and latencies of responses for novices, and then for experts. For each group, we will first examine overall accu-

curacy of response to the foils (i.e., correct acceptance of Old targets and correct rejection of all foils). We will then compare participants' "Old" responses and latencies across the foil types. Note that for all foils except F+/P+, we compared latencies for correct responses only. Because we expected a large number of false alarms for F+/P+ foils, for these foils, latencies for both correct and incorrect responses were used in the analyses.

Novices exhibited high overall accuracy for most of the foils, correctly accepting Old targets and correctly rejecting F-/P+, F-/P-, and F+/P- foils. They mostly false alarmed, however, on F+/P+ foils. The latter finding is expected because F+/P+ foils were categorically indistinguishable from Old targets, since both surface features and principled features present in Old targets were also present in F+/P+ foils. More specifically, results indicate that accuracy rates (i.e., hits for Old Targets and correct rejections for the other foils) for F+/P- ( $M = 0.69, SD = 0.35$ ), F-/P- ( $M = 0.93, SD = 0.20$ ), F-/P+ ( $M = 0.97, SD = 0.16$ ), and Old targets ( $M = 0.84, SD = 0.15$ ) were significantly higher than chance (all  $t(22) > 9.4, ps < .001$ ), whereas for F+/P+ ( $M = 0.36, SD = 0.26$ ) accuracy was significantly lower than chance,  $t(22) = -6.4, p < .001$ . These results indicate that these participants took the task seriously and were providing rather accurate responses.

Percentages of "Old" responses and latencies for novices are presented in Figure 1. A one-way repeated measures ANOVA points to significant differences among foils for novices ( $F(4, 88) = 53.9, MSE = 542.7, p < .0001$ ). Paired-samples t-tests indicated the following the following direction in the proportion of "Old" responses: Old targets  $>$  F+/P+  $>$  F+/P-  $>$  F-/P+ = F-/P-, all  $t(22) > 3$ , all Bonferroni adjusted  $ps < .05$  for differences.

Novices' latencies to different foils are also presented in Figure 1. These measures were also subjected to a one-way repeated measures ANOVA. The analysis indicates significant differences among the foils,  $F(4, 76) = 15.48, p < .001$ . Planned comparisons revealed that F+/P- latencies were significantly higher than those for Old targets,  $t(20) = 3.4, p < .005$ , whereas latencies of F-/P- and F-/P+ foils were significantly lower than those of the Old targets,  $t(21) > 3.5, ps < .005$ .

These data allow us to rule out Model 1 presented in Table 1 -- novices did not base their responses solely on the presence or absence of surface features. When surface features were absent (F-/P- and F-/P+ foils) participants produced fast and accurate "New" responses; however, when surface features were present, novices did not always produce "Old" answers. Rather, novices' responses were mediated by the presence or absence of principled features. In particular, when both surface and principled features were present (Old targets and F+/P+ foils) novices generally responded "Old". These responses were slower than those for F-/P- and F-/P+ foils. Finally, when surface features were present but principles were absent (F+/P- foils), participants in general accurately rejected these foils, but latencies for these correct rejections were significantly higher than latencies for Old targets. These findings support the notion of the attentional competition between the two types of features (see Table 1, Novice model 2), pointing to a relative difficulty for partici-

pants to inhibit the salient surface feature and reject the foil. Of course, these data raise an interesting question of whether or not experts would also exhibit attentional competition between deep relational and surface features.

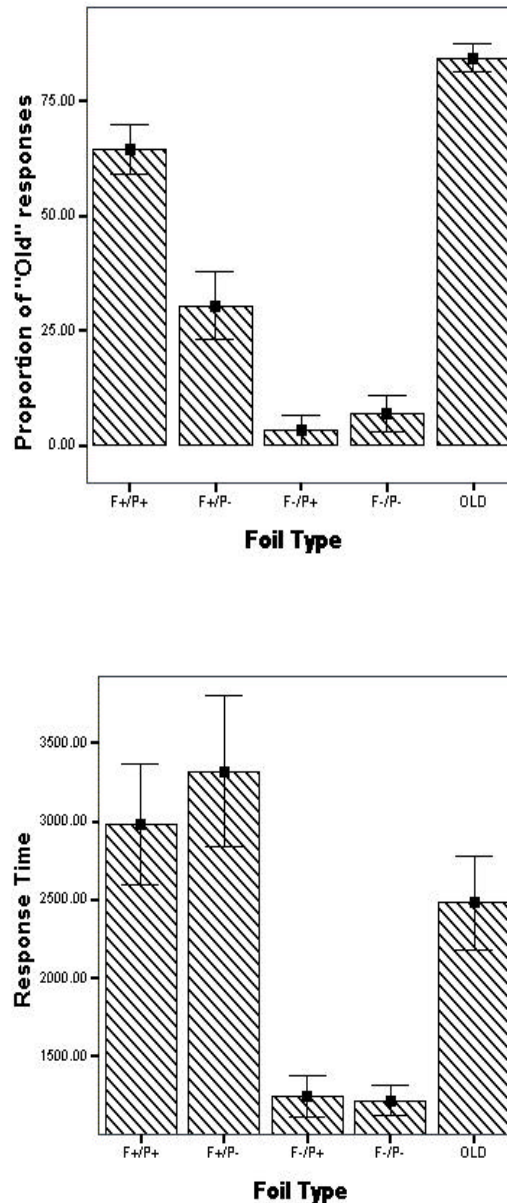
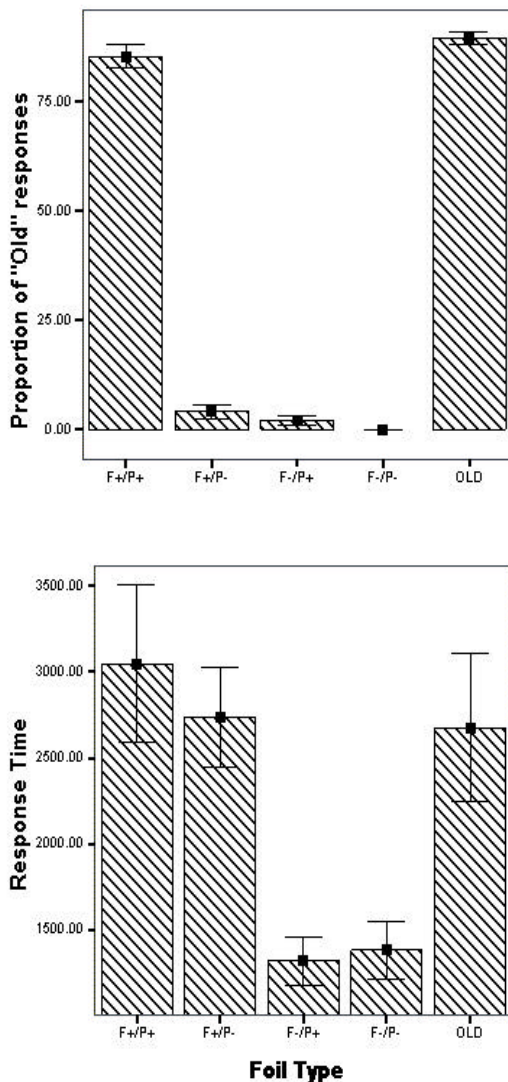


Figure 1. Proportion of novices' "Old" responses and response times (in milliseconds) across foil types in the recognition phase.

Similarly to novices, experts exhibited high overall accuracy for most of the foils, correctly accepting Old targets and correctly rejecting F-/P+, F-/P-, and F+/P- foils. They too mostly false alarmed, however, on F+/P+ foils. More specifically, accuracy rates (i.e., hits for Old Targets and correct rejections for the other foils) for F+/P- ( $M = 0.96$ ,

$SD = 0.06$ ), F-/P- ( $M = 1.00$ ,  $SD = 0.00$ ), F-/P+ ( $M = 0.98$ ,  $SD = 0.04$ ), and Old targets ( $M = 0.90$ ,  $SD = 0.05$ ) were significantly higher than chance (all  $t_s(11) > 58$ ,  $p_s < .001$ ), whereas for F+/P+ ( $M = 0.15$ ,  $SD = 0.09$ ) accuracy was significantly lower than chance,  $t(11) = -5.1$ ,  $p < .001$ . These results indicate that experts also took the task seriously and provided rather accurate responses.

Percentages of "Old" responses and latencies for experts are presented in Figure 2. A one-way repeated measures ANOVA points to significant differences among foils for experts ( $F(4, 44) = 768.5$ ,  $MSE = 34.2$ ,  $p < .0001$ ). Paired-samples t-tests indicated the following the following direction in the proportion of "Old" responses: Old targets = F+/P+ > F+/P- = F-/P+ = F-/P-, all  $t_s(22) > 23$ , all Bonfer-



oni adjusted  $p_s < .0001$  for differences.

Figure 2. Proportion of experts' "Old" responses and response times (in milliseconds) across foil types in the recognition phase.

Experts' latencies to different foils are also presented in Figure 2. These measures were also subjected to a one-way repeated measures ANOVA. The analysis indicates significant differences among the foils,  $F(4, 44) = 18.60$ ,  $p < .001$ . Planned comparison revealed that, in contrast to novices, F+/P- latencies for experts were not significantly different from those for Old targets,  $t(11) = 0.2$ ,  $p = .85$ , but that latencies for F-/P- and F-/P+ foils were again significantly lower than those of the Old targets,  $t_s(11) > 4$ ,  $p_s < .005$ .

The analysis of hits and false alarms allows us to eliminate Model 1 of expert responses presented in Table 1. Indeed, according to this model, experts should have responded "New" when principles were absent, and respond "Old" when principles were present. However, the F-/P+ foils almost invariably generated "New" responses, thus eliminating Model 1. Similarly, the analysis of latencies affords the elimination of Model 2. Recall that according to this model, experts should have more rapidly answered "New" when the principle was absent than when the feature was absent. However, the observed findings are consistent with Model 3 and not with Model 2, given that F-P+ foils were rejected faster than F+P- foils. Therefore, results of the experiment support Model 2 for novices and Model 3 for experts.

These findings point to important processing similarities and differences in experts and novices. First, both experts and novices exhibited serial processing. In addition, when construing problem representations, both experts and novices encode features first. At the same time, only novices experience competition between salient surface features and less salient deep principles. For the majority of novices, well known deep principles end up winning the competition; however, the competition takes time and effort. At the same time, experts represent both deep and surface features of the problem and do not experience such attentional competition. Recall that the experiment employed a very simple recognition task. In more resource demanding tasks, such as categorization, reasoning, or problem solving, deep relational features in novices may lose attentional competition to salient surface features. This loss would manifest itself in novices' tendency to focus on surface feature, while ignoring deep relational features (Chase & Simon, 1973; Chi, Feltoich, & Glaser, 1981; Gentner & Toupin, 1986; Kotovsky & Gentner, 1996; Larkin, 1983; Simon & Simon, 1978; Yarlas & Sloutsky, 1999).

The results have several potential implications. First, they lead to a better understanding of expertise, indicating that expert-novice differences persist even with most simple tasks (it is reasonable to expect that more complex tasks would result in more dramatic expert-novice differences). Second, the results have important educational implications, suggesting that salient surface features may deter rather than promote learning.

## Conclusion

The reported findings indicate that even when a task is very simple, experts and novices construct problem representations differently. While both experts and novices encode deep as well as surface features of the problem, only

for novices and not for experts, surface features compete with deep features, thus requiring additional resources to inhibit this attentional competition. These findings may or may not hold for less familiar deep principles or more complicated tasks. However, these results allow us to conclude that even when a task is very simple and deep principles are well known, experts and novices differ in processes underlying the construal of problem representations.

### Acknowledgments

This research has been supported by a grant from James S. McDonnell Foundation to the first author.

### References

- Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, 89, 369-406.
- Anderson, J.R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Byrne, R. M. (1989). Suppressing valid inferences with conditionals. *Cognition*, 31(1), 61-83.
- Case, R., & Okamoto, Y. (1996). The role of central conceptual structures in the development of children's thought.
- Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4, 55-81.
- Chi, M. T. H., Feltovich, P. G., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152.
- Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.
- Gelman, R., & Meck, E. (1986). The notion of principle: The case of counting. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, NJ: Erlbaum.
- Gelman, R., & Meck, E. (1992). Early principles aid initial but not later conceptions of number. In J. Bideaud, C. Meljac & J. Fischer (Eds.), *Pathways to number: Children's developing numerical abilities*. Hillsdale, NJ: Erlbaum.
- Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, 10, 277-300.
- Hiebert, J., & Levevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, NJ: Erlbaum.
- Johnson-Laird, P., & Byrne, R. (1991). *Deduction*. Hove, UK: Lawrence Erlbaum.
- Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, 67, 2797-2822.
- Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J.-P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, 106, 62-88.
- Kaplan, C. A., & Simon, H.A. (1990). In search of insight. *Cognitive Psychology*, 22, 374-419.
- Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, 67, 2797-2822.
- Larkin, J. (1983). The role of problem representation in physics. In D. Gentner & A. Stevens (Eds.), *Mental models*. Hillsdale, NJ: Erlbaum.
- Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11, 65-99.
- Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 14, 54-65.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, N J: Prentice-Hall.
- Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology*, 91(1), 175-189.
- Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. In R. S. Siegler, (Ed), *Children's thinking: What develops?* Hillsdale, NJ: Lawrence Erlbaum Associates.
- Yarlas, A. S., & Sloutsky, V. M. (1999). Representation of arithmetic principles by novices: Knowledge or attention? Manuscript under review.
- Yarlas, A. S., & Sloutsky, V. M. (2000). Problem representation in experts and novices: Part 1. Differences in the content of representation. *Proceedings of the 22<sup>nd</sup> Annual Conference of the Cognitive Science Society*.

## Prosodic Choice: Effects of Speaker Awareness and Referential Context

**Jesse Snedeker** (jessned@psych.upenn.edu),  
**Lila Gleitman** (gleitman@psych.upenn.edu),  
**Michaal Felberbaum** (felberbm@sas.upenn.edu),  
**Nicora Placa** (nicorap@sas.upenn.edu),  
**John Trueswell** (trueswel@psych.upenn.edu)

Institute for Research in Cognitive Science; 3401 Walnut Street, Suite 400A  
 Philadelphia, PA 19104 USA

### Abstract

These experiments were designed to discover whether untrained speakers produce prosodic cues that are sufficient to allow listeners to interpret ambiguous PP-attachments. A referential communication task was used to elicit productions of ambiguous sentences and determine whether listeners could use prosodic cues to correctly interpret these ambiguities in context. In Experiment 1, the referential context supported both potential interpretations of the ambiguity. Acoustic analyses indicated that Speakers produced potentially informative prosodic cues. Listeners' responses to the ambiguous sentences strongly reflected the demonstration the Speaker had seen, indicating that they were able to use this information. However, post-experiment interviews revealed that Speakers were aware of the ambiguous situations. Experiment 2 manipulated Speaker awareness by altering the Speaker's referential context to support only the intended meaning, and by making the resolution of the ambiguity a between subjects variable. Although Listeners' contexts were unchanged from Experiment 1, Listeners now showed no sensitivity to the Speakers' intended meaning. Acoustic analysis indicated that the strong prosodic cues provided in Experiment 1 were absent in Experiment 2. The experiments suggest that informative prosodic cues depend upon speakers' knowledge of the situation: speakers provide prosodic cues when needed; listeners use these prosodic cues when present.

### Introduction

One of the current challenges for research on prosody and syntactic ambiguity is to bring together what we know about the listener with what we know about the speaker. In doing so, we can begin to understand whether the prosodic cues that are available in speech can influence a listener's interpretation. The research reported here attempts to address this challenge by examining how a speaker uses prosody in the face of ambiguity and whether an accompanying listener is able to interpret the speaker's intended meaning. We will propose from this research that prosodic cues in adult-to-adult speech often depend upon the speakers' knowledge of the referential context. In particular, the choice to provide helpful prosodic cues depends upon whether or not the referential situation furnishes other cues that could help resolve the ambiguity.

Prior research on prosody and syntactic ambiguity has focused almost exclusively on either the speaker or the listener, and only rarely on the interaction between the two.

This division of labor has led to important advances in our understanding of prosody; we know a fair amount about what listeners can do with prosodic cues, and what prosodic cues speakers can produce.

Numerous language comprehension studies have demonstrated that prosodic manipulations of the linguistic input can influence comprehenders' on-line and off-line decisions about syntactic ambiguity (for reviews see, Warren, 1996; Kjelgaard & Speer, 1999). These studies have used a wide variety of experimental techniques, including cross-modal naming, lexical decision, word monitoring, and sentence judgments, and have found effects of prosody on the interpretation of a variety of temporary and global ambiguities. Likewise, studies of language production have found that the prosody of an utterance often reflects its syntactic structure (Cooper & Paccia-Cooper, 1980). Moreover, informed speakers can mark different meanings of an ambiguous string through prosodic grouping (Lehiste, 1976; Allbritton et al., 1996). These studies suggest that speakers of a language share knowledge about prosodic cues to syntax, and can use this information in decisions about production.

Curiously, most comprehension and production studies have relied upon distorted and/or artificial manipulations of prosodic information. In comprehension studies prosody is typically manipulated by splicing silent pauses into speech to indicate clause boundaries, manipulating synthesized speech, or asking trained speakers to produce particular prosodic variants of an utterance. Production studies have relied upon data from trained speakers, such as radio announcers, who have been explicitly instructed to contrast the alternate interpretations of an ambiguous sentence. Notably, few studies of prosody and syntax have examined how untrained listeners respond to the speech of untrained speakers in contexts in which the participants are attempting to communicate about a shared situation.

In naturally occurring speech, syntactic structure is only a weak predictor of prosodic variation (for review see Fernald & McRoberts, 1996) This is because prosodic patterns are affected by many other factors, including the length and stress pattern of words, speech rate and discourse factors such as contrastive stress (Selkirk, 1984). Unsurprisingly, a number of researchers have found that naïve speakers produce less consistent prosodic cues for syntactic disambiguation than the informed speakers typically used in comprehension experiments (Lehiste, 1973; Wales & Toner, 1979;



Allbritton, McKoon, & Ratcliffe, 1996). In the most relevant of these studies, Allbritton et al. (1996) compared situations in which speakers were uninformed or explicitly informed about potential ambiguities. In the uninformed case, untrained speakers (undergraduate students) and professional speakers (radio announcers) were asked to read paragraphs containing globally ambiguous sentences (e.g., “They rose early in May”) which had been disambiguated by the prior context. In the informed case, radio announcers were provided with the same globally ambiguous sentences without a disambiguating context, both of the meanings were explained to the speaker, and he/she was asked to read the sentence twice, to convey each of these meanings.

Recordings of these utterances were played for a separate group of subjects who were given both meanings and asked to identify the one that the speaker was attempting to convey. The findings from this judgment task revealed, in the words of the authors, that “most speakers trained or not, did not produce prosodically disambiguated utterances for most sentences. Trained, professional speakers reliably produced appropriate disambiguating prosody only when they were shown the two meanings of the sentence side by side and were explicitly asked to pronounce the sentence twice.”

There are three plausible explanations of the Allbritton et al. findings, each of which has different implications for the role of prosody in syntactic ambiguity resolution. First, it is possible, as the authors claim, that speakers only produce reliable cues when instructed to do so. However, this would suggest that prosodic cues to structure are rare in natural speech, raising questions about how listeners become sensitive to these cues. Second, the results could be interpreted as evidence that speakers only produce reliable cues when the surrounding context does not disambiguate the utterance. In the Allbritton et al. study, experimental naïveté and contextual constraint were confounded. Perhaps, as Lieberman (1967) suggested, speakers don’t bother to divide up an utterance into informative prosodic chunks if other cues are present to disambiguate structure. Finally, it is possible that speakers do not produce reliable prosodic cues when reading connected text, regardless of whether that text provides a disambiguating context.

Recently, Schafer, Speer, Warren & White (1999) have presented data which challenges the Allbritton findings. They elicited prosodic variants of temporary and global ambiguities from uninstructed subjects by having them play a game that used a set of scripted commands. These utterances were submitted to acoustic and phonological analyses and a judgement task parallel to that conducted by Allbritton et al. (1996). In all three analyses Schafer and colleagues found evidence that speakers produced consistent prosodic cues to the intended structure. They attribute the divergent findings to differences in the tasks that were used, suggesting that the subjects in the earlier study were reading and had no clear communicative intentions.

The current paper attempts to explore the role of prosodic cues in language production and comprehension. In particular, we examine the situations under which untrained speakers can produce reliable prosodic cues that will allow listeners to resolve attachment ambiguities. The critical sentences are ones that contain globally ambiguous preposi-

tional phrase attachments, such as “Tap the frog with the flower”. Out of context, the phrase “with the flower” can be taken as Instrument (VP-Attachment) indicating what to use for the tapping, or the phrase can be taken as a Modifier (NP-attachment) indicating which frog to tap.

These experiments were conducted using a referential communication task, in which a Speaker and a Listener were separated by a divider, allowing for only verbal communication between the two participants. Under discussion in these studies was the movement of objects, with Speaker attempting to have the Listener perform actions upon an identical set of objects on the other side of the screen. This situation provided two advantages of over other common tasks. First, the referential context was highly salient, and was defined by the set of objects in front of the speaker and listener. Memory considerations for referential factors (e.g., what a speaker remembers about a story) are not relevant in such a task since the reference world is co-present with the production task. Second, the separation of the Listener and Speaker allowed us to manipulate independently the referential context of the Speaker and the Listener, allowing us to disentangle referential effects on the task of production and the task of comprehension.

In Experiment 1, we examined the use of prosodic cues when the referential context of the Speaker supported either meaning of the target sentence. In Experiment 2, we examined prosodic cues when the referential context of the Speaker strongly supported the intended meaning of the utterance. If prosodic choice is affected by Speaker’s knowledge of the referential context, we would expect to see decreased use of helpful cues when the referential context provides other cues to disambiguate the utterance. If on the other hand, knowledge of the referential context is not relevant, we would expect similar performance across the two experiments.

## Experiment 1

### Methods

**Participants** Thirty-two pairs of participants from the University of Pennsylvania community volunteered for the experiment. They received extra course credit or were paid for their participation. In each pair, one participant played the role of Speaker and the other played the role of Listener. All Speakers were female whereas half the Listeners were male and half were female. All participants were native speakers of English.

**Procedure** During the experiment, the Speaker and Listener sat on opposite sides of a vertical screen. On each trial they were given identical bags containing toys, which they laid out on the trays in front of them. As the Speaker and Listener removed toys from their bags, the Experimenter introduced each toy using indefinite noun phrases (e.g., This bag contains a dog, a fan...).

Next, the Experimenter showed the Speaker a demonstration of the target action. This action could not be seen by the Listener. The Speaker then received a card containing a

written sentence describing this action. Speakers memorized the sentence and returned the card to the Experimenter. After seeing a second demonstration, the Speaker produced the sentence. The Listener responded by attempting to perform the correct action with his or her own set of toys. Speakers were told that the primary goal of the experiment was to say each sentence in such a way as to get the Listener to perform the same action on the other side of the screen. Each Listener was told that her job was to perform the action that she believed had been demonstrated to the Speaker.

During the course of the experiment, interaction between the Speaker and the Listener was limited. Once the Speaker produced the sentence, the Listener could not ask for any clarification. Listeners' actions were videotaped and the Speakers' utterances were audiotaped. After the study was completed the Listener and Speaker were separated and each was interviewed to assess their awareness of the experimental manipulation and the ambiguity in the critical items.

**Stimuli** On critical trials, the target sentence contained an ambiguous Prepositional Phrase attachment, as in (1a) and (1b) below. Identical bags of objects were given to both participants. On each trial the bag contained: 1) a Target Instrument, a full scale object that could be used to carry out the action (e.g., a large flower); 2) a Marked Animal, a stuffed animal carrying a small replica of the instrument (e.g., a frog holding a little flower); 3) an Unmarked Animal (e.g., an empty-handed frog); and 4) two unrelated objects (e.g., a giraffe in pajamas and a lego block). The set of toys supported both interpretations of the ambiguous sentence by providing a potential direct object (plain frog) and instrument (large flower) for the VP-attachment and a potential direct object for the NP-attachment (frog holding flower).

The Experimenter demonstrated one of two possible actions: an Instrument action (e.g., the Experimenter picked up the large flower and tapped the plain frog) or a Modifier action (e.g., using her hand, the Experimenter tapped the frog that had the small flower). Ambiguous sentences were compared with unambiguous sentences (1c and 1d).

1a. Tap the frog with the flower. (Amb, Inst)  
*Action involves the unmarked frog and the instrument.*

1b. Tap the frog with the flower. (Amb, Mod)  
*Action involves the marked frog and not the instrument.*

1c. Tap the frog by using the flower. (Unamb, Inst)  
*Action involves the unmarked frog and the instrument.*

1d. Tap the frog that has the flower. (Unamb, Mod)  
*Action involves the marked frog and not the instrument.*

Four presentation lists were constructed so that each of the 16 target trials appeared in only one of the four possible conditions on a given list but appeared in each of the conditions across lists (resulting in four target trials in each condition per subject pair). The target trials were interspersed with thirty distractor trials. Four additional lists were generated by reversing the order of trials in each list.

**Coding** The videotapes of Listeners' actions were edited to include only the actions on the sixteen target trials, and all audio was removed. Coders, who were blind to the condition of each trial, judged whether the Listener made an Instrument response (performed the target action using the Target Instrument or the miniature instrument).

## Results

**Listener's Actions** The percent of Instrument responses in each of the four conditions is presented in Figure 1. Listeners' actions in response to the ambiguous instructions were affected by the action demonstrated to the Speaker ( $F(1,16) = 63.42, p < .001$ ;  $F(1,12) = 77.31, p < .001$ ). When an Instrument action had been demonstrated to the Speaker, Listeners produced an Instrument action 66% of the time. When a Modifier action had been demonstrated, Listeners produced an Instrument action only 24% of the time.

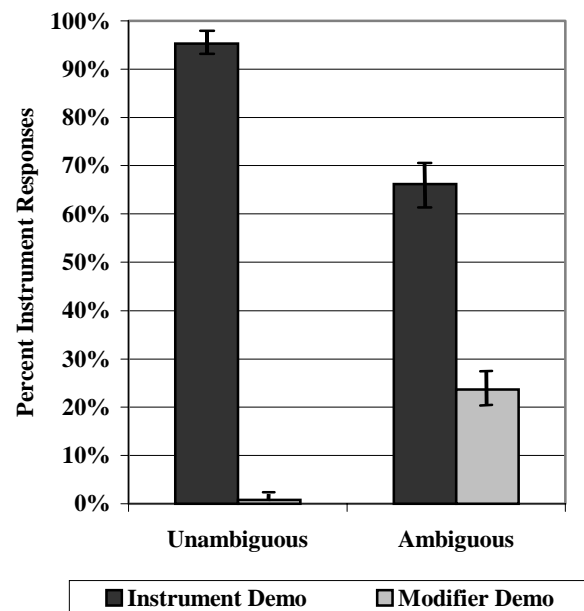


Figure 1: Experiment 1 Listener's Actions

Also, as expected, unambiguous instructions (the left-hand portion of Figure 1) resulted in extremely accurate performance by Listeners. The interaction between Ambiguity and Demonstration Type was reliable ( $F(1,16) = 81.91, p < .001$ ;  $F(1,12) = 113.71, p < .001$ ). As can be seen in the figure, this interaction arose because Listeners were more accurate at reproducing the demonstrated action when the utterance was syntactically Unambiguous than when it was Ambiguous. This pattern suggests that the prosodic cues produced by Speakers were highly informative to Listeners, but not as informative as unambiguous sentences.

**Speaker's Prosody** To verify that our Listeners were glean- ing this information from prosodic cues provided by the Speaker, we conducted acoustic analyses of the ambiguous target sentences. The audio recordings were digitized and a

speech waveform display was generated for each target utterance. Coders, who were blind to the condition, measured the duration of the Verb Composite (verb plus the postverbal pause, if any) and the Noun Composite (the direct object noun plus the following pause, if any). The onset or offset of a word was initially estimated by using visual information from the speech waveform display. This estimate was revised by listening to gated regions of the waveform.

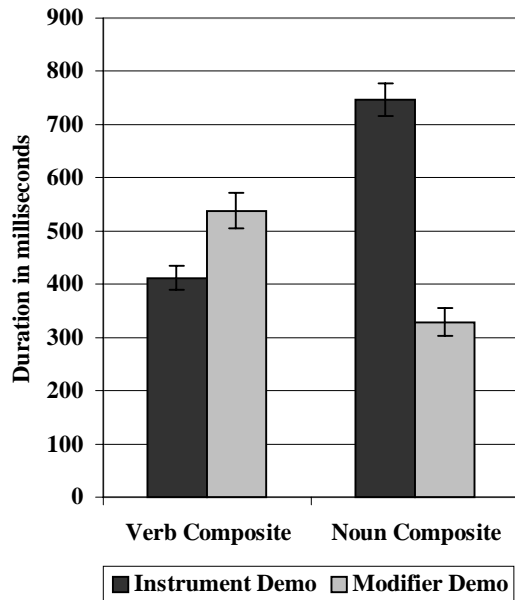


Figure 2: Experiment 1, Mean Durations from Speakers' Utterances

As Figure 2 suggests there is a reliable and substantial effect of demonstration on the mean duration of both the Verb Composite ( $F(1,24) = 12.92, p < .001; F(1,12) = 50.59, p < .001$ ) and Noun Composite ( $F(1,24) = 52.71, p < .001; F(1,12) = 290.42, p < .001$ ). When Speakers saw Instrument Demonstrations, they tended to lengthen the direct object noun and they paused between the noun that the with-phrase on 68% of the trials. This prosodic pattern suggests that the major phrase boundary is located between the direct object and the prepositional phrase and is thus consistent with a verb-phrase attachment of the prepositional phrase (instrument interpretation) but not with a noun-phrase attachment (modifier interpretation). In contrast, when Speakers saw Modifier Demonstrations, they tended to lengthen the verb and paused after the verb 40% of the time. This prosodic pattern suggests that the major phrase boundary is located between the verb and the direct object noun phrase and is more consistent with a noun-phrase attachment.

**Ambiguity Awareness** Listeners' actions in response to ambiguous instructions suggest that prosodic cues were a highly effective but imperfect means of syntactic disambiguation. However the results of the postexperimental interviews raised some concerns about the generality of these findings. 97% of the Speakers in our experiment and 91%

of the Listeners reported being aware of the ambiguity. As mentioned earlier, Allbritton and colleagues (1996) found that ambiguity awareness affected radio announcers' ability to generate useful prosody. Although our participants were not trained radio announcers, we thought it necessary to explore if ambiguity awareness, and more generally knowledge of the referential situation, were influencing the kinds of prosodic choices made by our Speakers.

## Experiment 2

In this experiment, we attempted to decrease Speaker awareness of ambiguity. This was accomplished by making two changes to the previous experiment. First, we altered the Speakers' referential context, so that only the intended meaning of the ambiguous phrase was supported. This was done in hopes that the alternate interpretation would not be considered by the Speakers if it was not suggested by the context itself. Second, we made the type of Demonstration a between subjects variable.

All other aspects of Experiment 2 were the same as Experiment 1. It is especially important to note that the Listeners' context was the same as that used in Experiment 1. And, as in Experiment 1, participants were told in advance that on each trial the Speaker and Listener would receive an identical set of toys. However, in Experiment 2 this was a deception, to be explained at the end of the study.

## Methods

**Participants** Thirty-two pairs of participants from the University of Pennsylvania community received extra course credit or were paid for their participation. All Speakers were female, 17 of the Listeners were male. All participants were native speakers of English and none had participated in Experiment 1. Two additional pairs of subjects participated but were not included in the analyses because of experimenter error (1) or failure to follow instructions (1).

**Procedure** The procedure was the same as Experiment 1 except that the contents of the bags were not listed aloud, to prevent the subjects from discovering that their bags contained different sets of objects. Instead a card listing the objects was included in each bag and the participants were told to check the contents of the bags against the card to insure that all of the toys were present.

**Stimuli** The stimuli and experimental design were the same as in Experiment 1, with the following exceptions. When the Experimenter performed an Instrument Demonstration, the Speaker's bag of toys did not include a Marked animal (e.g., the frog holding the flower) but instead included a second unrelated animal (e.g., an elephant wearing a hat). Hence, a modifier interpretation of the with-phrase should be less available to the Speaker. When the Experimenter performed a Modifier Demonstration, the Speaker's bag of toys did not include the Target Instrument (e.g., the large flower) but instead included a second unrelated object (e.g., a leaf). Hence, the Instrument interpretation of the with-phrase should be less available to Speakers in this context.

In addition, we excluded the unambiguous conditions in this experiment, because these sentences had been uniformly interpreted and coded correctly in Experiment 1. To equalize the number of ambiguous sentences that subjects received in each experiment, we divided the 16 critical sentences into two lists. The items on each list appeared in a pseudo-random order embedded in the same twenty-four distractor trials. In addition, reverse-order lists were generated.

Finally, the type of Demonstration was manipulated between subjects. In the Instrument Condition, all target items were ambiguous and involved an Instrument demonstration (and an Instrument context for the Speakers). In the Modifier Condition, all target items were ambiguous and involved a Modifier demonstration (and a Modifier context for the Speakers).

**Results**

**Ambiguity Awareness** Listeners in Experiment 2, like those in Experiment 1, usually reported that they were aware of the ambiguity. This is to be expected, given that the same referential contexts were presented to Listeners in both experiments, and it suggests that the between-subjects design does not, by itself, affect ambiguity awareness.

Speaker awareness of ambiguity did change across experiments. In particular, only one speaker in the Instrument condition (6%) reported being aware of the ambiguity. Interestingly, and in contrast, nine of the Speakers in the Modifier condition, or 56%, reported being aware of the ambiguity. This pattern was unexpected; we were hoping that few if any of the Speakers would be aware of the ambiguity. This difference may be related to the fact that Modifier attachments are dispreferred by readers, especially with action verbs (Spivey-Knowlton & Sedivy, 1994). Conflicts between lexical and referential cues in the Modifier condition may brought the ambiguity into awareness. As we shall

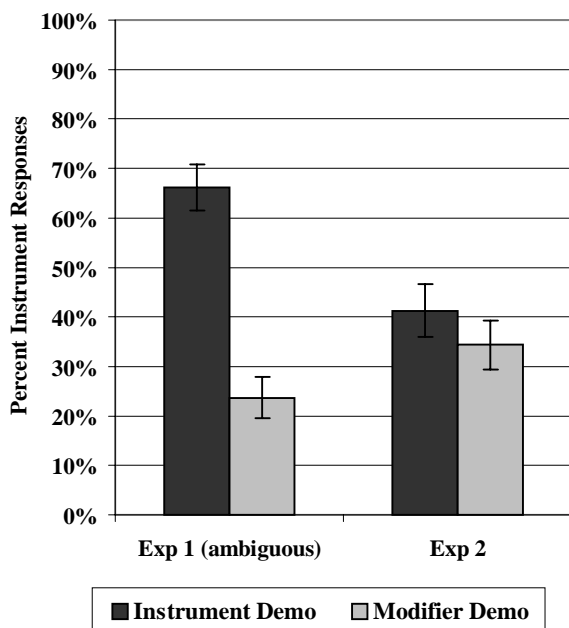
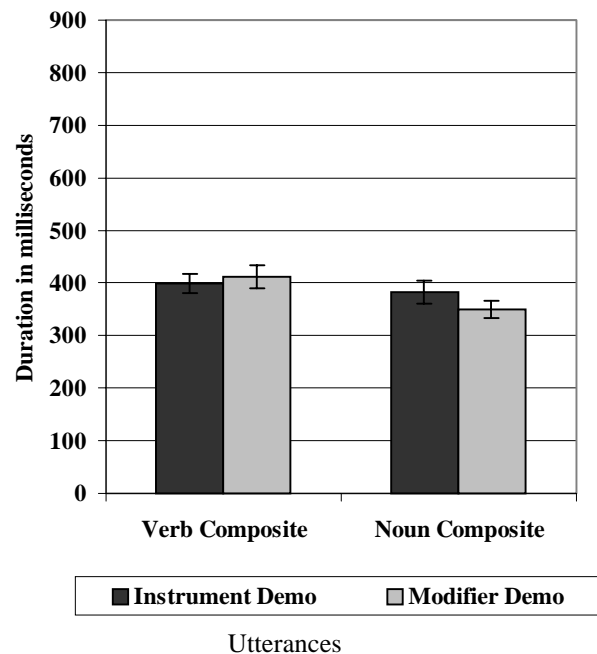


Figure 3: Listeners' Responses to the Ambiguous Sentences in Experiments 1 & 2  
see, this unexpected result is serendipitous, because it allows us to compare the performance of the Listeners who heard utterances from aware and unaware Speakers.

**Listener's Performance** The percent of Instrument responses for each conditions of Experiment 2 appear on the right hand side of Figure 3. Listeners in Experiment 2 were clearly unaffected by the type of Demonstration performed by the ( $F(1,24) < 1, p > .3$ ;  $F(1,12) = 1.88, p > .3$ ), suggesting that Speakers were not effective in helping Listeners resolve the ambiguity. To compare performance in the ambiguous conditions of Experiments 1 and 2, ANOVAs were conducted on the percent correct for items and subjects with Experiment as a between subjects and within items factor. Unsurprisingly, Listeners in Experiment 1 performed significantly better than those in Experiment 2 ( $F(1,62) = 11.76, p < .001$ ;  $F(1,14) = 19.91, p < .001$ ).

**Speaker's Prosody** The audiotapes were digitized and coded in the manner described above. As Figure 4 suggests, there was no reliable effect of condition on the duration of the Verb Composite ( $F(1,24) < 1, p > .6$ ;  $F(1,12) = 1.69, p > .2$ ). The effect of condition on the Noun Composite was not significant in the subjects analysis ( $F(1,24) = 1.66, p > .2$ ) and was small but reliable in the items analysis ( $F(2,1,12) = 7.71, p < .05$ ).

Figure 4: Experiment 2, Mean Durations from Speakers'



**General Discussion**

When the Speakers' context strongly supported the intended meaning of an ambiguous utterance, Listeners showed complete insensitivity to the intended meaning of the utterance. This stands in contrast to Experiment 1, where Listeners had

the same referential context, but were highly sensitive to the intended meaning. There are two possible explanations to these findings. First, the findings may be attributable to Speakers' awareness of the ambiguity (Allbritton et al, 1996). Speakers in Experiment 1 were almost always aware of the ambiguous sentences while those in Experiment 2 were usually unaware. Second, the findings may be due to the change in the referential context of the Speakers. In the first experiment, Speakers were given a context that supported either meaning of the ambiguous sentence, whereas in Experiment 2 Speakers were provided with a context that supported only the relevant interpretation. Perhaps speakers only produce informative prosody when the context doesn't disambiguate the sentence for them (Lieberman, 1967).

Because a subset of the subjects reported being aware of the ambiguity in the Modifier condition, we can test if ambiguity awareness per se is driving the results of Experiment 2. Ten additional subject pairs were tested to gather sufficient data for this comparison. We found no reliable differences between responses to aware speakers and those to unaware speakers ( $F(1,23) < 1, p > .8$ ;  $F(1,15) < 1, p > .9$ ). When the referential context disambiguated the sentence, awareness of the **potential** for ambiguity did not lead speakers to provide adequate prosodic cues.

The substantial difference in performance across the studies and the absence of an effect of awareness within Experiment 2, suggest that referential context itself is critical in determining whether speakers will produce strong prosodic cues. In Experiment 1, the Speaker's referential context supported both interpretations of the "with" phrase and thus the sentence was, in the absence of prosodic cues, ambiguous in context. Under these conditions, Speakers produced prosodic cues that were not only consistent with the intended structure but also inconsistent with the competing interpretation. Listener's were able to use these cues to determine the intended meaning, albeit imperfectly. In Experiment 2, the Speaker's referential context supported only the relevant interpretation, disambiguating the sentence and making strong prosodic cues unnecessary. Listeners, who did not have access to this disambiguating referential context, were able to find nothing in the Speakers' prosody to guide them. These data, therefore, support Lieberman's hypothesis that speakers only produce informative prosody when the context doesn't do the work for them.

This conclusion and these results appear to conflict with those of Schafer and colleagues who find that untrained, uninformed speakers produce consistent prosodic cues regardless of whether the context of the utterance provides disambiguating information (1999). To add to the confusion the tasks appear to be quite similar: both experiments use variants of the referential communication task and ask speakers to produce scripted, memorized commands to achieve concrete results. The two experiments, however varied in several critical respects. First, in the Schafer study, there is a higher degree of uncertainty about the listener's referential context. The speaker knows both that the listener has information about the context that the speaker lacks and that the listener's context will change as the experiment progresses. Second, the participants are given the set of commands, which contains both interpretations of the ambi-

guity, at the beginning of the study and are exposed to situations in which each meanings is applicable. Thus it seems likely that these speakers were aware of the globally ambiguous sentences and believed that there was the potential for referential ambiguity.

In this paper, we have suggested that a speaker's knowledge of the referential situation affects her ability to disambiguate otherwise ambiguous utterances. In particular, we propose that when a speaker recognizes that an utterance is ambiguous in context, she will disambiguate it by making prosodic choices that are consistent with the relevant interpretation and inconsistent with the alternatives.

### Acknowledgments

We thank Lisa Levine, Sarah Brown-Schmidt and Jared Novick for their assistance, input and patience. This work was supported by National Institute for Health Grant 1-R01-HD3750707-01 and a National Science Foundation Center Grant to the University of Pennsylvania Institute for Research in Cognitive Science.

### References

- Allbritton, D., McKoon, G. & Ratcliff, R. (1996). Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **22**, 714-735.
- Cooper, W.E. & Paccia-Cooper J. (1980). *Syntax and Speech*. Cambridge MA: Harvard University Press.
- Fernald, A. & McRoberts, G. (1996). Prosodic bootstrapping: A critical analysis of the argument and the evidence. In J. Morgan & K. Demuth (Eds), *Signal to Syntax* Mahwah, NJ: Erlbaum.
- Lehiste, I. (1973). Phonetic disambiguation of syntactic ambiguity. *Glossa*, **7**, 102-122.
- Lieberman, P. (1967). *Intonation, perception and language*. Cambridge MA: MIT Press.
- Kjelgaard, M. & Speer, S. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language*, **40**, 153-194.
- Schafer, A., Speer, S., Warren, P., & White, D. (1999). Intonational disambiguation in sentence production and comprehension. Paper presented at the Twelfth Annual CUNY Conference on Sentence Processing, New York, NY, March 1999.
- Selkirk, E.O. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.
- Spivey-Knowlton, M. & Sedivy, J. (1995) Resolving attachment ambiguities with multiple constraints. *Cognition*, **55**, 227-267.
- Wales, R. & Toner, H. (1979). Intonation and ambiguity. In W.E. Cooper and E.C.T. Walker (Eds.), *Sentence Processing: Psycholinguistic studies presented to Merrill Garrett*. Hillsdale, NJ: Erlbaum.
- Warren, P (1996). Prosody and parsing: An introduction. *Language and Cognitive Processes*, **11**, 1-16.

# Eye Movements During Comprehension of Spoken Scene Descriptions

Michael J. Spivey  
(spivey@cornell.edu)

Melinda J. Tyler  
(mjt15@cornell.edu)

Daniel C. Richardson  
(dcr18@cornell.edu)

Ezekiel E. Young  
(eey1@cornell.edu)

Department of Psychology  
Cornell University  
Ithaca, NY 14853 USA

## Abstract

A recent eyetracking experiment has indicated that, while staring at a blank white display, participants engaged in imagery tend to make eye movements that mimic the directionality of spatial expressions in the speech stream (Spivey & Geng, 2000). This result is consistent with a spatial mental models account of language comprehension (e.g., Johnson-Laird, 1983), adds a *motor* component to evidence for activation of *perceptual* mechanisms during visual imagery (e.g., Kosslyn, Thompson, Kim, & Alpert, 1995), and fits with claims regarding the embodiment of cognition (e.g., Varela, Thompson, & Rosch, 1991). However, some methodological concerns remain. We report some preliminary observations, and a controlled experiment, in which these methodological concerns are resolved. We demonstrate that, even when the speech includes no instructions to imagine anything, and even when participants' eyes are closed, participants tend to make eye movements in the same direction (and especially along the same axis) as the described scene when listening to a spatially extended scene description.

## Introduction

More than three decades ago, Donald O. Hebb (1968) suggested that the very same eye-movement scanpaths associated with *viewing* an object may be automatically triggered (via transcortical cell assemblies) when a person is *imagining* that object -- and some empirical support for this claim has recently been reported. When viewing a blank screen and being instructed to imagine a previously-viewed block pattern, observers produced scanpaths that bore some resemblance to the scanpaths elicited during original viewing of the actual block pattern (Brandt & Stark, 1997).

Such oculomotor behavior in the absence of visual input is consistent with the notion that, when imagining or remembering an object or event, we often develop a mental representation of that object or event that has a distinctly spatial structure to it. This spatial format of representation is thus able to take advantage of properties inherent to Cartesian space, such as topography and metric relationships. During the construction and interrogation of such spatial mental models (e.g., Bower & Morrow, 1990; Bryant, 1997; Johnson-Laird, 1983, 1996), cognition often uses linguistic input to activate memory representations, and imagery may then use those memory representations to partially activate perceptual representations (e.g., Farah, 1995; Finke, 1986; Kosslyn et al., 1995).

The present study demonstrates that, even in the absence of any visual stimulus at all, such "perceptual simulations" (Barsalou, 1999) often trigger corresponding oculomotor responses. In a sense, one might say that *thinking* of something often involves *pretending to look at it*. This finding contributes to the developing "embodied" view of the mind (e.g., Ballard, Hayhoe, Pook, & Rao, 1997; Brooks, 1995; Varela, Thompson, & Rosch, 1991), in which an adequate characterization of cognition requires special attention to the repertoire of actions available to the organism or agent.

## Looking at Objects that Aren't There

In a recent study, Spivey and Geng (2000, Experiment 1) had participants simply listen to pre-recorded instructions to imagine visual scenes while looking at a blank white projection screen and wearing a headband-mounted eyetracker. Each of the descriptions had a specific directionality (rightward, leftward, upward, and downward) to the manner in which new objects or events were introduced in the scene. In addition, a control scene description was presented, in which no particular directionality was present.

Pilot results with this methodology produced eye movement patterns very much in accordance with the directionality of the scene description, however most participants developed rather accurate suspicions of our experimental hypothesis. Although eye movements are relatively automatic, and usually not very susceptible to voluntary control, the concern remained that participants may not have produced such behavior if they hadn't known that their eye movements were being recorded.

To avoid potential strategy effects, we introduced a sham task (of following instructions to move objects around on a table), and referred to the imagery session as a break from the experiment during which the eyetracker would be turned off ("but don't take off the headband because then we'd have to recalibrate the tracker when we return to the experiment"), Although two participants suspected that their eyes were still being tracked, and two participants closed their eyes during the imagery session, the remaining six participants produced eye movement patterns that were remarkably consistent with the directionality of the scene descriptions. Figure 1 shows example data from the Control (left panel) and Rightward (right panel) scene descriptions.

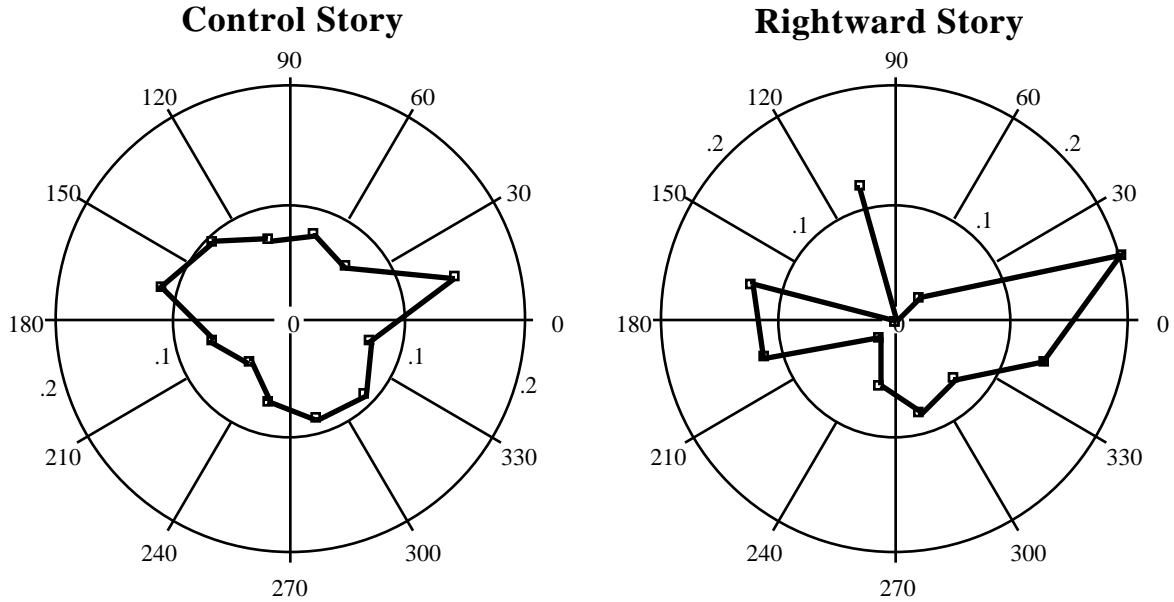


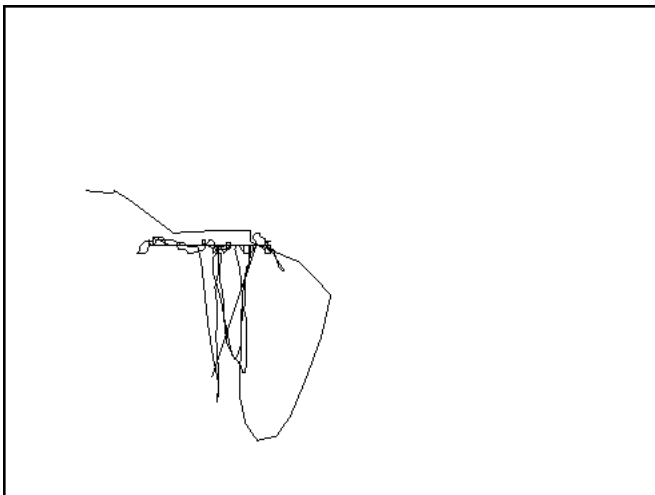
Figure 1: Polar coordinate plots of the average proportion of saccades in various directions while listening to stories. In the control story, participants made an approximately equal proportion of eye movements in all directions. In the rightward story, .2 and .14 of all saccades were in the two rightward directions (15° and 345°, respectively).

Despite remarkable direction-selectivity in saccades during the imagery instructions (for all four directionally-biased scene descriptions), some concerns about this methodology remained. To begin with, the scene descriptions began with an explicit instruction to "imagine" the scene, which may be importantly different from normal language use. Moreover, because they knew that eye movements were involved in the rest of the experiment, participants' eye movement behavior may still have been somehow unnatural. Ideally, these findings needed to be replicated under circumstances where there were no explicit instructions to imagine something, and where the participant had no idea that his/her eye movements were being recorded.

### Observation

We have been developing a methodology for recording a participant's eye movements with an ISCAN, Inc. remote eyetracking camera without the participant's knowledge. With a benign deception, great care is taken to prevent participants from discovering that their eye movements are being recorded, while still protecting their rights. For our purposes, the deception involves telling participants that the camera directed at them is recording subtle thermal changes in the face as a result of emotional arousal. They are encouraged to sit as still as possible during the experiment in order to allow an accurate thermal image.

### Control Story



### Rightward Story

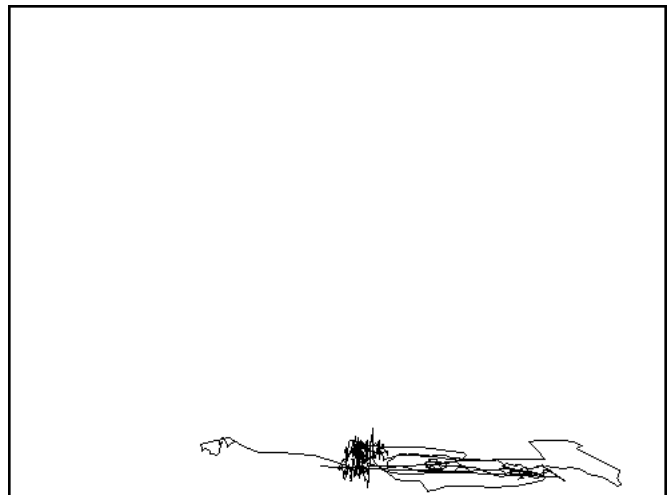


Figure 2: Example scanpaths from a participant while listening to the Control and Rightward scene descriptions.

For calibration, we use five pseudo-Rorschach images mounted on the corners and center of a black foam core board spanning 20 X 30 degrees of visual angle. With the cover story being that we must first collect some baseline thermal readings, participants are asked to look at each of the pseudo-Rorschach images and report what they see, during which time we record those five eye positions as calibration points. The ISCAN system then interpolates between those calibration points to provide an eye position signal with an accuracy of better than 1 degree of visual angle. In order to reduce the amount of visual information in the environment that participants might fixate during the time that they listen to the pre-recorded stories, the poster is then flipped over to display the black-colored board.

Table 1: Pre-recorded scene descriptions

---

### CONTROL STORY

"You are on a hill looking at a city through a telescope. Pressing a single button zooms a specific block into view. Another button brings a gray apartment building into focus. Finally a third button zooms in on a single window. Inside you see a family having breakfast together. A puppy appears and begs for a piece of French toast."

### RIGHTWARD STORY

There is a fishing boat floating on the ocean. It's facing leftward from your perspective. At the back of the boat is a fisherman with a fishing pole. The pole extends about 10 feet to the right beyond the edge of the boat. And from the end of the pole, the fishing line extends another 50 feet off to the right before finally dipping into the water."

### DOWNWARD STORY

"You are standing at the top of a canyon. Several people are preparing to repel down the far canyon wall across from you. The first person descends 10 feet before she is brought back to the wall. She jumps again and falls 12 feet. She jumps another 15 feet. And the last jump, of 8 feet, takes her to the canyon floor."

### LEFTWARD STORY

"There is a train extending outwards to the left. It is pointed to the right, and you are facing the side of the engine. It is not moving. Five cars down is a cargo holder with pink graffiti sprayed on its side. Another six cars down is a flat car. The train begins to move. Further down the train you see the caboose coming around a corner."

### UPWARD STORY

"You are standing across the street from a 40 story apartment building. At the bottom there is a doorman in blue. On the 10th floor, a woman is hanging her laundry out the window. On the 29th floor, two kids are sitting on the fire escape smoking cigarettes. On the very top floor, two people are screaming."

---

The pre-recorded auditory stimuli consist of ten short stories, each story lasting approximately thirty seconds. Half of the stories are filler stories which contain no directional cues, but which are intended to be slightly emotionally-engaging, in order to divert the participants from suspecting the actual experimental hypothesis. The five test stories are derived from Spivey & Geng (2000), but do not begin with an instruction to "Imagine...". The test stories contain systematic directional cues (see Table 1), and are not emotional in content. The order of the ten stories is pseudo-randomized, with one filler story interleaved between each pair of test stories. After about five minutes for calibration, the listening session begins with a filler story, and lasts another five minutes.

At the end of the session, participants are debriefed and thoroughly questioned as to their beliefs about the nature of the study. With the thermal camera cover story, very few of the participants have correctly guessed the experimental hypothesis -- almost all participants were surprised that their eye movements were being examined rather than patterns of thermal change on the face.

In developing this methodology, we have encountered numerous complications in acquiring and testing an accurate calibration of the eyetracking system without betraying the deception. Due to blinks and other movements, an accurate calibration with the remote eyetracking camera occasionally requires multiple fixations of some of the five calibration points. As the calibration period drags on with more and more inventive questions regarding the five pseudo-Rorschachs, participants can become suspicious of the cover story. Moreover, following an acceptable calibration, head movements during the listening phase of the experiment can make it difficult for the software, or a human controller, to maintain a centralized camera image of the eye.

Despite the complications, some preliminary observations from this methodology are available and worth reporting. It is already clear that listeners make a greater number of saccades during the directionally-biased stories compared to the Control Story. Although these eye movements are not always in the specific direction of the directionally-biased story, they do tend to be limited to the appropriate axis of orientation; see Figure 2. This should be expected if one considers the fact that a listener simply couldn't make rightward or upward eye movements indefinitely as the story continues to add rightward or upward expressions. At some point, an eye movement in the opposite direction is necessary to "re-center" the imagined scene in head-centered coordinates. (It is possible that this might occur less frequently if participants were allowed to turn their heads.) In any case, there is nothing stopping the listener from voluntarily "examining", or "looking back to", previously described elements of the scene.

Further development of this methodology, with more accurate tracks from additional naive participants, will allow averaging of saccades in polar coordinates (as shown in Figure 1). Preliminary data appear likely to confirm the findings of Spivey and Geng (2000), under circumstances where there are no explicit instructions to imagine anything, and participants are completely unaware that their eye movements are being recorded.



## Experiment

The described complications with the above methodology (as well as concerns about participants looking at objects in the visual field beyond the 20° X 30° black display board) point to a methodology with which it is easier to collect data from more reliably naive participants, but which produces data that is admittedly more difficult to analyze. In this experiment, participants were instructed to close their eyes while they listened to the same ten stories; the stories in Table 1 as well as the five filler stories. A standard video camera was directed at the participant's face, and the camera's image of the participants' closed eyes was later used to estimate incidence and direction of eye movements. (One could, in principle, record movements of closed eyes with electrooculography [surface electrodes near the eyes] or a search coil [a contact lens with a copper wire and loop, the position and orientation of which is precisely determined via an electromagnetic field in which the participant sits]. However, it might be difficult to convince such participants that their eye movements were not being recorded.)

## Method

**Participants** Eleven Cornell University undergraduates participated in the study for extra credit in Psychology courses. None of them had previously participated in an eyetracking experiment.

**Stimuli and Apparatus** This experiment used the same pre-recorded scene descriptions as described in the previous methodology. A standard video camera was positioned in front of the participant, with a black curtain as background. Stories were presented in a pseudo-randomized order, with each test story preceded by a filler story.

**Procedure** Participants were asked to relax in their seat, but to remain still and to close their eyes while listening to the ten short stories. They were informed that we would be examining their facial expressions as related to story content. Participants were told that their shoulders and face were being videotaped while the stories were being played. Upon achieving a well-focused image of the participant's closed eyes through the video camera, recording began. The stories were played from a cassette player. Each session lasted approximately five minutes. At the end of each session, participants were debriefed and thoroughly

questioned as to their beliefs about the nature of the study. None of the participants correctly guessed the study hypothesis, and all were surprised that their eyes movements had been of interest.

**Coding** Coding was made easier by focusing on the movement of a round spot of luminescence on each of the participant's eye lids. The spot of light reflectance on the eye lids corresponded to the protuberant round area of the cornea directly over the pupil sitting beneath the lid. Movements of the eye were considered to be clearly visible shifts in the spot of reflectance on each lid (as opposed to ambiguous twitches in the lid, or jitters in the positioning of the spot of reflectance that were too small in distance and too short in duration to be easily interpreted.) See Figure 3, where dark lines are added to indicate the points of inflection due to the corneal bulge. Such eye movements are much easier to discern when multiple frames are seen in real-time (<http://node15.psych.cornell.edu/home/eyesclosed.html>). A definite movement of the eye ball was considered to be a movement of the spot of reflectance lasting for 3 video frames or longer and of such a distance that the direction of movement was unambiguous to 2 independent trained coders. Assessment of start time and direction of eye movements involved pinpointing the movements on video tape by playing approximately 10 frames of tape at a time, then comparing the previous position of the spot of reflectance with the new position. It was frequently necessary to rewind or forward the position of the tape one frame at a time to specify as precisely as possible when each eye movement took place. Inter-rater reliability for the two coders was high, as measured by a Pearson correlation;  $r=.84$ . It is difficult to estimate how large a saccade (in degrees of visual angle) is detectable with this coding method. However, it is likely that many small eye movements are being made in this task, the direction of which cannot be determined with the present method.

## Results

For each scene description, proportion of detectable eye movements in each of eight directions was averaged across all eleven participants. Two of the eleven participants made no detectable eye movements during the entire experiment. Consistent with the previous observations, only three of the eleven participants made any detectable eye movements

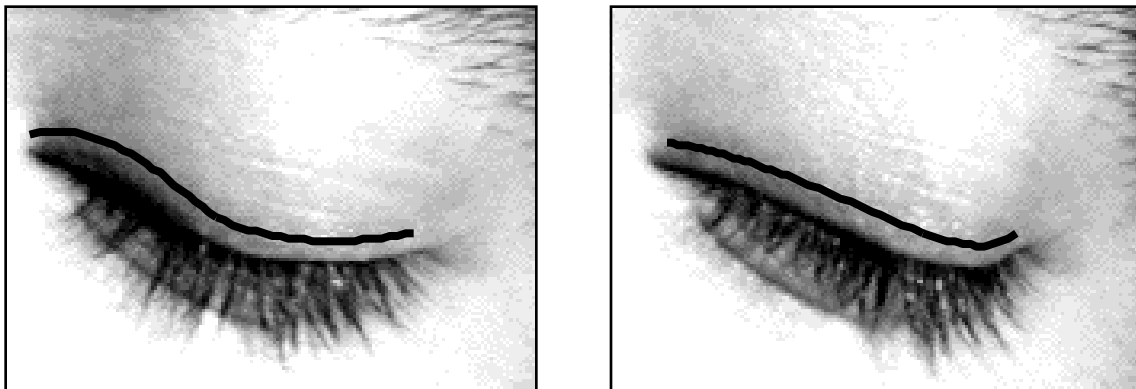


Figure 3: Examples of a detectable rightward eye position (left panel) and leftward eye position (right panel).

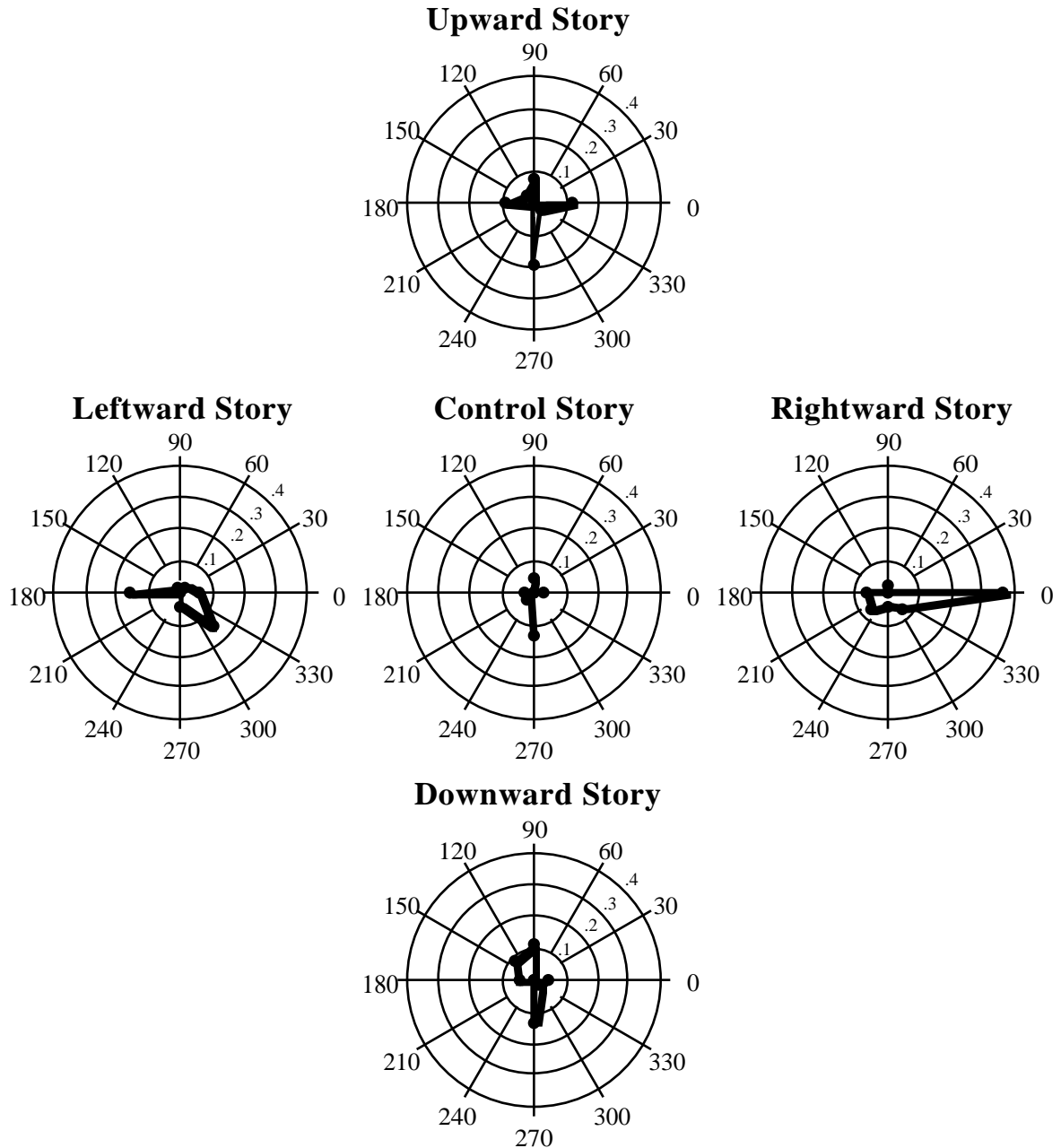


Figure 4: Polar coordinate plots of the average proportion of detectable eye movements in eight directions while participants listened to scene descriptions with their eyes closed.

during the Control Story. Unfortunately, this contributes to a rather unreliable estimate of the direction-selectivity profile for that story. Figure 4 shows polar coordinate plots of the eye-movement direction-selectivity profiles for the five scene descriptions. It is noteworthy that when the detectable eye movements were in the unpreferred direction, they were frequently in the exact opposite direction. Thus, even when the eye movements do not follow the specific directionality of the scene description, they nonetheless tend to be limited to the appropriate axis of orientation (horizontal or vertical), suggesting that an entire axis of a reference frame, rather than simply one direction, may be activated (e.g., Carlson-Radvansky & Jiang, 1998; Demarais & Cohen, 1998). In a

paired t-test, the average proportion of eye movements in a preferred direction was significantly greater than the average proportion of eye movements in the unpreferred directions;  $t(10) = 4.49, p < .01$ .

### Discussion

These results demonstrate that, even when participants' eyes are closed, they tend to move their eyes in directions that accord with the directionality of the scene being described. Although comprehension of these scene descriptions may involve some of the same mechanisms involved in imagery tasks (e.g., Kosslyn et al., 1995), the present results do not

rely on explicit instructions to imagine anything. We suggest that comprehension of scene descriptions employs a decidedly spatial format of representation, and that oculomotor coordinates may be an important component of that representation.

## Conclusions

With no visual information available, participants constructing mental models of complex scenes tend to make eye movements that mimic the kinds of eye movements that would be made when viewing that actual scene. In a similar vein, Spivey and Geng (2000, Experiment 2) and Richardson and Spivey (in press) report experiments in which participants attempting to recall visual or auditory information tend to make eye movements to the region of space (now empty) where that information was first encoded. In conjunction with results in visuomotor coordination by Ballard et al. (1997) and in attention by Pylyshyn (1994), this work suggests that the mind/brain exploits useful properties inherent to spatial formats of representation by relying on oculomotor pointers to spatial indexes out in the world or, in the present case, perceptual simulations of them. This perspective may point to a pivotal role for space, independent of perceptual modality, in all mental representations (cf. Bryant, 1997).

One could possibly interpret our findings as a tangential endorsement of a sort of 'Cartesian Theater' account of the mind (cf. Dennett, 1991): that when we imagine something, it is like viewing it "in our mind's eye", and that (perhaps epiphenomenally) our real eyes simply echo the motion of our internal spectating. However, it is well known that this kind of interpretation too easily falls into the infinite regress of a homunculus inside the mind. Therefore, we prefer to place these results in the light of Ryle's (1949) comment that "[A person picturing his nursery in his mind's eye] ... is not being a spectator of a resemblance of his nursery, but he is resembling a spectator of his nursery." That is to say, we would argue that our data are indicative of an embodied system that naturally activates 'lower level' motor actions to accompany 'higher level' cognitive processes because, rather than being separate functions that are triggered by a mental state, motor actions are fundamental components of the mental state.

In such an embodied view of the mind, action determines cognition as much as perception does. Indeed, a number of researchers have suggested that an important aspect of *perceiving* an environment is knowing *how to interact* with it (e.g., Brooks, 1995; Gibson, 1979; Milner & Goodale, 1994; Turvey & Carello, 1996). Perhaps we can add to this embodied view of perception that part of *perceiving* a scene is also knowing *how to look at it* -- even when it's not there.

## Acknowledgments

We are grateful to Michael Tanenhaus, Mary Hayhoe, and Dan Earley for helpful discussions of this work, and to Jessica Evett-Miller and Koji Park for assistance with data collection and analysis. Supported by a Sloan Foundation Fellowship in Neuroscience and a grant from the Consciousness Studies Program at the U. of Arizona.

## References

- Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*, 723-767.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577-660.
- Bower, G. H. & Morrow, D. G. (1990). Mental models in narrative comprehension. *Science*, *247*, 44-48.
- Brandt, S. A. and Stark, L. W. (1997) Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, *9*, 27-38.
- Brooks, R. (1995). Intelligence without reason. In L. Steels & R. Brooks (Eds.), *The artificial life route to artificial intelligence: Building embodied, situated agents*. Hillsdale, NJ: Erlbaum.
- Bryant, D. J. (1997). Representing space in language and perception. *Mind and Language*, *12*, 239-264.
- Carlson-Radvansky, L. A. & Jiang, Y. (1998). Inhibition accompanies reference-frame selection. *Psychological Science*, *9*, 386-391.
- Demarais, A. M. & Cohen, B. H. (1998). Evidence for image-scanning eye movements during transitive inference. *Biological Psychology*, *49*, 229-247.
- Dennett, D. (1991). *Consciousness explained*. Boston: Little, Brown and Co.
- Farah, M. J. (1995). Current issues in the neuropsychology of image generation. *Neuropsychologia*, *33*, 1455-1471.
- Finke, R. A. (1986). Mental imagery and the visual system. *Scientific American*, *254*, 88-95.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, Massachusetts: Houghton Mifflin.
- Hebb, D. O. (1968). Concerning imagery. *Psychological Review*, *75*, 466-477.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Cambridge University Press.
- Johnson-Laird, P. N. (1996). Space to think. In P. Bloom & M. Peterson (Eds.), *Language and space*. Cambridge, MA: MIT Press.
- Kosslyn, S. M., Thompson, W. L., Kim, I. J., & Alpert, N. M. (1995) Topographical representations of mental images in primary visual cortex. *Nature*, *378*, 496-498.
- Milner, A. D. & Goodale, M. A. (1995). *The visual brain in action*. Oxford: Oxford University Press.
- Pylyshyn, Z. (1994). Some primitive mechanisms of spatial attention. *Cognition*, *50*, 363-384
- Richardson, D. C. & Spivey, M. J. (in press). Representation, space, and Hollywood Squares: Looking at things that aren't there anymore. *Cognition*.
- Ryle, G. (1949). *The concept of mind*. London: Hutchinson.
- Spivey, M. J. & Geng, J. J. (2000). *Oculomotor mechanisms triggered by imagery and memory: Spontaneous eye movements to objects that aren't there*. Manuscript submitted for publication.
- Turvey, M. & Carello, C. (1995). Some dynamical themes in perception and action. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.
- Varela, F. J., Thompson, E., & Rosch E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.

# Heterogeneous reasoning in learning to model

**Keith Stenning** (K.Stenning@ed.ac.uk)

Human Communication Research Centre, Division of Informatics, University of Edinburgh  
2 Buccleuch Place, Edinburgh EH8 9LW, UK

**Melissa Sommerfeld** (msomme@leland.stanford.edu)

Cognition and Learning Lab, CERAS 105, Stanford Ca. 94305-3084, USA

## Abstract

Conceptual learning in maths and science involves learning to coordinate multiple representation systems into smoothly functioning heterogeneous reasoning systems composed of sub-languages, graphics, mathematical representations, etc.. In these heterogeneous systems information can be transformed from one representation to another by inference rules, and learning coordination is learning how and when to apply these rules. Heterogeneous reasoning has a particularly important role to play in teaching students how to apply formalisms to real world problems, rather than merely teaching formalism-internal calculation.

This paper analyses three learning incidents which happened in groups of students engaged in learning the mathematics and biology involved in modelling biological populations, from the perspective of the heterogeneous reasoning involved. Greeno, Sommerfeld & Weibe (2000) and Hall (2000) analyse incidents from the same curriculum intervention from other points of view, in this volume.

We observe both learning successes and failures that cannot be understood without understanding the seams joining the representation systems involved, and the inference rules and operations required to get from one to another. One conclusion is that even apparently homogeneous natural language has to be seen as heterogeneous in its fully contextualised application.

## Introduction

The coordination of multiple representation systems is frequently instrumental in conceptual learning (see e.g. van Someren et al. 1999; Barwise & Etchemendy 1994; Stenning, Cox & Oberlander 1995). In particular, learning mathematics and science concepts involves learning to coordinate multiple formalisms (numerical, graphical, algebraic, terminological), but it also involves learning how to apply formalisms in contexts. It is all too possible for students to succeed at the first and to fail at the second—to learn the internal operation of some formalism without learning how to apply it to new problems. Barwise and Etchemendy have used the term ‘heterogeneous reasoning’ for reasoning using multiple coordinated representations, and have applied heterogeneity of representation in order to improve students’ grasp of the application of formalisms in computer environments such as Hyperproof. Another related curriculum response to this problem of teaching the application of formalisms to real world problems has been a move toward project-based approaches which teach formalisms in close relation to their context of application—in particular teaching scientific concepts along with the mathematics that goes with them.

The purpose of this paper is to use some example episodes from project-based group learning to illustrate how the concepts of heterogeneous reasoning present themselves in the classroom setting in less formal contexts than Hyperproof. This investigation underscores how local the semantic interpretation of representations is in context. Words change meaning frequently and systematically, and the information they carry is moved into and out of other representations. The investigation also provokes examination of the relation between heterogeneity and localness of interpretation (e.g. Moravcsik 1998). With diagrams, it is usually quite evident to users that the diagram has a local interpretation and that the naive user needs to learn this local interpretation, even though there are regular features of such diagrammatic systems from use to use. With natural language, we are often so practiced at making the contextual interpretation of its local semantics that it is easy to fail to realise that this is what we do. Examining learning discourse in context raises the question whether heterogeneity should also be extended to cases where the linguistic part of the discourse has to be treated as multiply interpreted.

Our longer term aim is to coordinate this approach with others which focus on discourse practices and students’ recruitment of material from their diverse experiential worlds. Greeno, Sommerfeld & Weibe (2000) and Hall (2000) take these respective perspectives on material from the same group learning curriculum intervention. The advantage of pursuing several parallel analyses of the same data for cognitive theory may share something with the advantages of the project-based approach for the students. Applying several kinds of the theory to the same episodes turns up new questions about how the theories relate to each other, and thus may induce conceptual learning and improved ability to apply the theories in novel circumstances.

## Heterogeneous reasoning

Theories of human reasoning have begun to pay more attention to how representational systems are selected or constructed, and the variety of systems that may be used in solving a single problem, rather than conceiving of reasoning as a system internal activity. Barwise & Etchemendy have called this use of multiple coupled systems of representation *heterogeneous* reasoning, and have developed several computer environments for teaching heterogeneous reasoning. For example, Hyperproof presents a graphical window containing diagrams of a blocks-world inhabited by regular solids on a chequerboard, and a sentential window containing first order

logic sentences. The proof rules of the heterogeneous system incorporate the inference rules of the conventional sentential calculus, augmented by rules for moving information between diagram and sentences, in both directions. For example, the user can *observe* a feature of the diagram as the basis for inferring a sentence; or may *apply* information from a sentence by inferring (and constructing) a new feature of the diagram. *Observe* and *apply* are two (of about a half dozen) of the heterogeneous inference rules which coordinate the diagrammatic and sentential representations into a heterogeneous system.

Fundamental semantic distinctions between how diagrammatic and sentential representation systems express abstraction have been shown to play an essential part in analysing the learning that occurs as students master the construction of proofs in Hyperproof's heterogeneous environment, allowing the learning to be characterised as learning strategies of representation selection and use (Stenning, et al. 1999). Whether students benefit from the diagrammatic facilities of Hyperproof is determined to a great extent by their facility at grasping useful strategies for using Hyperproof's expressions of graphical abstraction.

Hyperproof reveals an important property of representational systems in use. Its semantics, both of its sentences and its diagrams are *partially interpreted*. That is, the system has some of its meaning fixed while other parts are defined in episodes of reasoning. This contrasts with the usual presentation of first order logic as an entirely uninterpreted language. In Hyperproof, even the sentences of the first order calculus have to be given a partial local interpretation because the predicates and relations have to coincide with those of the diagrams.

Partly diagrammatic systems of representation like Hyperproof reveal the need for coordinating diagrammatic and sentential representation systems, but lead to the further realisation that in situations of real language use, the apparently homogeneous languages in play are in fact often heterogeneous in the fundamental sense that many schemes of interpretation are in play at once. Even when natural language is the only modality, the reasoning systems in operation must be thought of as heterogeneous because the apparently single language can only be understood in terms of overlapping language fragments, each constituting a distinct system of representation.

To illustrate this point, this paper takes some classroom interactions of a group of students learning to model biological populations in terms of mathematical functions, and analyses the multiple partially interpreted representation systems which are in play. The students' representational resources and activities include at least the following: worksheet filling, graph drawing, computer operation, calculator use, group speech and gesture, reference material, and teacher interventions.

The educational issue in focus is the learning about modelling, and particularly learning about the process of formalisation and interpretation. A recurring theme is the struggle to coordinate formalism internal operations (calculation) and formalism external correspondences (semantics). We will analyse both successes and failures of coordination.

## The educational setting

The data we analyzed comes from an 8th grade Middle-school Mathematics through Applications (MMAP) classroom in the San Francisco Bay Area. The purpose of MMAP is to have students use math to address real-world problems, often with the assistance of computer applications. In the approximately 30 day unit we will discuss, called Guppies, students created mathematical models of biological population growth. As part of this unit students were to learn both about how to construct mathematical models of population growth and about the exponential functions that underlie them. Our analyses focus on a group of students, Manuel, Lisa, Kera & Nick whose improvement on pre/post assessments placed them about midway in learning of the half a dozen focus groups videotaped by Rogers Hall and his colleagues (Hall, 1999) during this unit in a variety of classrooms. These students are chosen to reflect roughly average performance for the class. For more information about the design of the study, please see Hall (2000, this volume).

## Three learning incidents

The three following incidents were chosen from videotapes because they illustrate both successful and problematic learning episodes. The initial incident from the pre-test phase sees the students make at least part of one of the fundamental conceptual discoveries of this field—that population models have a recursive characteristic that leads to exponential growth if unchecked—Malthus' equation.

The second incident, from the body of the course, is of interest because it contains an attempt to creatively diverge from the structure of the assigned worksheet by taking a short cut in the calculation. On the one hand this divergence reveals the germ of another important insight—that functions can be composed. But in the circumstances, the insight is not fully worked out and leads to error and confusion.

The third incident is chosen to illustrate that the confusion that is not resolved in the previous incident appears to persist into the much later post-test phase of the course. It consists of another attempt to calculate a birthrate for a new modelling problem.

In all of these incidents, the students struggle to coordinate multiple representations. We examine some of the coordinations in detail seeking to reveal how some episodes are successful and others not. For this short paper, the transcriptions are compressed by leaving out material which does not relate to our analysis.

### Pre-test insight—'babies have babies'

When the group discovers the recursive nature of population growth, they are engaged in constructing a model of a mouse population. They have obtained an initial number of 20 adults from the worksheet, and estimated a birthrate of four per couple. They are now calculating what the population will be after eight breeding seasons. The group initially adopt a linear model implicit in multiplication of a fixed birthrate. Only when they turn to the graphing activity dictated by the worksheet do they begin to think of the process which the calculation is intended to reflect.

60: M. so there's ... equals 40 babies each season

65: M. it's three hundred and twenty

- 66: K. (inaudible) is that including adults?  
 67: M. no, three hundred and twenty plus twenty  
 69: M. by the end of the winter  
 70: M. three hundred and forty mouse ...mice ... mices.  
 OK.  
 73: M. Now we need to make a graph of it  
 ...  
 182: M. so let's see ... the first season is over here  
 (making a mark on the graph)  
 183: L. xxxxxx wait a minute  
 186: M. and then sixty plus is going to be a hundred  
 189: L. wait a minute its forty (gestures a triangular  
 shape) OK its forty right?  
 190: L. and then you have to pair those up (brings hands  
 together) and then they have kids (spreads hands apart,  
 while K and M look at her confused)  
 192: M. oh yeaah (embarrassed, laughing at himself)  
 we were doing it ...  
 194: L. That's a lot of mice  
 195: K. gosh that's a lot of nasty mice

The interchange on lines 65/66 is an example of the frequent need to coordinate numbers with their semantics—adults still have to be included in the population, and “three hundred and twenty” is the number of babies in eight seasons just calculated. Similarly line 69 is a further reiteration of the semantics of the number “three hundred and twenty plus twenty”—the number represents a population at a time. Line 73 is an appeal to the authority of the worksheet for what has to be done next. What is interesting about this introduction of a new representation (the graph) is that it appears to be what triggers the new thinking that reveals the error (adopting the linear model) that they have all made. M. makes a mark of sixty on the Y axis at the origin representing the starting population. But L has realised that something is wrong (line 183). M continues calculating the next graph point. But L persists. She starts by reiterating the number and asking for acknowledgement of it (line 189). The number is the number of first season babies. She then states that these have to be paired up, and themselves reproduce (line 190). The gesture is intuitively an important part of her communication that she has a new insight, both for herself and for the group. M fairly rapidly sees their mistake too. They all realise that this is going to make the growth of the population much more rapid though they don't have any number for it. They immediately refer back to the experiential world of ‘nasty mice’. Perhaps the reality of reproduction lies behind the affective tone of the incident. It wasn't just a mathematical mistake, but a failure to apply the ‘facts of life’?

The original adoption of the linear model arises within the ‘mathematical world’. It is, in some sense, the obvious calculation to do—forty babies a season for eight seasons is going to give 320 babies. After all, multiplication is something we learn so as to avoid having to do multiple additions. It is not until the graphing activity makes them break this calculation down into a series of calculations that L sees the error. She thinks about what happens in the world of mice—about the semantics. Her insight is adopted rather rapidly.

### An attempt at creative construction—‘discovering function composition’

When the group brushes up against function composition, they are constructing one of their early models of a population. They have a worksheet entitled *Building the Birthrate* which gives them a procedure for calculating, or recording from reference sources, the various parameters of the situation (brood size for different ages, birthrate, survival rate). Parts of this worksheet and the computer interface are condensed into Figure 1.

The worksheet has its own sequence of activities, though it should be noted that this is not the sequence in which this group performs them. The worksheet's (see Figure 1) first table implements the calculation of the total population births in a season from data from reference sources. At Step 2 the percentage *survival rate* is entered from a reference source and, at Step 3, applied to the total from the table to give a number surviving. The lines represent page breaks in the work booklet. Step 4 then converts the total surviving fry into a percent *birthrate* for the computer. The relevant part of the interface appears next. The bottom table of the figure (over the page on the worksheet) keeps track of the model, and will hold several trial models later on.

The group's sequence of work is actually to start by fulfilling steps 1, 2 and 3, followed by entering the result into the computer and recording the model. Step 4 is circumvented initially and is only filled in retrospectively the next day.

The incident opens with M proposing to take a shortcut in the calculation. This is at first taken by L to be a mistake. She requests an explanation and receives one that she finds satisfying. However, she appears to appreciate that there is consequent bookkeeping which needs to be taken care of, but fails to deflect the group from continuing to the entry of data into the computer model.

444: M. hey wait wait wait ... no but listen. If 4% of the fry survive why don't we just forget about the fry survival and just put that amount for the, for how much are born ...

445: L. because the number born are not how much survived

446: M. yes. yes, the ones who survive are the ones we count, not the ones who are dead because we don't make room for the ones that are dead

453: M. OK you know how 4% the whoooooo fry who were born survive so why don't we just put 4% on the guppies birth because that's how many are going to survive

454: L. I get what you're saying because why put however many more guppies in when they're just going to die anyway?

455: K. so why not just put 4% because that's how many are surviving/ that's how many we're going to count

497: L. but what's that 4% ?

498: K. the ones that survive

499: M. The ones that actually survive fryhood

501: L. Yeah, I know, but how many of the guppies are 4% ?

502: M. we don't know, we'll let that mechanical thing

## Building the Birthrate

**Step 1**

age	# males	# females	# fry	total
young	2	1	4	4
mature	4	2	50	100
old	0	1	0	0
total	6	4		104

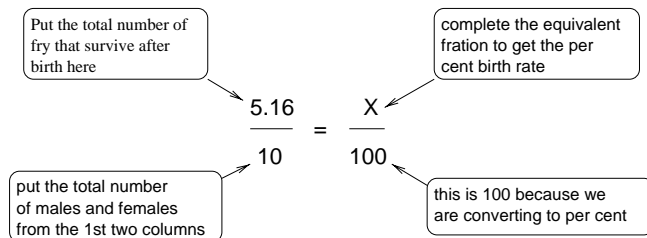
**Step 2** What percent of fry born survive? What happens to the ones who don't make it?

*5% of fry survive. They are eaten*

**Step 3** Use this survival percent and the total number born to calculate the number that survive.

*5.16*

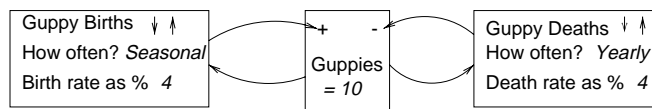
**Step 4** So what's the birthrate? Now that you have calculated an assumed number of fry that survive past birth, you need to convert this into something that Habitech can use as a birth rate. As you know, Habitech works with percents or constant numbers. You will be using a percent birth rate.



BASED ON YOUR ASSUMPTIONS YOUR BIRTHRATE IS

Congratulations! Now take this birth rate and the death rate you will use and head to Habitech to make your model. Remember this birth rate is based on certain assumptions. If you change an assumption, it will affect your model.

**Step 5** Entering numbers into the Habitech interface:



### Recording of Models

Initial #	Birth rate %	Death rate %	Years	Descr.
10	4%	4%	2 year	< 13

Figure 1: Parts of the worksheet and computer interface. The numbers in the tables, equation and the italicised answers were entered by the students

work and tell us

At 444, M opens with a proposal to collapse two stages of calculation into one. In fact, this proposal is perhaps something akin to what is embodied implicitly in the worksheet, and is potentially a creative proposal embodying a concept rather close to one of the core aims of this curriculum—the understanding of mathematical functions. M is proposing to compose two functions into a single function taking the argument of the first and the value of the last. L objects to this proposal and justifies her objection by pointing out that ‘the number born are not how much survived’. In fact we will see that in the terminology of the worksheet, the number of fry surviving expressed as a percentage of the whole population *is* the birthrate, which plays its part in this confusion. M appears to understand the objection and explains his proposal's departure from the worksheet with some success. L accepts the sense of the innovation even though she expresses reservations about its coordination with the worksheet. The activity is turned over to the superior calculating powers of ‘that mechanical thing’—the computer program Habitech.

Unfortunately, the ‘mechanical thing’ does not understand the creative proposal—L's reservations are well motivated, but, lacking a clear understanding herself, her intervention does not deflect the group (see Greeno, Sommerfeld & Weibe (2000) for further analysis). There are numerous problems of coordination between the representations in Figure 1. The survival rate of 5% at Step 2 gets copied into the model table as 4% (possibly a memory error, or a correction later). But the serious error is in shortcutting the calculation at Step 4 and entering the 4% rate directly into the birthrate box at the end. The algebraic ratio part at Step 4 is returned to only later next day when trying to comply with having the whole thing filled in.

What has gone wrong as the group struggles with the welter of representations and numbers? It is hard to give a crisp interpretation of a murky confusion, but we can suggest some of the contributing factors. An important source may be a divergence of the ordering of biological events and the calculation events that refer to them; another is the terminology. In the fish world, fry are born, and then the vast majority are eaten, and then at the end of the season they are counted. In the calculation world, first the number of births are calculated; then a survival rate is applied; and a census number of surviving fry results. So far so good. But turning the page after Step 3, and after recording model parameters on the next page, the students arrive at a further calculation of the ‘birthrate’, where ‘birthrate’ now means ‘birth-and-survival-to-year's-end rate’.

So, at Step 1, the birthrate is a set of numbers representing the brood size of the average guppy at different ages (namely the numerals 4, 50, 0); at Step 2, the birthrate is the number (namely the numeral 104) of fry born to the whole population. In steps 3 and 4 *birthrate* is the birth and survival to end of season rate expressed as a percentage of the whole population (namely the numeral 4). The same idea, a very tangible idea, is represented each time by a number, but each time the number counts different kinds of thing, and complex calculations constitute the inference rules which ‘move the number from box to box’.

Unfortunately, M's insight that two functions can be composed requires attendant housekeeping to keep the ontology straight. Perhaps a contributing factor is that because the pre-survival birthrate in Step 1 is never put into the form of a percentage (1040%), M does not appreciate that, after Step 3, it already has been implicitly composed with the survival rate, and the calculation at Step 4 is intended only to get back to a percentage form. The terminology unfortunately exacerbates this problem of 'backward causality'—first calculating a survival rate (using births) and then calculating a birthrate from that figure.

### Post-test—the persistence of a confusion

We now present an incident from the post-test in which the group displays evidence that the episode of confusion just described has not been fully resolved. Although in the intervening couple of weeks the group has made good progress in understanding population models, as is illustrated in Hall (submitted to this volume), it is of some concern that the particular confusions surrounding the derivation of birthrate from raw data appear to persist.

The group is working on the post-test problem of constructing a model for a mouse population preyed on by cats. This episode is from fairly early on when they are settling on a birth rate for mice and have not yet considered predation:

- 76: M. four, five or six? per adult?  
77: K. If we're going to go four, five or six, let's go four.  
78: L. actually lets use five. Its four through six. Let's use five.  
82: M. OK how do we find out the birthrate? (grabs a piece of paper) We do the ... five is what we decided on. How many did we start out with (looks at the computer)  
83: L. Twenty  
86: M. I'm not sure that this is right (as he writes  $5/20 = X/10$ )  
87: M. What's 500 divided by twenty?  
88: L. What are you doing?  
89: M. Finding out the birth rate  
90: L. Oh yeah.  
91: M. What's 500 divided by 20? (K hands him the calculator and M starts punching in numbers)  
92: M. 25% I could have figured that out myself (K laughs; M goes back to the computer) 25% right? (enters it into the birthrate) and how many die?

Segment 82 illustrates the pervasive struggle with the semantics of numbers. M accepts that they will use 5 (babies per litter per season) which one might think *is* a birthrate, but in this context, 'birthrate' is a specific number that can be entered into certain boxes on worksheet and computer screen. The birthrate, in this sense, they correctly appreciate they do not have, and this is precisely where they had problems before. The number they seek is a percentage. At 87, M has implicitly multiplied the 5 by 100 and is now explicitly going about dividing by twenty (the number in the initial population). L not surprisingly doesn't understand where the 500 came from and asks for clarification, but receives only the description at the completely unhelpful level "finding out the birthrate". The problem is then accepted as a calculation problem, and the semantics is left unaccessed. Why should

the number of babies in one litter divided by the total number adults in the population multiplied by 100 yield a percentage birthrate? The answer would appear to be that the based on some dim memory of a ratio formula (Step 4, Figure 1).

The group is content to continue to the next stage of the problem and does not question the reasonableness of the figure of 25%. This is testimony to the insulation of the numbers from what they mean. If each couple has 5 babies, the actual number is 250%. But the group do not discuss finding this number or acknowledge that adults have to be paired up. The group does not even apply the qualitative reasoning that since the parents are outnumbered by their babies, the birthrate must be more than 100%. Such qualitative inferences are only available if the numbers are treated as standing for something other than themselves—numbers. Even when the model actually turns out to extinguish the mice in short order, the problem is not traced to the low birthrate. It is all too easy for a problem to hide in a complex model. The whole point of models is that many parameters contribute to their outcome. But this means that there are many possible culprits when the outcome is unacceptable.

### Discussion

Nothing by way of inferences about the causalities or even correlations between the kinds of events observed here can emerge from an analysis of these few isolated examples. Nor is redesign of a curriculum usefully based on analyses of single incidents. It is clear from other studies of this curriculum that it is highly effective. Indeed, this very group of students shows a considerable mastery of modelling at the post-test phase. The group repeatedly alters parameters of complex models (including not just birth and survival rates but also predation) in the qualitatively correct direction in response to over- or under-shooting of the desired population outcome.

But we believe that these analyses do make clear just what a sea of semantic complexities the group swims in. They are awash with numbers, and those numbers have to travel from one representational system to another to achieve the problem solving task at hand. As they travel, they change their meanings and their names, and their values. Birthrate is rarely the same thing on two occurrences. The whole system cannot be understood as anything other than heterogeneous, and the interpretations as anything other than highly local. If we were to go through the transcript spelling out after each occurrence of a numeral, the type of the entities it enumerates, we would wind up with some splendid and totally incomprehensible sentences. Nor are numerals the only problem. Simply spelling everything out is *not* to be recommended other than as a way of exposing complexity. But we cannot understand the students' problems until this complexity is exposed.

From a theoretical perspective, this may seem either banal or outrageous. Once we are fluent at the skills of transformation required for coordinating the sub-systems of representation, the whole system appears to take on a transparency and homogeneity which is completely illusory. We cease to notice how the very same number means something quite different from occurrence to occurrence, as do many of the other words. We therefore can either forget that the system is heterogeneous (and respond with outrage to the claim), or we can, as theoreticians, claim that there is nothing deep in the



coordinations that are required (and respond with a yawn).

The students do not have the luxury of mastery. For example, one of the banal consequences of the instability of the meanings of the numerals is that there is a huge memory load as evidenced in the repeated mis-recalls of numbers from sheet to sheet of their workings. We do not believe that there is any way out of the heterogeneity. Learning mastery of the coordination of representation systems is a requirement of learning mathematics and science (and probably most other things). But what we can strive to do is to educate both teachers and students into the quirks of the representational furniture they find themselves surrounded by.

Our research experience in classrooms indicates that teachers are rather wary of taking an explicitly metalinguistic stance. They do not often point out the dangers of shifts in meaning of words during an argument. The critical thinking lecturer warns students about equivocation—the same term being used with different meanings in different occurrences in an argument—but only at college. Prevarication is treated as a fallacy, usually assumed to be eradicable, and therefore is perhaps thought to be eliminable from well-kept classrooms. Our analysis in terms of heterogeneity and localness of interpretation strongly suggests that prevarication is not eliminable. We cannot use unique terms for every meaning, and should not if we could. The use of the same term is often essential to anchor the term to the shared concept as the details shift through its various guises. Perhaps signalling when this is likely to be a problem would help? And perhaps teaching teachers to detect the seams that have become transparent for them between systems is an important aim?

But these observations from the classroom are just as important for theories of the semantics of representations. The conventional response to the kind of observations of language we have made here is that everyone knows that natural language is ambiguous. It is easy to acknowledge heterogeneity if a system contains language and diagrams—here the heterogeneity is on the surface. But the idea that natural language consists of many heterogeneous sub-systems is generally resisted, and explained away as polysemy at the lexical level. There are at least two problems with this explanation. The number of polysemous readings required is essentially infinite, and the meaning of one word is systematically related to that of others. Words in these discourses do not function atomistically—they are part of subsystems. If ‘birthrate’ is construed one way, then its contrasting terms such as ‘deathrate’ and ‘survivalrate’ will also be construed in related ways—at least until there is a shift to a different subsystem. Recently, (e.g., Moravcsik, 1998) theories of lexical meaning have paid more attention to the considerable distance between the generalities of the lexicon and the details of contextualised language use. These stratified theories are much more conducive to understanding real language use and the heterogeneous nature of most reasoning.

In learning to get from a real world problem into a formalism, and back out from the formal results of calculation to an implication about the real world, students must cross many experiential worlds and, when working in groups, negotiate complex patterns of authority for knowledge which determine what the group actually does. In these tapes, we again and again see transitions between the world of numbers and the

world of fish and mice. At one point the discourse is entirely numerical and insulated from the real world consequences, as witnessed by the acceptance of completely implausible values. At others, there is a sudden jumping out of the mathematical world to references to the death of a pet fish, or gee, that’s a lot of nasty meeces! Although formalisms distance proceedings from affective states, when we reason about the world, our reasoning should still be animated by affect. We will not understand conceptual learning until we can give an account of how representations, the social arrangements for authority in discourse, and our experiential worlds are all coordinated.

## Acknowledgements

We would like to acknowledge invaluable comment and discussion with Randi Engel, Muffie Weibe, Jim Greeno and Rogers Hall. We are grateful for their generous support in allowing access to their data gathered under an NSF grant to Hall. We also acknowledge fellowship support from grant GR #R000271074 from the Economic and Social Science Research Council (UK), and CSLI’s support for the first author’s research at Stanford.

## References

- Barwise, J. & Etchemendy, J. (1994). *Hyperproof*. CSLI Publications, Stanford.
- Greeno, Sommerfeld & Weibe (2000). Practices of questioning and explaining in learning to model. Proceedings of the 22nd Meeting of the Cognitive Science Society, Philadelphia.
- Hall, R. (2000). Working the interface between representing and represented worlds in middle school math design projects. Proceedings of the 22nd Meeting of the Cognitive Science Society, Philadelphia.
- Hall, Rogers (1999). Case studies of math at work: Exploring design-oriented mathematical practices in school and work settings. Final report to the National Science Foundation.
- Moravcsik, J. M. (1998). *Meaning, creativity, and the partial inscrutability of the human mind*. CSLI Publications: Stanford, California.
- Oberlander J., Monaghan, P., Cox R., Stenning K. & Tobin, R. (1999). Unnatural language discourse: an empirical study of multimodal proof styles. *Journal of Logic, Language and Information*, **8**, 363–384.
- Stenning, K. Cox, R. & Oberlander, J. (1995). Contrasting the cognitive effects of graphical and sentential logic teaching: reasoning, representation and individual differences. *Language and Cognitive Processes*, **10** (3/4), 333–354.
- van Someren, M., Reimann, P., Boshuizen, E. & de Jong, T. (1999). (eds.) *Learning with multiple representations*. Kluwer.

# Inducing hybrid models of task learning from visuo-motor data

Devika Subramanian  
devika@cs.rice.edu

Department of Computer Science, Rice University,  
6100 Main St MS 132, Houston TX 77005

## Abstract

We develop a new hybrid model of human learning on the NRL Navigation Task (Gordon et. al. 1994). Unlike our previous efforts (Gordon & Subramanian, 1997) in which our model was crafted from verbal protocols and eyetracker data, we demonstrate the feasibility of using visuo-motor data (time series of sensor-action pairs) gathered during training to construct models of a subject's strategy. The goal of our cognitive modeling is to provide a sufficiently detailed description of the subject's strategic misconceptions in real-time, in order to tailor a personalized, task training protocol. Using a small-parameter hybrid model that can be estimated directly and efficiently from the visuo-motor data, we study the deviation of the subject's action choices from that dictated by a near-optimal policy for the task. This model gives us a clear description of the subject's current strategy relative to the near-optimal policy, thus directly suggesting performance hints to the subject. We also provide evidence that our model parameters are sufficient to account for individual differences in learning performance.

## Introduction

Our goal is to build computational models of humans learning to perform complex visuo-motor tasks. By a model of human learning, we mean an explicit representation of the human's action policies (mapping from the perceptual inputs to motor actions) and its evolution over time. The models will be used in designing personalized training protocols to help humans achieve high levels of competence on these tasks. This intended use places constraints on the class of models we can consider and the methods for evaluating them. In particular, the models need to be detailed enough to pinpoint problems in a subject's learning; yet be coarse enough to be unambiguously built from the available visuo-motor learning data. Our criterion for evaluating models is empirical: (i) they must accurately identify incorrect aspects of the subject's strategy, and (ii) when used in place of the human, they must yield comparable performance.

A major challenge in this endeavour is the fact that the visuo-motor data are at an extremely low

level. One approach to modeling in such a situation is to start with a cognitive architecture, and then to find parameter settings for that architecture which recreate the available low-level data. This tactic is adopted by Newell in UTC, Anderson in ACT\* and in EPIC by Kieras and Meyer. We take an alternative approach here based on behavioral cloning (Sammut et. al., 1998). In our approach, the low-level visuo-motor data is taken as the ground truth, and using ideas from machine learning and data mining we "compress" the data in the form of a policy which maps sensors to actions. If there are high level regularities at the policy level in the learning data, they will be reliably extracted by our learning algorithms. This approach has the advantage that cognitive modeling constructs arise endogenously from the data, rather than being stipulated *a priori*.

Our task domain is the NRL Navigation task (Gordon, et al., 1994) developed by Alan Schultz at the Naval Research Laboratory (NRL). It requires piloting an underwater vehicle through a field of mines guided by a small suite of sonar, range, bearing and fuel sensors. Sensor information is presented via an instrument panel that is updated in real-time. The sensors are noisy. Decisions about motion of the vehicle (speed and turn) are communicated via a joystick interface. The task objective is to rendezvous with a stationary target before exhausting fuel and without hitting the mines. The mines may be stationary or drifting. A trial or episode begins with the vehicle being randomly placed on one side of a mine field and ends with one of three possible outcomes: the vehicle reaches the target, hits a mine, or exhausts its fuel. Reinforcement, in the form of a scalar reward dependent on the outcome, is received at the end of each episode. Since the mine configurations vary from episode to episode, it is fruitless for subjects to memorize a

sequence of actions that will get the vehicle to the target. To solve the task, subjects must learn a policy for choosing actions based on the sensor values presented to them.

The Navigation task belongs to the family of partially observable Markov decision processes. With the addition of the last action taken, we can transform it into a fully observable Markov decision process (MDP). This transformation lends theoretical tractability because deterministic optimal decision procedures exist for MDPs. However, the size of the state space is about  $10^{18}$  and there are 153 choices of action at each time step, which make the Navigation task extremely challenging both for humans as well as for present-day learning algorithms like reinforcement learning (Sutton, 1988).

There are four major sources of complexity in the Navigation task from a cognitive perspective: (1) the need for rapid decision making with incomplete information, (2) the sheer number ( $10^{18}$ ) of distinct sensor configurations for which an action choice has to be computed, and the need to learn a partition in the sensor space while acquiring a policy, (3) limited binary feedback at the end of each episode, and, (4) a tightly coupled action space in which the different components (turn and speed) cannot be learned independently. Together, these make the task difficult for our human subjects; one out of every three never acquires the task with our current training protocols.

Our data was gathered as follows. Five subjects ran the Navigation task with a configuration of 60 mines, small mine drift, and low sensor noise.<sup>1</sup> Subjects trained for five days, spending an hour each day running consecutive episodes. The number of episodes per hour varied from around 60 to 160. Each episode varied from 40 to 200 time steps. At the beginning of the first session, subjects were told they had to navigate through a minefield to get to a target location. They were allowed to interact with the task to get comfortable with the use of the joystick. We collected the time series of sensor action pairs as well eyetracker data for the entire training period. We also videotaped the subject and recorded all their verbal utterances. In this paper, we focus on the time series of sensor action pairs to

<sup>1</sup>Five undergraduates at San Diego State University participated in this experiment. The data collection was performed by our collaborators Diana Gordon and Sandra Marshall.

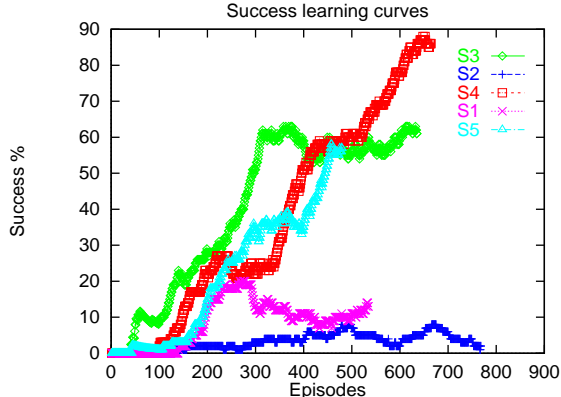


Figure 1: The evolution of success percentages on the Navigation task as a function of training for five subjects.

determine the strategy used by the subject.

In Figure 1 we show the learning curves of the five subjects. Note that the success learning curves are remarkably similar for the three subjects who eventually acquired the task. Subjects go through periods of relatively stable performance, punctuated by substantial improvements. The success curves for the subjects who fail to learn the task are also very similar. This raises hope for building a common computational model for all subjects, with a few parameters to account for individual variations.

The visuomotor performance data for the task is a time series in which each element is of the form: (*episode, timestep, range, bearing, s1, s2, s3, s4, s5, s6, s7, last\_turn, last\_speed turn, speed*). We have over thirty megabytes of visuomotor data for each subject. Extracting the policy used by the subject from this data is difficult for several reasons: (1) the high dimensionality of the data, and the need to find a small number of partitions in the sensor space that meaningfully cluster action choices, (2) noise in the motor data because of joystick hysteresis, (3) data is non-stationary, since the policy adopted by a subject changes with training.

### The approach: comparison against the optimal policy

The key to interpreting visuomotor data is a partitioning of the sensor space into a small number of equivalence classes, each of which is associated with an action choice policy. In this paper, we use the discretization of the sensor space adopted by a near-optimal policy to analyze the distribution of

1. *Part 1: Seek goal: (Sonar in direction of goal is clear)* Follow that sonar at half speed, unless it is the straight ahead sonar, then travel at full speed.
2. *Part 2: Avoid mine/gap finder: (Sonar not in direction of goal is clear)* Turn in place in the direction of the first clear sonar counted from the middle outward.
3. *Part 3: Avoid mine/gap finder: (No clear sonar)* If the last turn was nonzero, turn again in the same direction by that amount, else initiate a turn by summing the sonars to the left and right, and turning in the direction of the lower sum.

Table 1: The three-part near-optimal policy for the NRL Navigation Task. The italicised conditions for each part represent the equivalence class of sensor values that define the part.

actions chosen by our subjects. This approach allows us to determine the deviations of the subject’s strategy from that of the near-optimal policy, which can then be the basis of directed training. A potential disadvantage of the approach is that if there are other near-optimal policies that adopt very different discretizations, a subject using them would be misdiagnosed as making strategic errors<sup>2</sup>. We now describe the near-optimal policy that we discovered, and then present results of modeling the subject’s strategy viewed through its sensor space discretization.

### A near-optimal policy for the Navigation task

A near-optimal policy for the task is deterministic and is shown in Table 1. It must be emphasized that *discovering this solution was not easy!*. It took several months of work with a machine learning algorithm to arrive at this policy.

The near-optimal policy in Table 1 succeeds at least 99.7% of the time; its performance has not been matched by our best human subjects. There are three key properties of the near-optimal policy.

1. *task decomposition*: the policy decomposes the overall goal into the subgoals of avoid-mine and

<sup>2</sup>However, we were unable to determine other near-optimal policies for the NRL task after months of computation and investigation.

seek-goal, a decomposition which appears universal among our human subjects. However, the solutions to the sub-goals are tightly coupled and this is difficult for humans to learn.

2. *dependence between turn and speed choices*: Turning at zero (or close to zero) speeds is essential for success on this task. In addition, turning consistently in one direction while trying to find gaps in the minefield, is crucial.
3. *appropriate discretizations*: the near-optimal policy discretizes the sonar values that range from 0 to 220 into a binary distinction of clear/blocked with the threshold set at 50. The bearing sensor with 12 values is discretized into six, and the range sensor is ignored. The action space is discretized too: the turn action with 17 values is discretized into nine values, and speed with 9 values is discretized into three (zero, half speed, full speed).

The near-optimal policy partitions the state space into three mutually exclusive and collectively exhaustive components. The effective number of states considered by Parts 1 and 2 of the policy is  $2^7 * 6$  which is 768. This is because both parts consider the values of seven sonars, each of which is discretized into clear and blocked, and six values for bearing. The 768 states are really equivalence classes over  $\approx 10^{14}$  base states in the original sensor space. Part 3 examines the previous turn, and thus deals with an effective state space of size  $9 * 27$  which is 243.

### Model extraction algorithm

For ease of presentation, we first describe the model extraction method under the assumption that the visuo-motor sequence data represents a stationary process. This assumption will be relaxed at the end of this subsection. Using the discretizations and definitions of three parts of the near-optimal policy, we classify each sensor-action pair in the visuo-motor sequence as belonging to Part 1, Part 2 or Part 3 equivalence classes. For example, if the sonar in the direction of the goal is clear in the sensor vector, the sensor action pair is classified as a Part 1 pair.

Since the action decisions in Part 1 (resp. Part 2) of the near-optimal policy depend only on the current values of the discretized bearing, we estimate the conditional probability that the subject

chooses a particular discretized<sup>3</sup> action (turn and speed) given the value of the discretized bearing. For discretized action  $a$  in the set  $A$ , and discretized bearing  $b$  let  $n_{ab}$  be the number of times  $a$  is taken by the subject in a Part 1 sensor action pair with bearing  $b$ .

$$P(a|b) = \frac{n_{ab}}{\sum_{c \in A} n_{cb}}$$

The action selection scheme adopted by the near optimal policy for Part 3 sensor equivalence class is inherently sequential. Therefore, to fit Part 3 behavior, we use hidden Markov models (HMMs) (Rabiner, 1989). We identify sequences of sensor-action pairs that belong to Part 3 and train a three state left-to-right HMM on the data<sup>4</sup>.

The parametric hybrid model that we construct from the subject data is shown in Figure 2. Note that the model reflects the task structure. In particular, we use conditional action probability distributions to extract subject behavior on the seek-target subgoal of the task, and a combination of a conditional action probability distribution and an HMM to describe the solution of the coupled subgoal of avoid-mine. This model has few relatively few parameters and can be easily estimated online. It describes the subject’s policy viewed through the equivalence class filter imposed by the near-optimal policy. By comparing the subject’s model for the three parts against that of the near-optimal policy, we can read off strategic errors in the subject’s policy. Examples of such comparisons are offered in the next section.

To accommodate the non-stationary visuo-motor data sequence, we identify stationary subsequences from which the conditional probabilities are estimated and the HMMs are trained. We estimate conditional probability distributions for Part 1 and Part 2 and HMMs for Part 3 over small contiguous blocks<sup>5</sup> of episodes in the data sequence. We then use a standard measure of distance between

<sup>3</sup>The original action set has cardinality 153; the discretized set has nine turns and three speeds making a total of 27 actions.

<sup>4</sup>We experimented with a number of hidden states ranging from 2 to 10, and using log-likelihoods on a left-out test set, we determined that three was the best choice for number of hidden states.

<sup>5</sup>The size of the blocks is determined empirically, and we respect day boundaries in the construction of the blocks.

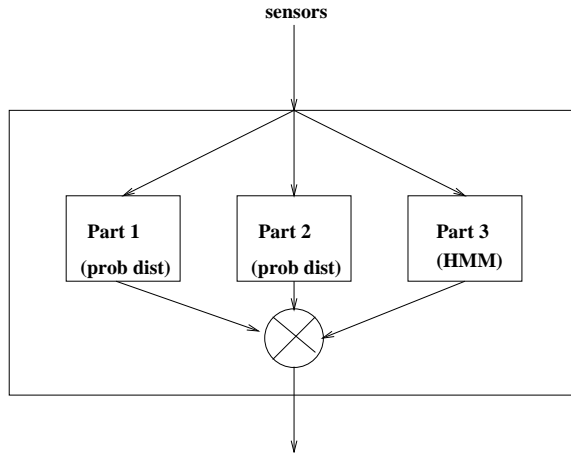


Figure 2: The structure of our hybrid model for the Navigation task.

distributions (KL-divergence<sup>6</sup>) to determine when a significant shift in the Part 1 and Part 2 distributions have occurred. For the HMMs for Part 3, we use KL divergence between both the transition probabilities and the output probabilities to determine when a significant shift has occurred. This procedure identifies points in the sequence that correspond to significant differences in the action selection distributions. These shift points are supported by verbal protocol data as well as eyetracker data. The sequences between shift points are taken as stationary, and the model extraction procedure described above is applied to them.

We now turn to the presentation of experimental results from the use of our model extraction technique on the visuo-motor data corpus for the NRL Navigation task.

## Modeling Results

Examination of the conditional probability distributions of part 1 and 2 and HMMs of part 3 from the three successful subjects reveals that they learn the following.

1. to follow the as-the-crow-flies strategy in the direction of the goal in states in Part 1.
2. to slow down significantly when turning in Parts 1, 2 and 3.
3. to turn minimally to avoid mines in states in Part 2.

<sup>6</sup> $KLdiv(p, q) = \sum_{s \in S} p * \log(p/q)$ , where  $p$  and  $q$  are discrete distributions defined over a set  $S$ .

action	day 1	day 2	day 3	day 4	day 5
$t = 0, s = 0$	0.334	0.370	0.192	0.090	0.078
$t < 0, s = 0$	0.104	0.083	0.106	0.052	0.031
$t > 0, s = 0$	0.083	0.075	0.081	0.021	0.035
$t = 0, s > 0$	0.408	0.454	0.552	0.695	0.646
$t < 0, s > 0$	0.042	0.005	0.015	0.052	0.081
$t > 0, s > 0$	0.028	0.014	0.053	0.090	0.129
KLdiv	3.528	4.220	2.894	2.369	2.011

Table 2: The evolution of the conditional action probability distribution for Subject 4 in Part 1 when bearing = 11 o'clock. The turn  $t$  and speed  $s$  choices are discretized into six categories for reading ease. Turns greater than zero are left turns, and turns less than zero are right turns. For a full explanation of this table, please see the text below.

- to turn in place consistently to find gaps in the minefield in Part 3.

We demonstrate the first point above with data from Part 1 for Subject 4. For this subject, shifts in Part 1 distributions correspond to day boundaries, so we present the evolution of his action selection policy for each day of training. Table 2 presents the conditional probability of Subject 4 taking an action  $a$ , given that the bearing (goal direction) is 11 o'clock. That is, the target lies slightly to the left of the current heading of the vehicle. The near-optimal policy dictates a mild turn to the left. The KL divergence between the subject's policy and the near-optimal policy is shown in the last row of the table. Note that the subject's policy initially diverges and then approaches the near-optimal policy between day 2 and day 3. Also note the rapid decline in the probability of pausing (turn and speed both equal to zero) as training proceeds, with the most dramatic reductions occurring between day 2 and day 3 and day 3 and day 4. The probability that the subject chooses a left turn goes down from day 1 to day 2, but then steadily increases from day 3 forward. All action probabilities except for straight ahead ( $t = 0, s > 0$ ) and left turn ( $t > 0, s > 0$ ) rapidly decay to zero, indicating that the subject is learning to follow bearing well in the Part 1 equivalence class.

It should be emphasized that while Part 1, Part 2 and Part 3 models for each subject co-evolve, they do not evolve at the same rate, and rarely do significant shifts in these probability models coincide. While Part 1 distributions evolve rather slowly and shifts in them occur aligned with day boundaries;

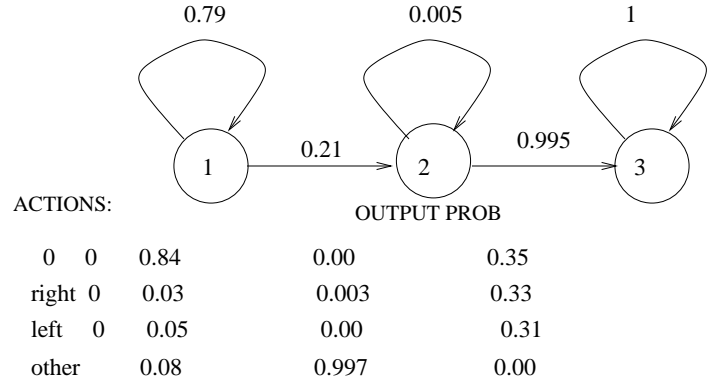


Figure 3: A hidden Markov model that generates and explains the behavior of Subject 5 in states where all sonars are blocked, day 2, episodes 45-67.

Part 3 HMMs evolve much more quickly. For example, for Subject 5, the Part 3 HMM we acquired on data from episodes 45-67 of day 2, differs significantly from the one learned from episodes 68-90 of day 2. These two HMMs are shown in Figures 3 and 4. The first HMM in Figure 3 is a mathematical description of the following strategy: pause (speed = 0 and turn = 0) for a while, and then make an average of two moves with non-zero speed and turn, and finally settle into oscillating back and forth between pauses, left and right turns at zero speed until time runs out. Note that the probability of left and right turns in the terminal hidden state 3 are about the same. In Figure 4, the HMM encodes the following very different strategy: pause for a while, make a left turn at zero speed, and then settle into an action pattern with a consistent preference for turning to the right at zero speed. That is, the subject no longer oscillates back and forth when hemmed in by mines, she sweeps them from left to right trying to find a gap between the mines. This behavior is fairly close to the near-optimal policy for Part 3. In fact, with practice we can get her to spend less time in the state labeled 1, completely eliminate state 2, and in state 3, we can zero out her tendency to pause and increase her probability to turn right. This analysis forms the basis for designing lessons to help the subject acquire greater competence at the task.

How good a fit to performance does the model in Figure 2 provide? The results on Subject 5 for day 2, for episodes 45-67 and episodes 68-90 are shown in Table 3. Note that although the magnitudes pro-

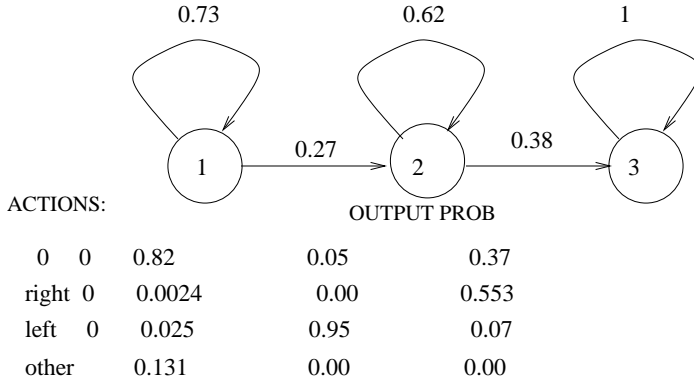


Figure 4: A hidden Markov model that generates and explains the behavior of Subject 5 in states where all sonars are blocked, day 2, episodes 68-90.

D 2, ep 45-67	Succ	Exp	Timeouts	Total
Subject 5	0	12	11	23
Model	0	17	6	23
D 2, ep 68-90	Succ	Exp	Timeouts	Total
Subject 5	0	2	13	15
Model	0	4	11	15

Table 3: The behavioural fit of the new hybrid model to Subject 5, day 2, episodes 45-90.

duced by the model only coarsely approximate those produced by the subject, the trends are captured. For example, both model and subject increase the number of timeouts and reduce the number of their explosions. To get better fits to the performance data, we are currently experimenting with distributions for Parts 1 and 2 conditioned additionally on the previous action.

## Conclusions and Related Work

Our work builds on several distinct pieces of work in the cognitive science as well as the machine learning community. The use of probabilistic models in generating hints for performance improvement is considered by (VanLehn, et. al., 1998). Our work uses a mixture of probabilistic models (conditional action distributions and HMMs) instead of Bayesian networks, and our models are automatically learned from visulmotor data. While the structure of the model is obtained from task analysis (Fredericksen and White, 1989), the parameters are learned by sampling the visulmotor data corpus. The idea of behavior cloning introduced by (Sammut et. al.,

1998) underlies our approach, however the specific techniques for partitioning and learning from non-stationary data are different and novel.

In sum, we have developed a new hybrid model for the NRL Navigation task and presented methods for automatically learning it from low level visulmotor data. The model succinctly represents the deviation of the subject’s policy from a near optimal policy, and allows directed design of new training instances. The model is expressive enough to capture individual differences in strategy. Our current work is to provide closer behavioral fits to the visulmotor data by using richer probabilistic representations.

## Acknowledgements

We sincerely thank our collaborators Diana Gordon of the Naval Research Lab and Sandy Marshall of San Diego State University, and our sponsors Helen Gigley and Susan Chipman of the Office of Naval Research for making this research possible.

## References

- Fredericksen, J. and White, B. (1989). An Approach to Training Based on Principled Task Decomposition. *Acta Psychologica*, 71:89-146.
- Gordon, D., Schultz, A., Grefenstette, J., Ballas, J., & Perez, M. (1994). *User’s guide to the navigation and collision avoidance task* (AIC-94-013). Washington, D.C.: Naval Research Laboratory.
- Gordon, D., & Subramanian, D. (1997). A cognitive model of learning to navigate. *Proceedings of the 19th Annual Conference of the Cognitive Science Society* (pp. 271-276). Lawrence Erlbaum Associates.
- Rabiner, L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *proc. IEEE*, 37(2):257-286.
- Sammut, C. and Harries, M. B. (1998). Extracting hidden context. *Machine Learning*, 32:101-126.
- Sutton, R. (1988). Learning to predict by the method of temporal differences. *Machine Learning*, 3(1):9-44.
- Gertner, A. S, Conati, C. and VanLehn K. (1998). Procedural help in Andes: generating hints using a Bayesian network student model. *Proceedings of the AAAI-98*, AAAI Press.

# Ungrammatical Influences: Evidence for Dynamical Language Processing

Whitney Tabor

tabor@uconnvm.uconn.edu

Bruno Galantucci

Bruno.Galantucci@uconn.edu

Department of Psychology

University of Connecticut

Storrs, CT 06269 USA

## Abstract

A distinguishing feature of self-organizing models of cognitive structure is that they permit incompatible structures to coexist at least temporarily. Here we report on a connectionist model of natural language processing which appears to temporarily construct incoherent structures. We then describe two reading-time studies which reveal people exhibiting the same tendency. In particular, both networks and people show sensitivity to the irrelevant structural interpretations of the underlined phrases in (1) and (2).

(1) We did not think the company would fire truck drivers without consulting the union first.

(2) The manager watched the waiter served pea soup by the trainee.

This kind of sensitivity is absent in parsing models which treat grammatical constraints as absolute because such models lack a principled method of generating incoherent parses. Connectionist networks make the right predictions by using feedback and self-organization. Our results push in the direction of seeking a solution to the tractability problems of parsing by using dynamical mechanisms in a parallel architecture.

## Introduction

Current sentence-processing research tends to focus on ambiguity-related processing in sentences like (1) – (3):

(1) The mechanic maintained the truck was working beautifully.

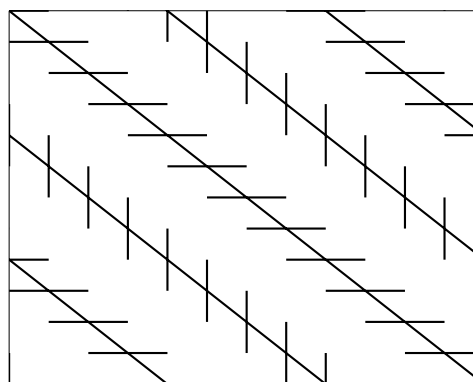
(2) The cop arrested by the detective was chagrined.

(3) The cook stirred the soup with the tomatoes.

Each of these sentences has a structural ambiguity which is resolved on the basis of structural or pragmatic information when the underlined words arrive. Reading time and eye tracking studies show that when biases favor the wrong interpretation initially, readers tend to slow down and/or make regressions in the disambiguating region, which suggests that they either choose the wrong parse initially or are biased toward it (see Frazier, 1988; Tanenhaus & Trueswell, 1995).

Such phenomena accord well with a model of sentence processing which assumes people construct phrase-structure parses incrementally based on the input up to the current point in time. On this view, the slow-down in the disambiguating region is due either to extra time

Figure 1: The Zöllner illusion.



spent on revising an incorrect parse, or to extra time spent on revising the weighting assigned to different possible parses maintained in parallel.

Focusing for a moment on cognitive processes outside of sentence processing, there is a good deal of evidence that people are reliably vulnerable to certain adverse influences when interpreting complex stimuli. In the Zöllner illusion (Held, 1971—Figure 1), lines on a page appear to be nonparallel even though retinal and depth of field information indicate parallelism. Similarly, in the Stroop effect (Stroop, 1966), a decision is supposed to have been made ahead of time to interpret the stimulus along one particular dimension of contrast (e.g. color), and yet when the stimulus is presented, people are often led astray by irrelevant verbal information.

These cases are different from the classic sentence processing examples listed in (1) through (3) in that they show people temporarily failing to rule out an interpretation that could be ruled out absolutely, given the information at hand. What would be the analogous cases in sentence processing?

## Definition of Ungrammatical Influences

There is a class of sentences in which one parse of a word sequence can be completely ruled out on grammatical grounds and yet (we hypothesize) people are influenced by it anyway. The following are examples of such hypothesized “Ungrammatical Influences”:

- (4) a. They won’t fire truck drivers on Sunday.  
b. They won’t hire truck drivers on Sunday.



- (5) a. The manager watched the waiter served pea soup by the trainee.  
 b. The manager watched the waiter given pea soup by the trainee.

Each of the (a) examples has a familiar construction within it that is irrelevant to the only grammatical parse of the sentence. But by the time this distractor construction is encountered, it can be ruled out on grammatical grounds. Our hypothesis is that people are influenced by this “ruled out” parse nevertheless. Thus the (a) examples should be processed differently from the (b) examples which lack the distractors. In (4), the sequence of words “fire truck” forms a familiar compound in English, but coming on the heels of a modal verb, “would”, the word “fire” can only reasonably be interpreted as a verb, not a noun. Similarly, in (5a), the second verb “served” must be interpreted as a passive verb introducing a reduced relative clause which modifies the noun phrase, “the waiter”. But, taken in isolation, “the waiter served pea soup” makes a sensible transitive construction with an active verb.

Our hypothesis is that readers will be distracted by these pockets of coherent structure, even though the structures are incompatible with prior information.

## Models

We find that an often-studied connectionist network, the Simple Recurrent Network (or SRN), behaves in accordance with the hypothesis that Ungrammatical Influences exist. This prediction distinguishes it from most current models of sentence processing.

Elman (1991) showed that a recurrent connectionist network trained by and approximation of backpropagation through time (Rumelhart, Hinton, and Williams, 1986) on word prediction could extract much of the structure of a natural-language-like generating process from a corpus generated by the process.

We trained such a network on the output of Grammar 1 (see Table 1). The network was trained on the task of predicting next words in a constantly growing corpus of strings generated by Grammar 1. The sentences were presented to the network one word at a time. Each input unit corresponded to a possible current word and each output unit corresponded to a possible next word (Elman, 1990, 1991). The learning rate was set to 0.01 throughout and no momentum was used.

The network’s output layer had normalized exponential units. During training, error on a given word was thus defined as the Kullback-Leibler Divergence between the vector of network output activations and the output encoding of the next word that occurred in the corpus (Rumelhart, Durbin, & Chauvin, 1995). We stopped training when the network had successfully distinguished the underlying states of the grammar. At this point, it had seen on average about 500,000 words in sequence.

Since optimal training of such networks causes the output activations to converge on the expected value of the outputs given the inputs, we computed the Kullback-Leibler Divergence between the output activation pattern and grammar-derived probability distributions for

Table 1: A simple phrase structure grammar for generating Noun Noun compounds and Noun/Verb ambiguities.

0.50 S	→ SVP
0.50 S	→ SNP
0.17 SVP	→ to waste N[Obj] is unforgivable
0.17 SVP	→ to bear N[Obj] is necessary
0.17 SVP	→ to mail N[Obj] is costly
0.17 SVP	→ to place N[Obj] is challenging
0.16 SVP	→ to cart N[Obj] is toilsome
0.16 SVP	→ to fuel N[Obj] is ignoble
0.17 SNP	→ the waste baskets are large
0.17 SNP	→ the bear cubs are round
0.17 SNP	→ the mail men are persistent
0.17 SNP	→ the place mats are flat
0.17 SNP	→ the cart wheels are shaky
0.16 SNP	→ the fuel tanks are full
NObj	→ baskets, mats, cubs, wheels, tanks, men

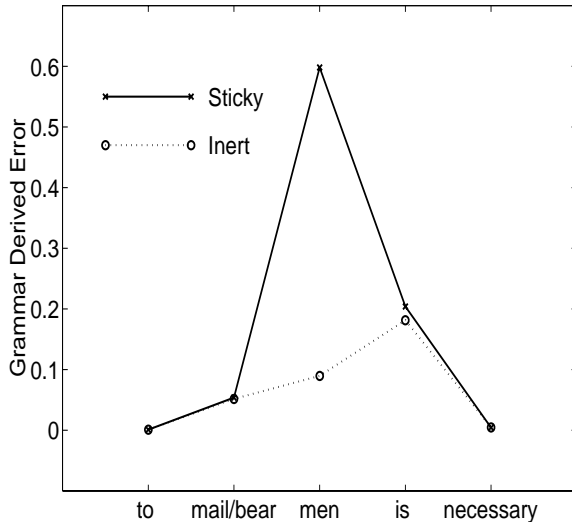
each string of interest. The average Divergences over the six test and control sentences of the form (4) from the grammar are shown in Figure 2.

We repeated the simulation on 10 networks that started learning with different random initial weights. The contrast between the Sticky and Inert conditions occurred in every case. In every case, if we had stopped training earlier (before the network sorted out the differences between states of the underlying grammar), the effect would have been even more pronounced: that is, the network was overwhelmed by the local coherence of the Sticky cases, initially failing to recognize when they occurred in the infinitive context. The effect was somewhat unstable if we trained the network longer on the same materials, and sometimes reversed itself. We suspect that this instability might be reduced if the distractor compounds were not such a prominent feature of the grammar. In real language corpora, coincidences of the Sticky type appear to be quite rare. There are no instances, in the million word Brown Corpus, of coincidental juxtaposition of the 20 sticky pairs used in Experiment 1.

Following Juliano & Tanenhaus (1994), we make an analogy between the network’s error scores and reading times in the self paced reading task (Just, Wooley, & Carpenter, 1982) that is often used to study human sentence processing. The network model thus predicts that readers can be distracted by irrelevant interpretations of pairs of words, and that this distraction will lead to higher reading times on the distracting items.

It appears that the Simple Recurrent Network is prone to be distracted by Ungrammatical Influences. By contrast, standard models of syntactic processing, which assume incremental construction of phrase-structure parses, do not predict such effects, for such models insist on global coherence of each parse they construct. There

Figure 2: Simulation 1: Divergence from grammar-derived expected values. Sticky sentences contain irrelevant Noun-Noun compounds immediately after the main verb. Inert sentences do not.



is one class of hybrid Connectionist-Symbolic models which may, with some modifications, predict Ungrammatical Influence effects: it is the class consisting of the Competitive Attachment Processor (“CAPERS”) of Stevenson (1994) and the Dynamical Unification-Space parser of Vosse and Kempen (1999). These parsers build phrase structure trees by positing variable-strength bonds between nodes in a phrase-marker, and allowing incompatible attachment possibilities to compete with each other under a set of constraints which favor globally coherent structures. Both of these frameworks currently assume that words are brought into the “Unification Space” one at a time, and that some resolution is reached before additional words are incorporated. Thus they do not permit local coherences between successive words to give rise to detached substructures. Nevertheless, it is natural to consider the possibility of allowing them to do so. If one were to permit arbitrary local bonding, then these dynamical structure-building models would probably (modulo the setting of some noise and decay parameters) predict Ungrammatical Influence effects.

What, then, is at stake when we ask the question if Ungrammatical Influences exist? Distinguishing properties of the SRN and the hybrid connectionist models are the use of dynamical (continuously adjusting) feedback and self-organization. These models contrast with chart parsers, pushdown automata and other incremental symbolic parsing systems which maximize the use of constraining information at each point in time. Research on incremental symbolic parsing has strained to grapple with tractability problems associated with the combinatorial growth of parse structures. It seems, at first blush, that opening the door to the inclusion of local coherences, as the Ungrammatical Influences hypothesis

suggests, will only make matters worse. But this impression may be misleading. The coincident emphasis on feedback mechanisms, which allow efficient elimination of incoherent parses through competition, may be just what is needed to permit a parallel processing solution to the tractability problem. Thus, the significance of finding empirical evidence for Ungrammatical Influences is that it would push us in the direction of seeking such a solution.

We turn, now, to empirical investigation of the hypothesis.

## Experiment 1

Tabor and Richardson (1999) compared examples like those in (4a-b) above.

### Method

#### Subjects

Thirty-two undergraduates from Cornell University participated in the experiment. All were native speakers of English. The subjects received course credit for their participation. The experiment lasted for about 30 minutes. The data from one subject was removed from the analysis because of a corrupted file problem.

#### Materials.

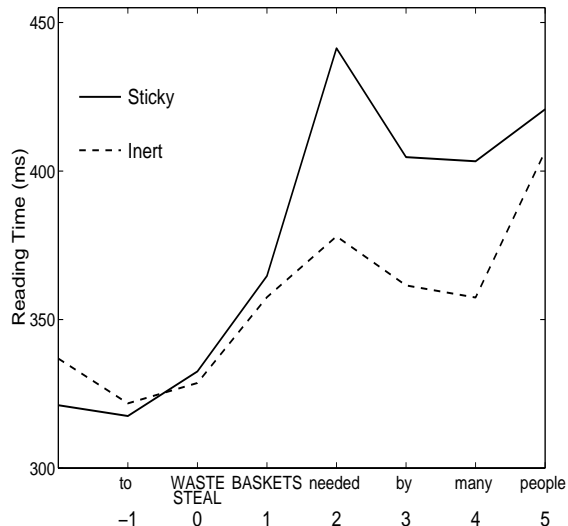
Sixteen target sentences and 16 controls were created. Each target sentence included a clause beginning with a syntactic pattern that strongly constrained the next word to be a verb (e.g., *Some people cannot...* (NP Aux: 7 stimuli), *We decided to...* (NP V[inf] to: 7 stimuli), *on a proposal to...* (P NP to: 1 stimulus), *need a truck to...* (V NP to: 1 stimulus)). This next word (labeled “Word 0” in Figure 3) was lexically ambiguous between a verb sense and a noun sense. In its verb sense, it fit naturally with the preceding and following sentential context, both syntactically and semantically. In its noun sense, this word formed a compound with the word after it (“Word 1”), but this compound did not fit the surrounding context either syntactically or semantically. In 15 of the 16 cases, the compound was a Noun-Noun compound. In one case (“fail safe”) the compound was an Noun-Adjective compound. The control sentences were exactly the same as the target sentences except that Word 0 did not form a familiar compound with Word 1. In 14 out of the 16 controls, Word 0 was ambiguous between a verb sense and a noun sense (the two exceptions were “attend” and “flunk”). This control ambiguity was important for ruling out the possibility that any contrast we might observe between target and control sentences might be due to contrasting ambiguity in Word 0.

#### Procedure.

The sentences were presented using the moving-window self-paced-reading method of Just, Carpenter, and Wooley (1982). Readers read sentences one word at a time, pushing a spacebar to see each successive word. Reading times are measured as intervals between spacebar presses.

The 20 targets and controls were sampled randomly and distributed among 80 filler items. The experiment was preceded by a sequence of six practice trials.

Figure 3: Graph of mean reading time versus position for Experiment 1.



**Results.** All subjects scored better than 80% on the comprehension questions.

We computed the base 10 logarithms of the raw reading times to normalize their distribution. We performed a linear regression with characters-per-word as independent variable and subjects as random factor. The analyses we report below were performed on the standardized residuals from this regression analysis (Trueswell, Tanenhaus, & Garnsey, 1994).

Figure 3 shows average self-paced reading times at word positions -2 through 5. For each region of interest, subject and item means were subjected to separate analyses of variance (ANOVAs), each with a single factor: Stickiness. The means were not significantly different across the two conditions at any word prior to Word 2 or beyond Word 4. The effect of stickiness was significant in both subject and item analyses in the region defined by Words 2, 3, and 4 together ( $F(1, 30) = 10.77, p < .005$ ;  $F(1, 15) = 4.79, p < .05$ ). The stickiness effect was also significant in the subject analysis at Word 2 alone ( $F(1, 30) = 5.82, p < .05$ ), at Word 3 alone ( $F(1, 30) = 8.78, p < 0.01$ ), and at Word 4 alone ( $F(1, 30) = 6.38, p < .05$ ). Stickiness was marginally significant in the item analysis at Word 3 alone ( $F(1, 15) = 4.35, p = .054$ ) and at Word 4 alone ( $F(1, 15) = 3.51, p = .08$ ).

#### Discussion.

These results support the claim that Ungrammatical Influences involving two word sequences exist.

But there is an alternative explanation of the outcome should be considered. An early indication of the existence of Ungrammatical Influences came from a priming experiment on the modularity of the lexicon. Tanenhaus, Leiman, & Seidenberg (1979) found that even the irrelevant meaning of a syntactically ambiguous word (e.g. “rose”) would cause priming for a short interval (< 200ms) after the word was read in a syntactically

constraining context (e.g., “They all rose.”). These results are naturally accounted for in a model that assumes that an activation based lexicon operates partially independently of a phrase-building parser. An ambiguous word activates nodes corresponding to all its senses in the lexicon, and irrelevant nodes are only clamped down when syntactic information is later brought to bear. The results of Experiment 1 may reflect such lexical “automaticity”, since the two-word locally coherent structures are Noun-Noun compounds, which are arguably lexical items (e.g., Mohanan, 1986). Perhaps the parser correctly chooses to treat these sequences as Noun-Verb collocations, but activation of the compound sense in the lexicon creates interference which slows reading down.

Thus Experiment 1 does not decisively demonstrate the existence of Ungrammatical Influences. The next experiment is designed to probe for the existence of Ungrammatical Influences in a case that does not conform to the lexical activation model’s predictions.

## Experiment 2: English clauses

### Experiment 2

The examples in (5a) contain a potentially distracting local coherence in the form of a clause. It is less convincing that clauses are stored as lexical units since they occur in so many combinations and their meanings can generally be computed compositionally.

#### Method.

##### Subjects.

47 subjects were recruited from classes and through advertisement on the campus of the University of Connecticut. All were native speakers of English. They received either money or course credit for their participation. The experiment lasted for about 30 minutes.

##### Materials.

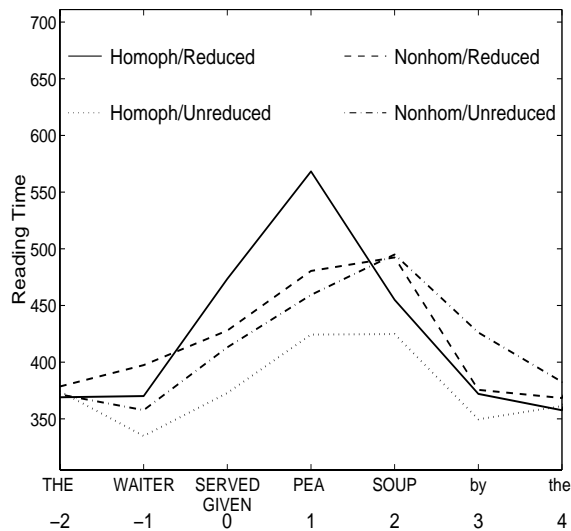
Eighteen experimental items were created. Each item involved four conditions as in (6):

(6) The	manager	watched	the	waiter...			
		0	1	2	3		
a.		served	pea	soup	by...	(R / H)	
b.	who	was	served	pea	soup	by...	(UR / H)
c.			given	pea	soup	by...	(R / NH)
b.	who	was	given	pea	soup	by...	(UR / NH)

Each item included a noun phrase in a non-subject position which was modified by a relative clause in passive voice. Two dimensions of contrast in the relative clause gave rise to four conditions for each item. The relative clause was either reduced (R) or unreduced (UR); its past participle verb was either homophonous and homographic (H) with the corresponding past tense form or distinct from it (NH). Relative clauses like these have been extensively studied in the case where they occur as modifiers of nouns in subject position in a finite clause as in (7) (e.g., Ferreira and Clifton, 1986; Trueswell, Tanenhaus, and Garnsey, 1994).

(7) The waiter served pea soup by the trainee ate ravenously.

Figure 4: Reading times in the four conditions of Experiment 2.



The evidence indicates that when it is semantically sensible to interpret the verb following the subject noun as the main verb of the clause, readers have a strong tendency to do so. Consequently, they become confused starting around the words “by the trainee ate” in a case like (7) because these words disambiguate in favor of the relative clause reading. In a case like (6a), however, the syntax of the words prior to the reduced relative clause precludes the possibility of a main verb reading of the relative clause verb (“served”). If readers were to compute such a reading, then, this would be a case of an Ungrammatical Influence.

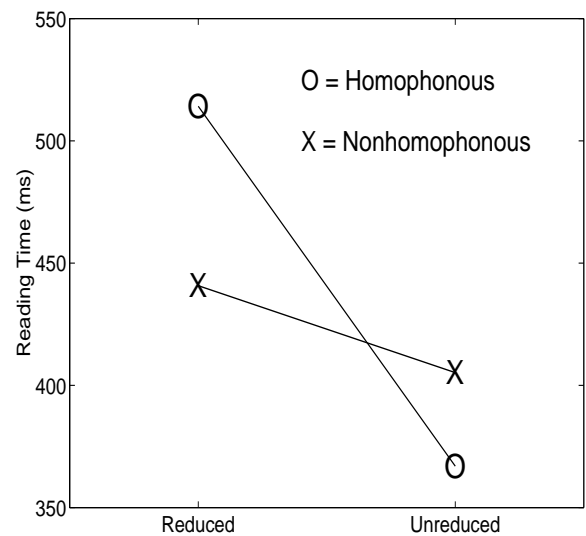
We are looking for an effect of Reduction in the Homophonous case. If this effect obtains and the Unreduced cases are read faster than the Reduced cases, the Ungrammatical Influences hypothesis will not be contradicted. However, it would be premature to take such a result on its own as evidence for the existence of Ungrammatical Influences. Greater speed of processing is expected at the relative clause verb in (b) simply because the syntax is more constraining at this point in case (b) than case (a). That is, it is generally the case that processing speed is faster at grammatical events that are more expected (Jurafsky, 1996; Tabor, Juliano, and Tanenhaus, 1997). Thus, we expect a slowing effect of Reduction in the Nonhomophonous case as well ((d) vs. (c)). For this reason, we have employed the more complex 2 x 2 design. We expect that reduction will slow processing in both cases (a) and (c), but it will slow it more in (a) than in (c). If this interaction occurs, then we will have convincing evidence of the existence of Ungrammatical Influences.

#### Procedure

The procedure was the same as for Experiment 1.

**Results.** All subjects scored better than 80% correct on the comprehension questions and all the data were used in the analysis.

Figure 5: Interaction between Homophony and Reduction in Experiment 2 (Words 0 to 2).



For each region of interest, subject and item means were subjected to separate analyses of variance (ANOVAs), each with two factors: Homophony and Reduction. There was a main effect of Reduction in the region defined by Words 0 to 2 ( $F(1, 46) = 16.83, p < .001$ ;  $F(1, 17) = 7.71, p = .013$ ). There was a main effect of Homophony in the region defined by Words 2 to 3, that was significant in the subject analysis only ( $F(1, 46) = 21.10, p < .001$ ). In both subject and item analyses, there was a significant interaction between Homophony and Reduction over the region defined by words 0 to 2 ( $F(1, 1) = 26.83, p < .001$ ;  $F(1, 1) = 6.99, p = .018$ ). The interaction was also significant at Word 0 ( $F(1, 46) = 12.31, p = .001$ ;  $F(1, 17) = 11.66, p = .004$ ), and significant by subject at Word 1 ( $F(1, 46) = 6.03, p = .018$ ;  $F(1, 17) = 3.85, p = .069$ ) and Word 2 ( $F(1, 46) = 4.25, p = .045$ ). Figure 4 is a graph of reading times for Experiment 2. Figure 5 is graph of the interaction. As Figure 5 indicates, Reduction slowed reading times in both the Homophonous and the Nonhomophonous conditions, but the slowing was significantly greater in the homophonous case.

#### Discussion

The existence of the interaction, with Reduction slowing the Homophonous case more than the Nonhomophonous case, supports the Ungrammatical Influences hypothesis.

There is one aspect of the outcome for which we do not have a clear explanation. The distracting effect of the local structural ambiguity affects reading times earlier in Experiment 2 than in Experiment 1, relative to the locally ambiguous region. We speculate that this difference in timing stems from the fairly unusual syntax of the grammaticality correct interpretation of the Experiment 2 sentences. Reduced relative structures with ditransitive verbs are especially unusual, so readers may be working hard to interpret the sentences in the first place,

and an additional distraction from an Ungrammatical Influence may easily disrupt processing. By contrast, the syntactic structures of Experiment 1 are very common modal+Infinitive or “to”+Infinitive collocations, so readers may not detect the distracting influence until it has had more time to “sink in”. This interpretation again supports a dynamical treatment of information in parsing: some information takes longer to emerge than other information.

### Conclusion

We have focused on the hypothesized phenomenon of Ungrammatical Influences: the syntactic parser is expected to be influenced by local, phrasal coherences that are incompatible with the structure of preceding syntactic material. Two experiments supported the existence of Ungrammatical Influences in parsing. Such effects push the theory of parsing strongly in the direction of dynamical, self-organizing models: Ungrammatical Influences occur because the parser is letting all local coherences among words compete to combine into a maximally coherent structure, rather than deductively eliminating parses based on top-down well-formedness constraints.

Although the present experiments suggest treating Ungrammatical Influences as a kind of interference effect (consistent with the class of Limited Resource models of parsing). Ungrammatical Influences may not always get in the way of parsing. Galantucci, Flores D’Arcais, and Tabor (1999) found that when sentences required people to establish reference for a pronoun, and there was a natural candidate embedded in the internal structure of a compound word (e.g., *The killjoy<sub>i</sub> did not manage to kill it<sub>i</sub> after all.*), processing was facilitated, even though grammatically, the binding is disallowed. These results, combined with the results discussed in this paper suggest that the theory of grammar needs to take up in earnest the problem of incoherent structure representation.

### Acknowledgments

Daniel Richardson was a major contributor to the early stages of this work. Many thanks also to David Perkowski and Kate Finerty who helped with the design and running of the experiments. Helpful comments were provided by Ted Gibson, Neal Pearlmutter and the members of Pearlmutter’s lab group at Northeastern University as well as by the members of the University of Connecticut Linguistics/Psychology lunch talk group.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7: 295–225.

Frazier, L. (1988) Sentence Processing: A Tutorial Review. In M. Coltheart (Ed.), *Attention and Performance* (pp. 559–586). Hillsdale, NJ: Lawrence Erlbaum Associates.

Galantucci, B., Flores d’Arcais, G.B., Tabor, W. (1999). Italian V+N compounds: evidence for syntactic processing. Presentation at the XIX Conference on Neuropsychology, Brixen, Italy.

Held, R. (1974). *Image, object, and illusion, readings from Scientific American*. San Francisco: W. H. Freeman.

Juliano, C. & Tanenhaus, M.K. (1994). A constraint-based lexicalist account of the subject-object attachment preference. *Journal of Psycholinguistic Research*, 23(6): 459–471.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20: 137–194.

Just, M.A., Carpenter, P.A., & Wooley, J.D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111, 228–238.

Mohanan, K.P. (1986). *The theory of lexical phonology*. Boston: D. Reidel Pub. Co.

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In D.E. Rumelhart & Y. Chauvin, eds. *Backpropagation: Theory, Architectures, and Applications*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing, Volume I* (pp. 318–362). MIT Press, Cambridge, Massachusetts.

Stevenson, S. (1994). Competition and recency in a hybrid network model of syntactic disambiguation. *Journal of Psycholinguistic Research*, 23(4):295–322.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18: 643–662.

Tabor, W. & Richardson, D. (1999). Ungrammatical influences in sentence processing. Poster session presented at the 12th Annual CUNY Sentence Processing Conference, New York, NY.

Tanenhaus, M.K., Leiman, J.M., & Seidenberg, M.S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior*, 18:427–440.

Tanenhaus, M. K. & Trueswell, J. C. (1995). Sentence comprehension. In Miller, J. & Eimas, P., (Eds.) *Handbook of Perception and Cognition: Volume 11* (pp. 217–262). Academic Press, San Diego.

Trueswell, J.C., Tanenhaus, M.K. & Garnsey, S.M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33, 285–318.

Vosse, T. & Kempen, G. (1999). Syntactic structure assembly in human parsing. Poster session presented at the 12th Annual CUNY Sentence Processing Conference, New York, NY.

# Mapping the Syntax/Semantics Coastline

Whitney Tabor  
tabor@uconnvm.uconn.edu  
Sean Hutchins  
sonicmoose@aol.com

Department of Psychology  
University of Connecticut  
Storrs, CT 06269 USA

## Abstract

A number of language processing studies indicate that violations of syntactic constraints are processed differently from violations of semantic constraints (Brain imaging: e.g., Ainsworth-Darnell et al., 1998; Ni et al., in press; Speeded grammaticality judgment: McElree & Griffith, 1995; Eye-tracking: Ni et al., 1998). Although these results are often taken as support for the view that the processor employs two separate modules for enforcing the two classes of constraints, we find (in keeping with Rohde & Plaut, 1999, and Tabor & Tanenhaus, 1999) that a nonmodular connectionist network can learn a quantitative distinction between the two types of constraints. But prior connectionist studies have been inexplicit about why the distinction arises. We argue that it stems from the distinct distributional correlates of the different types of information: syntax involves gross distinctions; semantics involves subtle ones. We also describe the Bramble Net, an attractor network which derives grammatical categories and models an approximation of the syntax/semantics distinction in qualitative terms. These results support Elman's (1990) suggestion that grammatical structures may arise by self-organization, rather than by hardwiring. They also help clarify what the grammatical structures are in a self-organizing connectionist network, and emphasize the usefulness of dynamical systems theory in grammatical explanation.

## Introduction

### Definition of syntax vs. semantics

By the distinction between syntax and semantics we mean the fundamental one that Chomsky (1957) identified when he contrasted (1a) with (1b).

- (1) a. Colorless green ideas sleep furiously.  
b. Furiously sleep ideas green colorless.

The modificational relationships between the words in (1b) are not evident to a native English speaker, and one cannot identify any coherent phrasal hierarchy. We thus label (1b) as syntactically anomalous. By contrast, native speakers have no trouble deciding on a parse tree for (1a), but the meanings of the complex phrases are odd and seemingly contradictory. We thus call (1a) semantically anomalous.

By employing some of the basic apparatus of Generative Grammar, we can make a finer characterization of the two types. If we assume that phrases are organized

around grammatical heads which select the semantic attributes of their complements, then (1a) can be diagnosed as an amalgam of *selection violations*. *Subcategory errors* involve incorrect selection of an argument-structure constellation, typically of a verb, (e.g., in \*Ermin put the book). *Agreement errors* involve inconsistencies between elements that are required to share a common feature like number or gender (e.g., \*They eats.) We refer to other mistakes in the sequencing of categories (e.g. \*See dog dog) as *category errors*. The last three types are standardly considered syntactic errors.

### Evidence for the distinction

Drawing a fundamental distinction between syntax and semantics has several advantages.

First, it is only by factoring out the variation in sentence quality due to semantic contrast that it is possible to discern the simple approximation of the range of a language that its phrase structure rules provide (Chomsky, 1957). These rules receive independent justification from the observation that they permit a compositional treatment of meaning that largely accords with human judgment (Frege, 1892).

Second, several recent language processing studies indicate distinct processing responses to syntactic and semantic anomaly. McElree and Griffith (1995) used a speeded grammaticality judgment task to find out how quickly people could detect syntactic and semantic anomalies. They found that detection of syntactic anomaly (both subcategorization violation and category violation) rose above the level of chance about 100 ms. sooner than detection of semantic anomaly (selection violation). Ni et al. (1998) and Braze et al. (submitted) used an eye-tracker to monitor participants as they read sentences that were semantically (selection violation) and syntactically (agreement violation) anomalous. They found that readers slowed down at both kinds of anomalies, but for syntactic anomalies the distribution of their regressive eye movements spiked abruptly on the anomaly itself or shortly after, while for semantic anomalies it was strongly skewed toward the end of the sentence. Ainsworth-Darnell et al (1998), tied together many previous EEG studies by demonstrating independent responses to the two types of anomalies in individual participants. Ni et al. (in press) showed distinct regions of brain response to the two types using fMRI.

## Models

The distinction between syntactic and semantic anomaly seems to be well supported both theoretically and empirically. It is therefore desirable to have a good understanding of how it is instantiated in mental representations. The standard view, coming from Generative Linguistic Theory, assigns separate modules the jobs of checking the two types of anomaly. But this model leaves open the question of how a learner decides whether to attribute an observed distributional systematicity to a syntactic or semantic module. For example, why is “Dogs moo” classified as a semantic anomaly, while “Dogs barks” is a syntactic one?

Connectionist models have exhibited an ability to glean both syntactic and semantic information from text data. Elman (1990, 1991) trained a Simple Recurrent Network or “SRN” on the task of predicting each next word in a simple, English-like corpus. He found that a hierarchical cluster analysis of the trained-network’s hidden unit space contained clusters corresponding to both syntactic classes (Noun, Verb, and various transitivity classes of verbs) and semantic classes (Animate, Large, Edible, etc.). Rohde and Plaut (1999) studied a similar simulation and found that the inclusion of semantic-like lexical cooccurrence biases significantly enhanced the ability of the network to learn complex phrase structures. Moreover, the average lowest transition likelihoods in natural grammatical sentences were higher than the average lowest in grammatical but semantically odd (selection violation) sentences, which in turn were higher than the average lowest in ungrammatical sentences (including verb subcategorization, agreement, and other category sequencing violations). Allen & Seidenberg (in press) used a continuously settling recurrent network and included a bidirectional mapping from form to meaning. The resulting fixed point dynamics provided good generalization behavior.

These results indicate that connectionist networks can derive a distinction between syntactic and semantic structure, while encoding both in a common metric space. But the results raise many questions about what syntactic and semantic structure consist of in such self-organizing models. While, the resemblance of network cluster structures to linguistic categories is suggestive and the alignment of graded network properties with category levels (well-formed, semantically odd, ungrammatical) are encouraging, the findings do not provide much insight into why the resemblances hold or what general properties of the networks produce these results. We performed several additional simulations to better understand how connectionist networks represent syntactic and semantic structure.

### Simulation 1

Following Elman (1991) and Rohde and Plaut (1999), we employed a SRN with three hidden layers, and recurrent connections only in the middle hidden layer. The 30 input units were clamped on or off, one at a time, with each unit uniquely coding the appearance of a particular word. The hidden units (10 in layer 2, 20

Table 1: The grammar for simulation 1. All productions have equal likelihood of being used. The lexical classes expand to between 1 and 4 individual lexical items.

S	→	N[human]	V[eat]	N[food]	p
S	→	N[human]	V[perceive]	N[inanimate]	p
S	→	N[human]	V[destroy]	N[breakable]	p
S	→	N[human]	V[cogitate]		p
S	→	N[human]	V[perceive]	N[human]	p
S	→	N[human]	V[pursue]	N[human]	p
S	→	N[human]	V[move]	N[inanimate]	p
S	→	N[human]	V[move]		p
S	→	N[animate]	V[eat]	N[food]	p
S	→	N[animate]	V[perceive]	N[animate]	p
S	→	N[animate]	V[pursue]	N[animate]	p
S	→	N[animate]	V[act-on]	N[animate]	p
S	→	N[animate]	V[move]	N[inanimate]	p
S	→	N[animate]	V[move]		p
S	→	N[inanimate]	V[move]		p
S	→	N[aggressive]	V[destroy]	N[fragile]	p
S	→	N[aggressive]	V[eat]	N[human]	p
S	→	N[aggressive]	V[eat]	N[animate]	p
S	→	N[aggressive]	V[eat]	N[food]	p

in 3, 10 in 4) had fixed sigmoid activation functions. The target at each point in time was an activation of 1 on the output unit corresponding to the next word in the training sequence. We wanted the outputs to converge on probability distributions over next words, so the output units as a group had the softmax (normalized exponential) activation function. We thus employed the multinomial cost function (Rumelhart et al, 1995) and the delta rule was used to adjust the hidden-to-output weights. The remaining feedforward units were trained using additional backpropagation (Rumelhart, Hinton, & Williams, 1986), and the recurrent connections were trained on the approximation to backpropagation through time (BPTT) in which the gradient is estimated on the basis of only a single previous time step of the hidden units (see Pearlmutter, 1995).

We used probabilistic context free rewrite rules to construct a simple grammar similar to the one used by Elman 1990 for training a syntax network (Table 1). The grammar generated only nouns, verbs, and end-of-sentence markers (“periods”). The verbs were either transitive or intransitive. Both the nouns and verbs fell into a number of semantic classes (See Table 1). We defined a selectional violation to be a sentence in which a verb had the right transitivity, but the noun features were not consistent with the grammar (e.g., N[inanimate] V[eat] N[food]). We defined a subcategorization violation to be a sequence in which a strictly intransitive verb took an object, or a strictly transitive verb did not.

The grammar was used to generate strings of words at random. These were strung together end to end and presented to the network one word at a time. The network was trained with a learning rate of 0.01. Momentum was

Table 2: Means of the grammaticality measure. All within-language comparisons are significant ( $p < .001$ ).

Language	Class	N	Mean	SD
SVO	Well-formed	662	-1.56	0.35
SVO	Sel Viol	2002	-4.18	1.08
SVO	Subcat Viol	1098	-5.21	1.27
SOV	Well-formed	662	-1.60	0.34
SOV	Sel Viol	2002	-5.37	1.66
SOV	Subcat Viol	1098	-6.81	0.82

not used.

The grammar was used to compute exact target distributions for every juncture between words in the training corpus (see Rohde & Plaut, 1999). The Kullback-Leibler divergence ( $E$ ) between the network’s output and the correct distribution was computed at each word in the training corpus ( $E_w = \sum_i t_i \ln t_i/o_i$  where  $t_i$  is the target for unit  $i$  and  $o_i$  is its output on word  $w$ ). Training was stopped when the cumulative divergence error over a large sample of patterns was consistently small enough that we could conclude that the network was not conflating any of the target distributions with one another (approximately 1 million word presentations).

Rohde & Plaut (1999) studied a measure of sentence goodness based on the network’s output predictions. They found that the mean goodness (log of the product of the two lowest output activation transitions) of normal grammatical sentences was higher than that of selection violation sentences, and the selection violation sentences, in turn, had a higher mean than syntactic violation sentences. Because our sentences were much shorter than theirs, we used a simplified version of their goodness measure (log of the single worst transition) and tested it on well-formed sentences, selection violations, and subcategorization violations. We also found a clear stratification (See the “SVO” rows in Table 2).

One of the consequences of defining syntactic category descriptions independently of semantic classifications is that category order is expected to be able to vary independently of the contrast between semantic and syntactic violation. Generative theory thus predicts that the distinction between selection and subcategorization will persevere across languages with different fundamental word orders. To see if the network made a similar separation, we tested it on the output of a grammar exactly like Grammar 1 except that the order of constituents was systematically Subject (Object) Verb (SOV) rather than Subject Verb (Object) (SVO). Indeed a similar relationship between goodness values obtained in the SOV case (Table 2).

A disadvantage of Rohde and Plaut’s goodness measure is that it does not explicitly characterize the effects on processing of making a low-probability transition. The experiments of Ni et al. (1998) and Braze et al. (submitted) indicate that people react to the anomaly of a sentence at or after the anomalous word or words (in Rohde and Plaut’s terms, after they have made a low-

Table 3: Distances to closest grammatical state. All within-language comparisons are significant ( $p < .001$ ).

Language	Class	N	Dist	SD
SVO	Well-formed	662	0.040	0.029
SVO	Sel Viol	2002	0.176	0.206
SVO	Subcat Viol	1098	0.360	0.266
SOV	Well-formed	159	0.020	0.025
SOV	Sel Viol	1000	0.288	0.329
SOV	Subcat Viol	1000	0.625	0.394

probability transition). We studied the response of the network to anomalies by examining the hidden unit representations. To do this, we presented a long sequence (2000 words) of grammar-generated words to the network and recorded the hidden unit states associated with each word. Tabor et al. (1997) called this kind of sample a *Visitation Set*. We then tested the network on ill-formed sentences by finding the hidden unit location visited following the transition with the lowest output activation over the course of the sentence (the *low-point*). Table 3 shows the mean distance in hidden unit space between the low-point and the nearest point in the Visitation set for samples of selection violation sentences and subcategorization violation sentences. For comparison, a new random sample of grammatical sentences was also tested against the visitation set.

The minimum distance measure parallels Rohde and Plaut’s grammaticality measure, and points to a useful way of characterizing the effect of anomaly on the network: there is a subset of the hidden unit space that the network sticks to during grammatical processing. This subset is approximated by the Visitation Set. Selection violations throw the network off the track somewhat. Syntactic violations throw it off more substantially.

This geometrical contrast between the anomaly types has a simple explanation in terms of the distributional distinction between selection and subcategorization. Subcategorization refers to more abstract classes than selection. Thus more instances of training are involved in the development of subcategorization contrasts than in the development of selection contrasts, and subcategorization distinctions produce larger separations in hidden unit space. Violations are cases where the information provided by the current word clashes with the information provided by the preceding context. The network responds to such clashes by averaging the conflicting signals. In the case of selection violation, this averaging interpolates between nearby structures. In the case of syntactic violation, the averaging interpolates between widely separated, major clusters. As a result, syntactic violations tend to result in greater displacement from familiar territory. We hypothesize that the empirical results of McElree & Griffith (1995), Ni et al. (1998), and Braze et al. (1999), which found syntactic violations more readily detected than semantic, stem from this contrast: wildly divergent states are easier to distinguish from normal states than slightly divergent ones.



## Simulation 2

Samples of geometric relationships in the SRN’s hidden unit space do not make it clear what the network’s total generalization behavior is, nor whether its coverage of a language can match that of symbolic phrase structure rules. Nor do relative distance measures alone explain the eye-tracking and brain-imaging results indicating qualitatively distinct responses to semantic and syntactic anomaly. Our previous work on sentence processing (Tabor et al, 1997; Tabor & Tanenhaus, 1999) suggests that the study of dynamical settling networks can clarify the structural principles underlying connectionist sequence-learning. We designed the Bramble Network (BRN) to explore this hypothesis. The BRN is similar to the simple version of the SRN that has one input layer, one recurrently connected hidden layer, and one output layer. But the BRN has two sets of recurrent connections in the hidden layer. One set, the discrete weights, works like the recurrent connections in the SRN, changing the hidden activations discretely every time a new word is read. The other set, the continuous weights, undergoes continuous settling according to Equation (1).

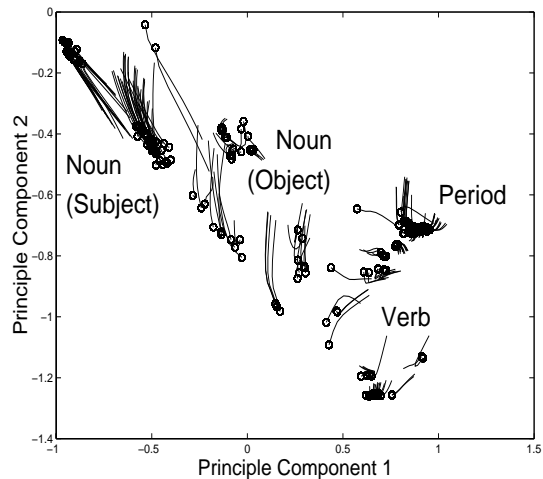
$$\frac{dv_i}{dt} = net_i - v_i \quad (1)$$

where  $v_i$  = unit state,  $net_i = b_i + \sum_j w_{ij}\sigma(v_j)$ ,  $b_i$  = unit bias,  $w_{ij}$  = weight from  $j$  to  $i$ , and  $\sigma(x) = \tanh(x)$ .

In the BRN, the input and context units are updated first. Then the input-to-hidden weights and the discrete hidden-to-hidden weights are used to compute an initial state of the hidden units. Continuous settling is carried out via the continuous weights among the hidden units. Finally, the hidden-to-output weights map the final state of the hidden units to the output.

The discrete weights in the BRN are updated just as in the SRN. We also assume that settling only occurs for brief periods of time (1 cycle) before the discrete weights are updated. This makes it easier for the network to discover dependencies across words. The continuous weights are updated according to a principle of stability maximization. That is, for continuous weights, we define the error on unit  $i$  as  $E_i = (dv_i/dt)^2$  so that  $dE_i/dw_{ij} = 2\sigma(v_j)(net_i - v_i)$ . This equation says: change the weights in the direction that minimizes the magnitude of recent activation change. Continuous weight learning is applied only when the network has almost converged to a stable state. It thus moves the stable state in the direction of the initial state, causing bifurcations when widely separated initial states are associated with a single attractor. The overall effect is that the attractors of the continuous weights tend to track the centers of masses of clusters defined by the discrete weights (cf. Tabor, Juliano, & Tanenhaus, 1997). We found it most effective to train the network with a mixture of fast (1 cycle) discrete weight training and slow (approximating convergence) continuous weight training. A similar result was produced more quickly when we did all the discrete training first and then followed it with the continuous training. The simulation we report below used this batch technique.

Figure 1: Principal component projection of the visitation set for the Simulation 2 network.



As in Simulation 1, the network was trained on output from Grammar 1. In this case, we trained it directly on the output of the grammar for 200,000 words of discrete training (learning rate = 0.002, momentum = 0.9) and then 120,000 words of continuous training (learning rate = 0.05, no momentum). At this point, both discrete and continuous training had successfully distinguished the states of the grammar.

To gain insight into the organization of the trained BRN’s processing, we saved the trajectories associated with a random sample of 200 words in sequence from Grammar 1. We performed Principal Component Analysis (Jolliffe, 1986) on this set of points in order to make the structure visible. The trajectories are graphed in Figure 1. (The two principal components shown account for 87% of the variance). Note that there are regions corresponding the major lexical classes (Noun, Verb, and Period). There are also discernible subclusters within the lexical classes. These correspond to both syntactic (e.g. Subject versus Object, Transitive vs. Intransitive) and semantic (e.g. Big vs. Small, Edible vs. Inedible) classes as well as some clusters whose determinants have not yet ascertained.

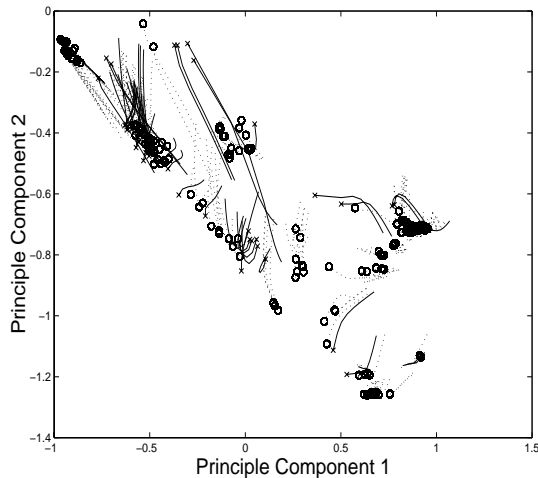
We tested the network on the same sets of good and anomalous sentences that were used in Simulation 1. We defined convergence times for the network by using Euler integration to compute trajectories with  $\Delta t = 0.05$ , and stopping a trajectory when the distance between successive points on the trajectory passed below a threshold (0.005) or when a maximum of 200 steps was reached. The number of steps in the trajectory was taken as a model of reading difficulty. Table 4 shows mean convergence times for several string classes of interest.

When we designed this model, we expected convergence times to provide a good model of human reading times. This prediction is partially sustained in the contrast between normal sentences in their most familiar sequence (71.43) and selection violations (84.52), for much processing evidence supports the claim that readers slow

Table 4: Mean convergence times (MCT) for Simulation 2. All comparisons significant with  $p < .001$  except between selection violations and the sample from all well-formed sentences.

Class	N	MCT
Well-formed (Randomly generated by grammar)	265	71.43
Well-formed (Randomly sampled from list of all well-formed strings)	220	83.69
Sel Viol	250	84.52
Subcat Viol	274	122.85
Syntactic Viol	251	155.13

Figure 2: The trajectories the network follows upon processing selection violations (solid lines) against a background of normal processing (dotted lines).



down when they encounter less familiar sequences (see Jurafsky, 1996). In a loose sense, the model's very high reading times for syntactic anomalies are also consistent with empirical evidence, for Ni et al. (1998) and Braze et al. (submitted) found readers making substantial regressive eye movements at syntactic anomalies, which implies that they take quite a long time to read past the anomalies. However, it is not clear whether the BRN can predict the McElree and Griffith results showing fast detection of syntactic anomalies. It needs to be able to tell quickly when it's not in a familiar attractor basin. We leave this as a question for future work.

Figures 2–4 show a sample of selection violations, subcategorization violations, and category violations (trajectories end on the x's) against the background of normal processing (end on the o's). The sample of anomalous events was generated by picking the longest trajectory in each sentence. These graphs reveal an interesting structure around which the computation is organized. There appears to be a stable connected manifold (con-

Figure 3: The trajectories the network follows upon processing subcategorization violations (solid lines).

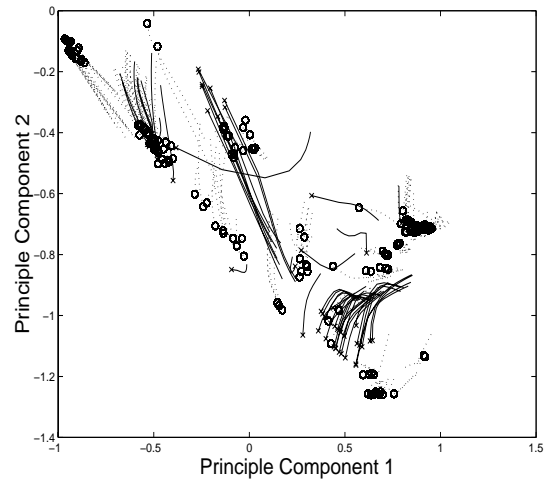
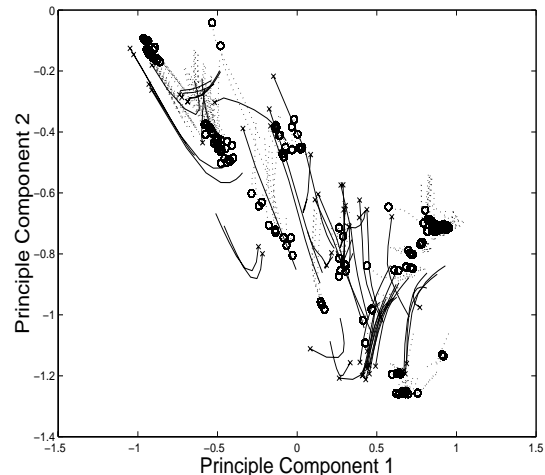


Figure 4: Figure 7. The trajectories the network follows upon processing category violations (solid lines).



tinuous structure that attracts nearby trajectories) running from the upper left of the figure to near the lower right.

There also appear to be pieces of connected manifolds extending to the various other regions where normal processing trajectories end. Perhaps the combination of these manifolds is the locus of grammatical processing. Even semantically anomalous transitions and subcategorization anomalies land by and large on this manifold, though the anomalous cases tend to land on different parts from the normal cases. By contrast, the category violations generally lead to attractors that are separate from the manifold. This suggests that the highly relativistic network model does make a qualitative distinction between types of sentences, and its distinction lines up approximately with current notions of syntactic vs. semantic structure. It is true that the dividing line seems to be different from that of standard linguistic theory, for it is between subcategorization and category error,

rather than between selection and subcategorization error. This difference may stem from a difference between our training grammar and natural language: in natural language, subcategorization constraints are generalizations over more populous classes of items than they are in Grammar 1.

### Conclusions

These graphical results suggest an interesting possibility: the skeleton of a language may be a connected manifold in a dynamical system. Such a finding would be appealing because a connected manifold contains an infinity of points, more than we could ever observe. Thus, identifying such a skeleton could be a way of characterizing one aspect of the unbounded nature of linguistic generalization. Such an insight would be similar to the sort of insight that Generative Theory strives for when it posits a phrase structure or transformational architecture. The trouble with current Generative models, however, is that the steps leading to their creation are very controversial (witness the plethora of current syntactic theories), the data themselves are controversial (note the disagreement about grammaticality judgments), and much of the decision-making that goes into building models of specific parses is not made explicit (note the paucity of implemented parsers that employ modern syntactic theory). The dynamical connectionist approach may be an effective alternative, for it is based on a relatively uncontroversial mathematical theory, it uses performance data rather than competence data and thus does not depend on grammaticality judgments, and the process of choosing a parse is explicit. Moreover, unlike the natural language parsers that have been implemented for practical application, the connectionist theory makes contact with fundamental questions about the principles that underlie linguistic representation.

### Acknowledgments

This work was supported by University of Connecticut Research Foundation Grant # 477138.

- Ainsworth-Darnell, K., Shulman, H. and Boland, J.E. (1998). Dissociating brain responses to syntactic and semantic anomalies: Evidence from event-related potentials. *Journal of Memory and Language*, 38, 112-130.
- Allen, J. & Seidenberg, M.S. (in press). The emergence of grammaticality in connectionist networks. In B. Macwhinney (Ed.), *Emergentist Approaches to Language*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Braze, D., Shankweiler, D., Ni, W., & Palumbo, L.C. (1999). Readers' eye movements distinguish anomalies of form and content. Manuscript, Department of Psychology, University of Connecticut.
- Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton and Co.,
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.

- Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7, 195-225.
- Frege, G. (1892). ber Sinn und Bedeutung. *Zeitschrift fr Philosophie und philosophische Kritik* 100, 25-50.
- Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20, 137-194.
- McElree, B. & Griffith, T. (1995). Syntactic and thematic processing in sentence comprehension: Evidence for a temporal dissociation. *Journal of Experimental Psychology: Language, Memory, and Cognition*, 21(1), 134-157.
- Ni, W., Constable, R.T., Mencl, W.E., Pugh, K.R., Fulbright, R.K., Shaywitz, S.E., Shaywitz, B.A., & Gore, J.C. (in press) An Event-related Neuroimaging Study Distinguishing Form and Content in Sentence Processing. *Journal of Cognitive Neuroscience*.
- Ni, W., Fodor, J.D., Crain, S., & Shankweiler, D. (1998). Anomaly detection: eye movement patterns. *Journal of Psycholinguistic Research*, 27(5), 515-539.
- Pearlmutter, B.A. (1995). Gradient calculations for dynamic recurrent networks: a survey. *IEEE Transactions on Neural Networks*, 6(5), 1212-1228.
- Rohde, D.L.T. & Plaut, D.C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72, 67-109.
- Rumelhart, D.E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In D.E. Rumelhart & Y. Chauvin (Eds.), *Backpropagation: Theory, Architectures, and Applications*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing, Volume I* (pp. 318-362). Cambridge, Massachusetts: MIT Press.
- Tabor, W. & Tanenhaus, M. K. (1999). Dynamical Models of Sentence Processing. *Cognitive Science*, 23(4), 491-515.
- Tabor, W., Juliano, C., and Tanenhaus, M. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes*, 12(2/3), 211-271.

# Word learning as Bayesian inference

Joshua B. Tenenbaum  
Department of Psychology  
Stanford University  
jbt@psych.stanford.edu

Fei Xu  
Department of Psychology  
Northeastern University  
fxu@neu.edu

## Abstract

We apply a computational theory of concept learning based on Bayesian inference (Tenenbaum, 1999) to the problem of learning words from examples. The theory provides a framework for understanding how people can generalize meaningfully from just one or a few positive examples of a novel word, without assuming that words are mutually exclusive or map only onto basic-level categories. We also describe experiments with adults and children designed to evaluate the model.

## Introduction

Learning even the simplest names for object categories presents a difficult inference problem (Quine, 1960). Given a typical example of the word “dog”, e.g. Rover, a black labrador, the possible inferences a learner might make about the extension of “dog” are endless: all (and only) dogs, all mammals, all animals, all labradors, all black labradors, all black things, all running things, this individual animal (Rover), all dogs plus the Lone Ranger’s horse, and so on. Yet, even children under five can often infer the approximate extension of words like “dog” given only a few relevant examples of how they can be used, and no systematic evidence of how words are not to be used (Carey, 1978; Markman, 1989; Regier, 1996). How do they do it?

One influential proposal has been that people come to the task of word learning equipped with strong prior knowledge about the kinds of viable word meanings (Carey, 1978; Clark, 1987; Markman, 1989), allowing them to rule out *a priori* the many logically possible but unnatural extensions of a word. For learning nouns, one of the most basic constraints is the *taxonomic assumption*, that new words refer to taxonomic classes, typically in a tree-structured hierarchy of natural kind categories (Markman, 1989). Given the one example of “dog” above, the taxonomic assumption would rule out the subsets of all black things, all running things, and all dogs plus the Lone Ranger’s horse, but would still leave a great deal of ambiguity as to the appropriate level of generalization in the taxonomic tree that includes labradors, dogs, mammals, animals, and so on. Other, stronger constraints try to reduce this ambiguity, at the cost of dramatically oversimplifying the possible meanings of words. Under the *mutual exclusivity* constraint, the learner assumes that there is only one word that applies to each object (Markman, 1989). This helps to circumvent the problem of learning without negative evidence, by allowing the inference that each positive example of one word is a negative example of every other

word. Having heard Sox called “cat” as well as Rover called “dog”, we can rule out any subset including both Rover and Sox (e.g. mammals, animals) as the extension of “dog”. But some uncertainty in how far to generalize always remains: does “dog” refer to all dogs, all labradors, all black labradors, or just Rover himself?

Inspired by the work of Rosch et al. (1976), Markman (1989) suggested the even stronger assumption that a new word maps not to just any level in a taxonomy, but to an intermediate or *basic* level. Basic-level categories are intermediate nodes in a taxonomic tree that maximize many different indices of category utility and are widely recognized throughout a culture (Rosch et al., 1976). Whether children really have a bias to map words onto basic-level kinds is controversial (Callanan et al., 1994), but it is certainly a plausible proposal. Moreover, the basic-level constraint, together with the taxonomic constraint and mutual exclusivity, actually solves the induction problem, because each object belongs to one and only one basic-level category. However, this solution only works for basic-level words like “dog”, and in fact is counterproductive for all the words that do *not* map to basic level categories. How do we learn all the other words we know at superordinate or subordinate levels? Some experimenters have found that seeing more than one labeled example of a word may help children learn superordinates (Callanan, 1989), but there have been no systematic theoretical explanations for these findings. Regier (1996) describes a neural network learning algorithm capable of learning overlapping words from positive evidence only, using a weakened form of mutual exclusivity that is gradually strengthened over thousands of learning trials. However, this model does not address the phenomenon of “fast mapping” (Carey, 1978) – the meaningful generalizations that people make from just one or a few examples of a novel word – that is arguably the most remarkable feat of human word learning.

To sum up the problem: taking the taxonomic, mutual exclusivity, and basic-level assumptions literally as hard-and-fast constraints would solve the problem of induction for one important class of words, but at the cost of making the rest of language unlearnable. Admitting some kind of softer combination of these constraints seems like a reasonable alternative, but no one has offered a precise account of how these biases should interact with each other and with the observed examples of a novel word, in order to support meaningful generalizations from just one or a few examples. This paper takes some first steps in that direction, by describing one possible learning theory that is up to the task of fast mapping

and applying it to model a simple experimental situation. Our experiments use real, everyday objects with an intuitively clear taxonomic organization, but they require subjects to learn multiple words at different levels of generality which violate the strict versions of mutual exclusivity and the basic-level constraint. Our theory is formulated in terms of Bayesian inference, which allows learners to combine probabilistic versions of *a priori* constraints with the statistical structure of the examples they observe, in order to acquire the sort of rich, multi-leveled vocabulary typical of natural languages.

The paper is organized as follows. Section 2 describes our basic word learning experiment and presents data from adult participants. Section 3 describes the Bayesian learning theory and its application to modeling the data in Section 2. Section 4 concludes and discusses some preliminary data from a parallel experiment with children.

## Experiments with adult learners

Our initial experiments were conducted with adult learners, although the studies have been designed to carry over to preschoolers with minimal modification. The experiment consists of two phases. In the *word learning* phase, participants are given one or more examples of words in a novel language and asked to pick out the other instances that each word applied to, from a large set of test objects. In the *similarity judgment* phase, participants judge the similarity of pairs of the same objects used in the first phase. The average similarity judgments are then submitted to a hierarchical clustering algorithm, in order to reconstruct a representation of the taxonomic hypothesis space that people were drawing on in the word learning phase.

**Participants.** Participants were 25 students from MIT and Stanford University, participating for pay or partial course credit. All participants carried out the word learning task and the first nine also participated in the similarity judgment phase that followed.

**Materials.** The stimulus set consisted of digital color photographs of 45 real objects. This set was structured hierarchically to mirror, in limited form, the structure of natural object taxonomies in the world. Objects were distributed across three different superordinate categories (animals, vegetables, vehicles) and within those, many different basic-level and subordinate categories. The 45 stimuli were divided into a test set of 24 stimuli and a training set of 21 stimuli.

The training stimuli were grouped into 12 *nondisjoint* sets of examples. The first three sets contained one example each: a dalmatian, a green pepper, or a yellow truck, representing the three main branches of the microworld’s taxonomy. The remaining nine sets contained three examples each: one of the three objects from the single-example sets (the dalmatian, green pepper, or yellow truck), along with two new objects that matched the first at either the subordinate, basic, or superordinate level of the taxonomy. For example, the dalmatian was paired with two other dalmatians, with two other dogs (a mutt and a terrier), and with two other animals (a pig and a toucan) to form three of these nine multiple-

example sets.

The test set consisted of objects matching the labeled examples at all levels: subordinate (e.g., other dalmatians), basic (non-dalmatian dogs), and superordinate (non-dog animals), as well as many non-matching objects (vegetables and vehicles). In particular, the test set always contained exactly 2 subordinate matches (e.g. 2 other dalmatians), 2 basic-level matches (labrador, hush-puppy), 4 superordinate matches (cat, bear, seal, bee), and 16 nonmatching objects.

**Procedure.** Stimuli were presented on a computer monitor at normal viewing distance. Participants were told that they were helping a puppet who speaks a different language to pick out the objects he needs. Following a brief familiarization in which participants saw all 24 of the test objects one at a time, the experiment began with the word learning phase. This phase consisted of 32 trials in which learners were shown pictures of one or more labeled examples of a novel monosyllabic word (e.g. “blick”) and were asked to pick out the other “blicks” from the test set of 24 objects by clicking on-screen with the mouse. On the first three trials, participants saw only one example of each new word, while on the next nine trials they saw three examples of each word.<sup>1</sup> Subject to these constraints, the 12 example sets appeared in a pseudo-random order that counterbalanced the order of example content (animal, vegetable, vehicle) and example specificity (subordinate, basic, superordinate) across participants. The frequencies with which each test objects was selected by participants when asked to “pick out the other blicks” were the primary data.

In the similarity judgment phase that followed these trials, participants were shown pairs of objects from the main study and asked to rate their similarity on a scale of 1 to 9. They were instructed to base their ratings on the same aspects of the objects that were important to them in making their choices during the main experiment. Similarity judgments were collected for all but six of the 45 objects used in the word learning experiment; these six were practically identical to six of the included objects and were omitted to save time. Each participant in this phase rated the similarity of all pairs of objects within the same superordinate class and one-third of all possible cross-superordinate pairs chosen pseudo-randomly, for a total of 403 judgments per participant (executed in random order). Similarity ratings for all nine participants were averaged together for analysis.

**Results and discussion.** The results of the word learning phase are depicted in Figure 1. Figure 1a presents data collapsed across all category types (animals, vehicles, and vegetables), while Figures 1b-d show the data for each category individually. The four plots in each row correspond to the four different kinds of example sets (one, three subordinate, three basic, three superordinate), and the four bars in each plot correspond to test objects matching the example(s) at each of four different levels of specificity (subordinate, basic, superordinate, nonmatching). Bar height (between 0 and 1)

<sup>1</sup>The last 20 trials used different stimulus combinations to explore a different question and will not be analyzed here.

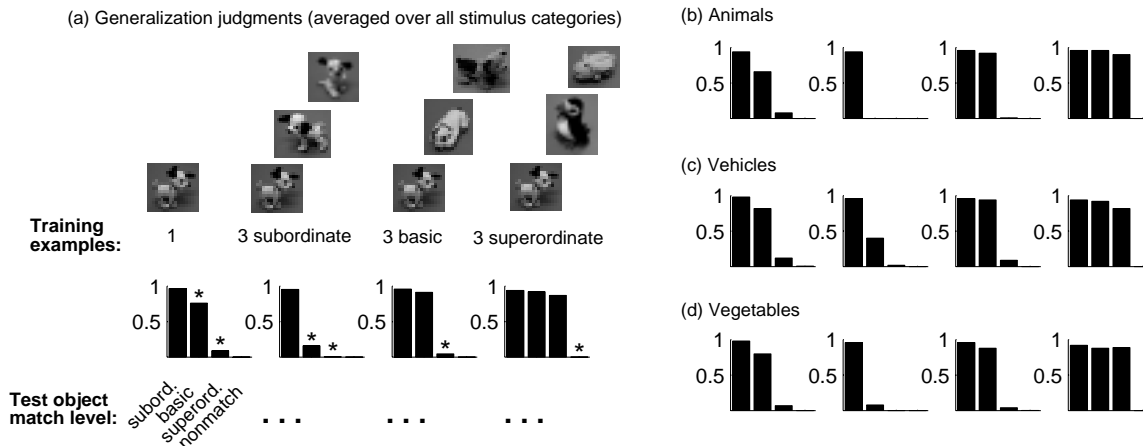


Figure 1: Generalization judgments averaged across categories (a) and broken down into individual categories (b-d).

represents the average probabilities with which participants chose to generalize to the corresponding kind of test object. In Figure 1a, asterisks denote probabilities that are significantly lower than the probabilities to the immediate left ( $p < .05$ , one-tailed paired t-tests ( $df = 24$ ) with Bonferonni correction for 12 comparisons), indicating significant gradients of generalization.

The first plots in each row represent trials in which only a single labeled example was provided. Across all three major categories, participants generalized almost always (97% of trials) to test objects matching the example at the subordinate level (e.g., other dalmatians), often but not always (76% of trials) to basic-level matches (e.g., non-dalmatian dogs), rarely (9% of trials) to superordinate matches (e.g., non-dog animals), and practically never ( $< 1\%$  of trials) to nonmatching test objects (e.g., vegetables or vehicles). Thus, generalization from one example appears to fall off according to a gradient of exemplar similarity, with a threshold located around the basic level.

A different pattern emerges in the last three plots of each row, representing trials on which three labeled examples were provided. Instead of a gradient of generalization decreasing with similarity to the example, there appears in most cases to be a sharp transition from near-perfect generalization to near-zero generalization. The cut-off occurs at the level of the most specific category containing all three labeled examples. That is, given three dalmatians as examples of “blicks”, participants generalized to all and only the other dalmatians; given three dogs, to all and only the dogs, and so on.

Two aspects of these results are consistent with the existing literature on word learning in children. First, we found what appears to be a basic-level bias in generalizing from one example. This interpretation is complicated by the fact that our participants already knew a very familiar word in English for each of the basic-level categories used in our study, “pepper”, “truck”, and “dog”. The tacit knowledge that objects are almost always named spontaneously at the basic level (Rosch et al., 1976) may have increased participants’ propensity to map words in a new language onto these basic-level

categories, and this bias could exist over and above any preference children or adults might have to map words for unfamiliar objects onto basic-level categories. Second, we found that giving participants more than one example had a dramatic effect on how they generalized to new objects, causing them to select all objects at the most specific taxonomic level spanned by the examples and no objects beyond that level. This finding is consistent with developmental studies in which children given two examples from different basic-level categories were significantly more likely to generalize to other objects of the same superordinate category, relative to children given only a single example (Callanan, 1989).

Our results also differ from the developmental literature in important ways. First, we found a qualitative difference in generalization from one labeled example versus several labeled examples. While generalization from a single example decreased according to a gradient of similarity to the test objects, generalization from three examples followed more of an all-or-none, threshold pattern. Second, we found that people could use multiple examples to infer how far to generalize a new word at any level of specificity in a multi-level taxonomy of object kinds, not just at the basic or superordinate levels.

Figure 2 shows the results of a hierarchical clustering (“average linkage”) analysis applied to participants’ similarity judgments from the second phase of the experiment. Each leaf of the tree corresponds to one object used in the word learning phase. (For clarity, only objects in the training set are shown.) Each internal node corresponds to a cluster of stimuli that are on average more similar to each other than to other, nearby stimuli. The height of each node represents the average pairwise dissimilarity of the objects in the corresponding cluster, with lower height indicating greater average similarity. The length of the branch above each node measures how much *more* similar on average are that cluster’s members to each other than to objects in the next nearest cluster, i.e., how distinctive that cluster is.

This cluster tree captures in an objective fashion much of people’s intuitive knowledge about this domain of objects. Each of the main classes underlying the choice of

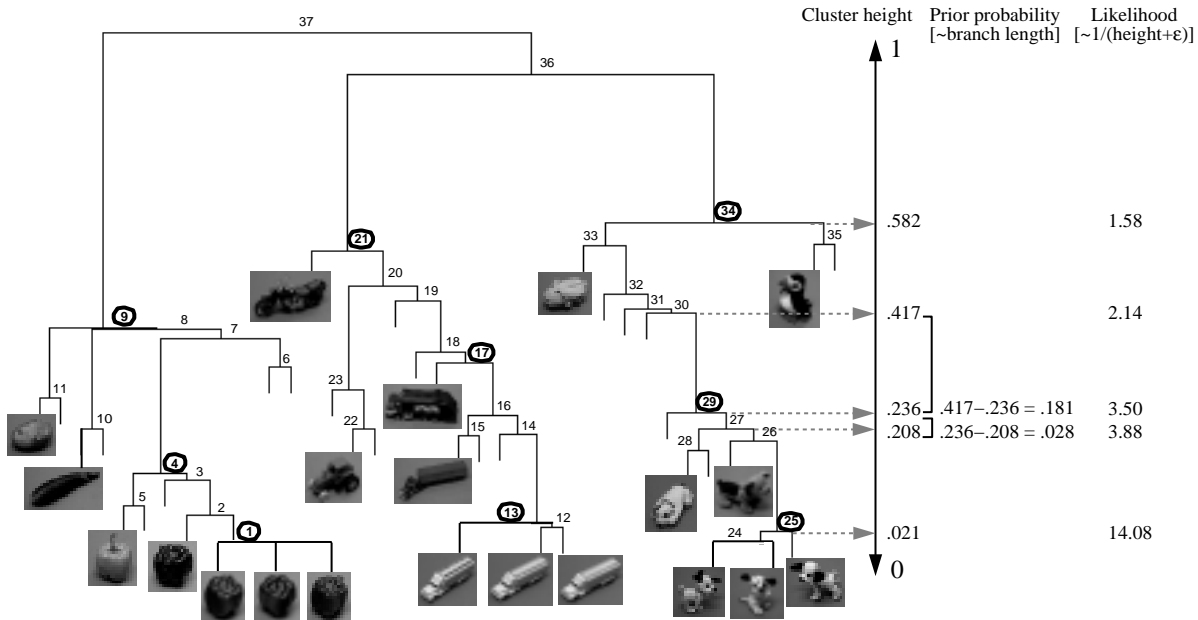


Figure 2: Hierarchical clustering of similarity judgments yields a taxonomic hypothesis space for word learning.

stimuli (vegetable, vehicle, animal, pepper, truck, dog, green pepper, yellow truck, and dalmatian) corresponds to a node in the tree (marked by a circled number). Moreover, most of these clusters are highly distinctive, i.e., well-separated from other clusters by long branches, as one would expect for the targets of kind terms. Other naturally “nameable” nodes include cluster #23, containing the tractor, the bulldozer, and the crane, but no other vehicles, or cluster #33, containing all and only the mammals. Still other clusters reflect more subtle similarities. For example, cluster #18 includes all of the trucks and also the yellow schoolbus. While the schoolbus does not fall into the class of trucks, it intuitively comes much closer than any other non-truck object in the set. This intuitive taxonomy of objects will form the basis for the formal Bayesian model of fast mapping described next.

## A Bayesian model

We first describe the general approach, saving the details for below. We assume that the learner has access to a hypothesis space  $\mathcal{H}$  of possible concepts and a probabilistic model relating hypotheses  $h \in \mathcal{H}$  to data  $X$ . Let  $X = \{x^{(1)}, \dots, x^{(n)}\}$  denote a set of  $n$  observed examples of a novel word  $C$ . Each hypothesis  $h$  can be thought of as a pointer to some subset of objects in the world that is a candidate extension for  $C$ . The Bayesian learner evaluates these hypotheses by computing their *posterior* probabilities  $p(h|X)$ , proportional to a product of *prior* probabilities  $p(h)$  and *likelihoods*  $p(X|h)$ :

$$p(h|X) \propto p(X|h)p(h) \quad (1)$$

The prior, along with the structure of the hypothesis space, embodies the learner’s pre-existing (though not necessarily innate) biases, such as the taxonomic or basic-level assumptions. The likelihood captures the statistical information inherent in the examples. The poste-

rior reflects the learner’s degree of belief that  $h$  is in fact the true extension of  $C$ , given a rational combination of her observations  $X$  with her relevant prior knowledge about possible word meanings.

**The hypothesis space.** Tenenbaum (1999) introduced this Bayesian framework for learning simple concepts with hypotheses that could be represented as rectangular regions in a multidimensional continuous feature space. Here we adapt that framework to the task of word learning, assuming that the hypotheses can be represented as clusters in a tree-structured taxonomy (e.g., Figure 2). Such a hypothesis space is clearly not appropriate for learning all kinds of words, but it may be a good first approximation for learning common nouns under the taxonomic assumption. Assuming a tree-structured hypothesis space makes the model more tractable but is by no means a requirement of the Bayesian framework. In principle, any subset of objects could be a hypothesis under consideration.

**Priors and likelihoods.** Both priors and likelihoods can be defined in terms of the geometry of the cluster tree. The crucial geometrical feature is the height of node  $h$  in the tree, which is scaled to lie between 0 (for the lowest node) and 1 (for the highest node) and measures the average dissimilarity of objects within  $h$ .

We take the prior  $p(h)$  to be proportional to the branch length separating node  $h$  from its parent:

$$p(h) \propto \text{height}(\text{PARENT}(h)) - \text{height}(h). \quad (2)$$

This captures the intuition that more distinctive clusters are *a priori* more likely to have distinguishing names. For example, in Figure 2, the class containing all and only the dogs (#29) is highly distinctive, but the classes immediately under it (#27) or above it (#30) are not nearly as distinctive; accordingly, #29 receives a much higher prior than #27 (proportional to .181 vs. .028).

The likelihood function comes from assuming that

the observed positive examples are sampled at random (and independently) from the true concept to be learned. Imagine that each hypothesis consisted of a finite set of  $K$  objects. Then the likelihood of picking any one object at random from a set of size  $K$  would be  $1/K$ , and for  $n$  objects (sampled with replacement),  $1/K^n$ . Hence set size is crucial for defining likelihood. While we do not have access to the “true” size of the set of all dogs in the world, or all vegetables, we do have access to a psychologically plausible proxy, in the average within-cluster dissimilarity (as measured by cluster height in Figure 2). Moving up in the tree, the average dissimilarity within clusters increases as they become larger. Thus equating node height with approximate cluster size, we have for the likelihood

$$p(X|h) \propto \left[ \frac{1}{\text{height}(h) + \epsilon} \right]^n, \quad (3)$$

if  $x_i \in h$  for all  $i$ , and 0 otherwise. (We add a small constant  $\epsilon > 0$  to  $\text{height}(h)$  to keep the likelihood from going to infinity at the lowest nodes in the tree (with height 0). The exact value of  $\epsilon$  is not critical; we found best results with  $\epsilon = 0.05$ .) Equation 3 embodies the *size principle* for scoring hypotheses: smaller hypotheses assign greater likelihood than do larger hypotheses to the same data, and they assign exponentially greater likelihood as the number of consistent examples increases. This captures the intuition that given a dalmatian as the first example of “blick”, either all dalmatians or all dogs seem to be fairly plausible hypotheses for the word’s extension (with a likelihood ratio of  $14.08/3.50 \approx 4$  in favor of just the dalmatians). However, given three dalmatians as the first three examples of “blick”, the word seems much more likely to refer only to dalmatians than to all dogs (with a likelihood ratio now proportional to  $(14.08/3.50)^3 \approx 65$  in favor of just the dalmatians).

**Generalization.** Given these priors and likelihoods, the posterior  $p(h|X)$  follows directly from Bayes’ rule (Equation 1). Finally, the learner must use  $p(h|X)$  to decide how to generalize the word  $C$  to new, unlabeled objects.  $p(y \in C|X)$ , the probability that some new object  $y$  belongs to the extension of  $C$  given the observations  $X$ , can be computed by averaging the predictions of all hypotheses weighted by their posterior probabilities  $p(h|X)$ :

$$p(y \in C|X) = \sum_{h \in \mathcal{H}} p(y \in C|h)p(h|X). \quad (4)$$

To evaluate Equation 4, note that  $p(y \in C|h)$  is simply 1 if  $y \in h$ , and 0 otherwise.

**Model results.** Figure 3a compares  $p(y \in C|X)$  computed from the Bayesian model with the average generalization data from Figure 1a. The model achieves a reasonable quantitative fit ( $R^2 = .93$ ) and also captures the main qualitative features of the data: a similarity-like gradient of generalization given one example, and more all-or-none, rule-like generalization at the most specific consistent level, given three examples. The main errors seem to be too little generalization to basic-level matches given one example or three subordinate examples, and

too much generalization to superordinate matches given three basic-level examples. All of these errors would be explained if participants in the word learning task had an additional basic-level bias that is not captured in their similarity judgments. Figure 3b shows the fit of the Bayesian model after adding a bias to the prior that favors the three basic-level hypotheses. With this one free parameter, the model now provides an almost perfect fit to the average data ( $R^2 = .98$ ). Figures 3c and 3d illustrate respectively the complementary roles played by the size principle (Equation 3) and hypothesis averaging (Equation 4) in the Bayesian framework. If instead of the size principle we weight all hypotheses strictly by their prior, Bayes reduces to a similarity-like feature matching computation that is much more suited to the generalization gradients observed given one example than to the all-or-none patterns observed after three examples ( $R^2 = .74$  overall). If instead of averaging hypotheses we choose only the most likely one, Bayes essentially reduces to finding the most specific hypothesis consistent with the examples. Here, that is a reasonable strategy after several examples but far too conservative given just one example ( $R^2 = .78$  overall). Similarity-based models of category learning that incorporate selective attention to different stimulus attributes (Kruschke, 1992) could in principle accommodate these results, but not without major modification. These models typically rely on error-driven learning algorithms, which are not capable of learning from just one or a few positive examples and no negative examples, and low-dimensional spatial representations of stimuli, which are not well-suited to representing a broad taxonomy of object kinds.

## Conclusions and future directions

Research on word learning has often pitted rule-based accounts (Clark, 1973) against similarity-based accounts (Jones & Smith, 1993), or rationalist accounts (Bloom, 1998) versus empiricist accounts (Quine, 1960). In contrast, our work suggests both a need and a means to move beyond some of these classic dichotomies, in order to explain how people learn a hierarchical vocabulary of words for object kinds given only a few random positive examples of each word’s referents. Rather than finding signs of exclusively rule- or similarity-based learning, we found more of a transition, from graded generalization after only one example had been observed to all-or-none generalization after three examples had been observed. While special cases of the Bayesian framework corresponding to pure similarity or rule models could accommodate either extremes of this behavior, only the full Bayesian model is capable of modeling the transition from similarity-like to rule-like behavior observed on this task. The Bayesian framework also brings together theoretical constraints on possible word meanings, such as the taxonomic and basic-level biases, with statistical principles more typically associated with the empiricist tradition, such as the size principle and hypothesis averaging. No one of these factors works without the others. Constraints provide sufficient structure in the learner’s hypothesis space and prior probabilities to enable reasonable statistical inferences of word meaning from just



Examples: 1      3 subordinate      3 basic      3 superordinate

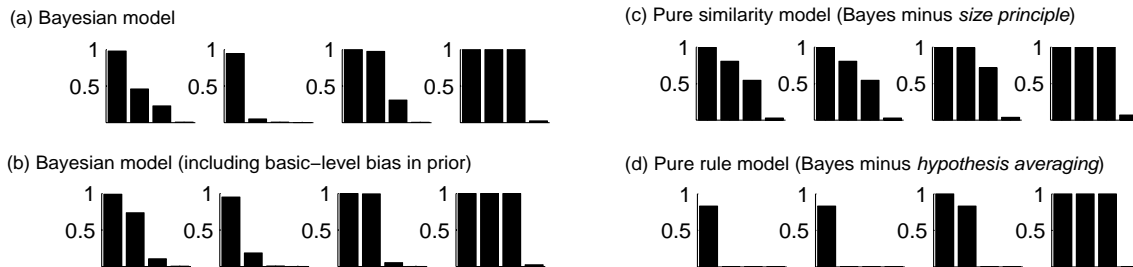


Figure 3: Predictions of the basic Bayesian model and three variants for the data in Figure 1.

a few random positive examples.

Still, the hardest questions of learning remain unsolved. Where does the hypothesis space come from? Are constraints on the hypothesis space learned or innate? In ongoing work, we are exploring how unsupervised learning algorithms might be used to bootstrap a hypothesis space for supervised concept learning. For example, can clustering algorithms like the one we used to construct our taxonomic hypothesis space still be successful when applied to more primitive perceptual representations of objects, instead of adult humans' similarity judgments? Generalizations of the Bayesian framework also hold some promise as bootstrapping mechanisms, in virtue of their ability to propagate probabilistic information from raw data up to increasingly higher levels of abstraction. Perhaps we begin life with a hypothesis space of hypothesis spaces – each embodying different possible constraints on word meanings – and grow into the most useful ones – those which consistently contain the best explanations of the word-to-world pairings we encounter – through the same mechanisms of Bayesian inference used to learn any one novel word.

Examples: 1      3 subordinate      3 basic      3 superordinate

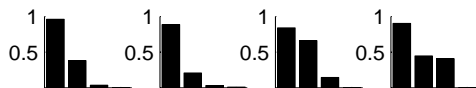


Figure 4: Data from child word learners.

We are also working to extend this line of research to studies of child learners, and to studies of both adults and children learning words for novel objects. Figure 4 shows some promising pilot data from a study with 4-year-old children, using familiar objects in a design approximately parallel to the above adult study. Like the adults, children given three examples of a novel word adapt their generalizations to the appropriate level of specificity, although their superordinate generalizations are less consistent. When given just one example, children show a gradient of generalization much like the adults, but with significantly fewer responses at the basic level and above. If anything, children's overall patterns of responses look more like the Bayesian model's predictions *without* the added basic-level bias (Figure 3a) than *with* that added bias (Figure 3b). Consistent with Callanan et al. (1994), this suggests that a strong basic-level bias may not be a fundamental building block of

early word learning – at least, not as distinct from the more general bias in favor of labeling distinctive clusters that the Bayesian model assumes – but rather develops later as the child gains experience about how words are typically used. This issue is one aspect of a broader question: to what extent should differences between child and adult word learners be attributed to differences in their hypothesis spaces, probability models (e.g., priors), or learning algorithms? We hope to answer these questions as we conduct more extensive studies with child learners.

## References

- Bloom, P. (1998). Theories of word learning: Rationalist alternatives to associationism. In Bhatia, T. K. and Ritchie, W. C. (eds.), *Handbook of Language Acquisition*. Academic Press.
- Carey, S. (1978). The child as word learner. In Halle, M., Bresnan, J., and Miller, G. A. (eds.), *Linguistic Theory and Psychological Reality*. MIT Press.
- Callanan, M. A. (1989). Development of object categories and inclusion relations: Preschoolers' hypotheses about word meanings. *Developmental Psychology*, 25(2):207–216.
- Callanan, M. A., Repp, A. M., McCarthy, M. G., and Latzke, M. A. (1994). Children's hypotheses about word meanings: Is there a basic level constraint? *Journal of Experimental Child Psychology*, 57:108–138.
- Clark, E. V. (1973). What's in a word? On the child's acquisition of semantics in his first language. In Moore, T. E. (ed.), *Cognitive Development and the Acquisition of Language*. Academic Press.
- Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In MacWhinney, B. (ed.), *The 20th Annual Carnegie Symposium on Cognition*. Erlbaum.
- Jones, S. S. and Smith, L. B. (1993). The place of perception in children's concepts. *Cognitive Development*, 8:113–140.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44.
- Markman, E. M. (1989). *Categorization and Naming in Children: Problems of Induction*. MIT Press.
- Quine, W. V. (1960). *Word and Object*. MIT Press.
- Regier, T. (1996). *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. MIT Press.
- Rosch, E., Mervis, C. B., Gray, W., Johnson, D., and Boyes-Braem, P. (1976a). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.
- Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In Kearns, M. J., Solla, S. A., and Cohn, D. A. (eds.), *Advances in Neural Information Processing Systems 11*. MIT Press.

## Aspectual Coercion and the Online Computation of Sentential Aspect

**Marina Todorova** ([todorova@cogsci.jhu.edu](mailto:todorova@cogsci.jhu.edu))

Department of Cognitive Science, JHU; 3400 N. Charles Street  
Baltimore, MD 21218 USA

**Kathy Straub** ([kath@cogsci.jhu.edu](mailto:kath@cogsci.jhu.edu))

Department of Cognitive Science, JHU; 3400 N. Charles Street  
Baltimore, MD 21218 USA

**William Badecker** ([badecker@cogsci.jhu.edu](mailto:badecker@cogsci.jhu.edu))

Department of Cognitive Science, JHU; 3400 N. Charles Street  
Baltimore, MD 21218 USA

**Robert Frank** ([rfrank@cogsci.jhu.edu](mailto:rfrank@cogsci.jhu.edu))

Department of Cognitive Science, JHU; 3400 N. Charles Street  
Baltimore, MD 21218 USA

### Abstract

We investigate the comprehension of sentences where an aspectual incompatibility between a verbal predicate (*send a check*; completive reading) and a verbal modifier (*for years*; durative reading) is resolved through the operation of aspectual coercion. Aspectual coercion modifies the aspectual properties of the predicate in the direction required by the verbal modifier; here the result is an obligatory iterative interpretation for the combined string (*send a check for years*). We find that sentences where the iterative interpretation arises as a result of coercion (*Howard sent a large check to his daughter for years*) show a significant reading delay in the coercion and post-coercion regions as compared to sentences where an iterative interpretation is achieved by other means (*Howard sent large checks to his daughter for years*). Such delay does not occur with substitution of an aspectually neutral modifier (*last year*). We propose that the observed delay is a processing reflex of aspectual coercion deriving either from an initial misanalysis of the aspectual representation of the utterance, or from the need to postulate a null iterative operator in order to arrive at a coherent interpretation of the coerced sentence.

### Aspectual Coercion

This study investigates the processing correlates of aspectual coercion. Aspectual coercion has been proposed in the linguistic and computational literature (e.g. Moens & Steedman, 1988) as an operation that resolves a mismatch between the aspectual properties of the verbal predicate, on one hand, and a (temporal) sentential operator, on the other. In English, the operation of coercion does not have an overt morphological reflex. Therefore, it is generally considered to lack a structural counterpart in the syntax. Instead, the effects of coercion are purely semantic: The verbal predicate obligatorily receives a specific aspectual interpretation, which differs from its most natural (or default) aspectual reading. An inquiry into the processing correlates of aspectual coercion promises to provide valuable information

about the mechanisms of semantic processing in general, and details of computing the aspectual interpretation of an utterance, in particular. This study examines the effects of durative adverbial modifiers on the aspectual interpretation of the predicate. It is hypothesized that coercion triggered by such modifiers is associated with a specific processing cost.

### Aspect

The grammatical category of *aspect* relates to the internal temporal structure of an event. Aspectual distinctions are anchored around the presence or absence of logical boundaries in the denotation of events. For example, the eventuality denoted by the verb ‘find’ seems to contain a logical endpoint, namely the moment at which one becomes aware of the existence of some novel object. It is implausible that the act of finding extends beyond this endpoint; similarly, we cannot say that an event of finding has been instantiated unless this endpoint has been realized. By contrast, the state denoted by the verb ‘love’ can plausibly extend indefinitely in time. This does not mean that loving cannot reach a terminal point; rather, such a point is not a logically necessary component of the verb’s meaning. For the purposes of this paper, we will call those aspectual readings that contain a necessary and/or realized event boundary *telic*; aspectual readings that are open-ended (i.e. indeterminate with respect to an endpoint) will be called *atelic*.<sup>1</sup>

---

<sup>1</sup>Strictly speaking, the labels *telic* and *atelic* are usually applied to the lexical-conceptual structure of events; the properties of sentential utterances are described as *bounded* vs. *unbounded*; *perfective* vs. *imperfective*, etc. We keep only one set of labels for simplicity; however, it is worth pointing out that we consider *telic* at the sentential level interpretations where the logical endpoint of an event is understood to have been instantiated, i.e. roughly the idea described by the traditional notion of perfectivity.

While telicity is encoded in the conceptual structure of events, it can be modified by the larger context within which an event is embedded. Thus we have a distinction between *lexical* and *grammatical* aspect. Lexical aspect categorizes verbs into aspectual classes based on their meaning (Dowty, 1978; Vendler, 1967). The atelic lexical classes encompass *states* and *processes*:<sup>2</sup> Verbs such as *love* or *write* describe temporally unbounded eventualities, or, alternatively, eventualities with homogenous reference: A subpart (subinterval) of a state of loving is still a state of loving, and a subpart of an act of writing is still an act of writing. The telic lexical class is that of *events*: Verbs such as *find* denote eventualities that involve some change or transition between different states of affairs. The transition corresponds to the logical boundary of the event. Events have non-homogenous reference: it is hard to conceive of subparts of the event of finding an object, and if we imagine a situation where the object is found after an active search, then any subparts of this (larger) event are instantiations of searching, rather than finding.

The aspectual reading of a fully articulated utterance is not always transparently related to the lexical aspect of its main verb. Rather, the computation of sentential aspect is influenced by the presence of nominal arguments and temporal sentential modifiers. Within the verbal predicate, the presence of an object and its cardinality have important consequences for the resulting aspectual reading: Singular and/or definite (count) objects support telic interpretations, whereas bare plural and/or mass noun objects support atelic interpretations of the verbal predicate. To illustrate, by itself the verb *'write'* denotes an unbounded process; when combined with a singular object (*write a book*), it denotes a bounded event in the course of which some change/transition is effected (i.e., a new object - a book - comes into existence). This type of predicate is traditionally referred to as an *'accomplishment'*. However, if the same verb is combined with a bare plural object (*write books*), it receives an atelic aspectual reading: The predicate now denotes an iterative or habitual process of book-writing. Similarly, punctual eventive verbs, such as *'send'*, receive an iterative interpretation when combined with a bare plural or mass noun object: The predicate *'send letters'* denotes a process that can potentially repeat itself over an indefinitely long period. Since there is no special instance of iteration that is regarded as the terminal point of the iterative event, iterative interpretations are open-ended. The importance of the cardinality of the object for the aspectual reading of predicates leads some authors to propose that aspectual properties are computed over the verb-argument complex (Verkuyl, 1993).

To illustrate the effect of temporal operators, the English progressive operator alters the aspectual properties of its input into those of an ongoing process (i.e. an atelic eventuality). Consequently, even though the primitive predicate *'send a letter'* is associated with a telic reading, its progressive version, *'I'm sending a letter'*, is an atelic process that does not allow an inference to the completion of the ongoing event. The modification of the aspectual properties of a

verbal predicate by sentential operators is known as aspectual coercion. This paper focuses on the processing cost associated with one particular instance of coercion, which arises in the presence of adverbial material denoting temporal span.

### Adverbial Coercion

A long-standing observation in the aspect literature maintains that adverbs denoting extent in time are sensitive to the aspectual category of the predicate that they modify. Adverbs of duration, such as *'for X time'* or *'throughout'*, combine with atelic predicates: *John wrote letters for an hour*, whereas adverbs of completion, such as *'in X time'*, combine only with telic predicates: *John wrote a letter in one hour*.

However, this generalization is not entirely correct. Adverbs of duration can occur in combination with any aspectual type of predicate; the output of such combination, however, is necessarily interpreted as an atelic eventuality. Thus, even though the primitive *'write a letter'* is a telic accomplishment predicate, its modified counterpart *'write a letter for an hour'* is interpreted as an atelic process of letter-writing that lasted one hour. The absence of telicity in this expression is made evident by the fact that the sentence *'John wrote a letter for an hour'* does not entail that at the end of the hour the letter in question has actually been written. Similarly, the primitive punctual predicate *'send a letter'* can be modified with a durative expression *'for several years'*. In this case, the overall interpretation shifts to the eventuality denoted by the predicate repeating itself over and over (with some pragmatically plausible frequency) for the extent of several years. Thus, it is clearly the case that adverbs of duration act like coercing operators for some predicates. This behavior is not surprising if we assume that the denotation of durative adverbs picks out a temporal interval within which an event unfolds: Since an interval interpretation is necessarily atelic, all input to the durative adverbial must acquire atelic properties. The specific reinterpretation that occurs as a result of combination of a predicate with a durative modifier is still somewhat dependent on the basic properties of the input event. An accomplishment predicate (*write a letter*) contains a simple process within its denotation; therefore, reinterpretation usually amounts to *'stripping'* off its culmination phase and understanding the predicate as an instance of the underlying process that did not reach its endpoint (Moens & Steedman 1988). Punctual events, on the other hand, have neither internal structure nor internal temporal extent. The only way in which they could be forced into an interval interpretation is by introducing a process of repetition of the punctual event. This is what happens in an example such as *John sent a letter to the company for several years*.

To summarize, several factors can potentially contribute to the overall aspectual reading of an utterance: The aspectual class of the main verb, the cardinality of its object, and the input specifications of modifying adverbial material. This situation creates a certain degree of instability within the parsing system, since material encountered later in an utterance can conflict with properties of the semantic representation that have been built up on the basis of material encoded earlier in the utterance. For instance, if the parser is

<sup>2</sup>There exist various classifications of verbal lexical aspect; the one adopted here is due to Mourelatos (1981).

assumed to incrementally compute a telic representation for the *entire utterance* upon encountering a telic verb (or its combination with a singular object), subsequent modification with a durative adverbial should trigger (potentially observable) aspectual reanalysis immediately after the durative modifier is encoded. Alternatively, one might hypothesize that in the absence of overt aspectual markers, such as a progressive or a perfect operator, aspectual commitment is postponed until all relevant material has been encoded. On that view, sentential aspect is left underspecified for the duration of the sentence. Where no further material relating to aspect becomes available, the aspectual reading for the sentence is determined over the properties of the entities that make up the predicate. However, if additional salient entities, such as a temporal modifier, emerge, these are taken into consideration in the initial computation of sentential aspect. Coercion in this model would amount to nothing more than a selection of the appropriate aspectual value based on all the lexical information provided, though we might expect to observe increased sentence ‘wrap-up’ processing time as the correct aspectual properties of the utterance are calculated, especially when factors informing the computation of aspect are in conflict. Further complications can arise if issues of plausibility/frequency are taken into consideration. It could be, for example, that particular verb + object combinations (given the importance of the cardinality of the object) increase the probability that the overall interpretation of the utterance will be of a certain kind (telic or atelic), and lead to an early adoption of the respective aspectual interpretation. For some verbs, one aspectual usage may be more frequent than the other (for example, the eventuality denoted by the verb ‘break’ may be less likely to be represented as an (iterated) process than the eventuality denoted by the verb ‘kick’, especially when it is understood to affect the same unique object). The kinds of aspectual reinterpretations triggered by coercion may also involve varying degrees of reanalysis: reinterpreting an event as a sequence of iterations is representationally different from reinterpreting an event as incomplete. Clearly, we cannot begin to unravel all of these issues at the same time. The present paper concentrates on one specific question: Is there a processing cost associated with the coercion which occurs when verbs denoting punctual events are forced to assume a repetitive interpretation, and if so, how does this inform our understanding of the mechanisms underlying semantic parsing?

### The Psycholinguistics of Aspect

To date, very little research examining the psycholinguistic implications of the coercion process has been reported. In one important exception, Piñango, Zurif and Jackendoff (1999) examined processing costs associated with coercion using a cross-modal lexical decision task. They investigated the effect of interpreting a durative temporal adverbial following a punctual verb (*kick*) vs. a non-punctual verb (*examine*). When the presentation of a lexical decision target coincided with the disambiguation point of the underlined temporal adverbial in coercion contexts like *The man kicked the little bundle of fur for a long time to see if it was alive*, they observed slower responses in comparison to decisions made in the corresponding location for non-coercing contexts

(i.e., *The man examined the little bundle of fur for a long time to see...*). Piñango et al. attribute the longer decision times to the increased processing costs associated with the coercion operation.

Although these results are suggestive, one potential problem undermines their interpretation. The creation of minimal pairs by systematically alternating verbs introduces other interpretive differences to which the observed processing variation might be attributed: Sentences in the “coerced condition” entail an iterative interpretation, unlike most sentences in the “non-coerced condition”. This difficulty suggests that further, more rigorous examination of the on-line implications of the coercion process is indicated.

The experiment presented here expands on Piñango et al.’s strategy of contrasting coercion and non-coercion contexts. We examine the processing correlates of the specific type of semantic coercion which arises from the combination of punctual verbs (e.g. *send*) with a durative adverbial (*for X time*), culminating in an iterative reading of the entire utterance. Since it is unclear whether iterative interpretations are computationally more demanding than non-iterative ones, the critical alternation in the materials that we employ hinges on varying the cardinality of the direct object (as opposed to varying the verb) as the factor which controls the initial repetitive vs. non-repetitive aspectual status of the predicate. It should be recalled that bare plural direct objects impose an iterative reading (*send letters*), whereas singular direct objects impose a single-instance reading of the event denoted by the predicate (*send a letter*). In the case of the bare plurals, the repetitive event interpretation is signaled by the plural object prior to the introduction of the durative adverbial and that adverbial simply specifies the temporal span over which the repetitive event occurs. Thus, the interpretation of bare plurals modified by durative adverbials (*sent letters for many years*) is straightforward, since the (atelic/iterative) aspectual reading of the predicate is consistent with the aspectual input specifications of the modifier.

In contrast, the introduction of a durative adverbial modifier following a singular object (*send a letter for many years*) triggers aspectual conflict between the telic predicate and the durative adverbial. This incompatibility is hypothesized to be resolved via the coercion process, through which the predicate is reanalyzed as an iterative event spanning the specified interval. If the reanalysis process suggested as a correlate of aspectual coercion is computationally demanding, we should expect to observe increased processing load at or subsequent to the introduction of aspectual incompatibility (as seen in sentences containing iterative events over singular objects). Evidence of this cost should be observed when we compare parallel regions of the singular vs. bare plural items, just because the coercion operation is hypothesized to occur only over the predicates containing singular objects.

Further, since aspectual coercion is triggered by a specific type of temporal modifier, we would not expect to see evidence of coercion with adverbs that are indifferent to the aspectual properties of the predicates that they have scope over. This expectation is independent of the cardinality of the direct object of the predicate (*sent a letter last year* vs. *sent letters last year*). Although these sentences come to

mean different things (singular vs. multiple instances of letter-sending, corresponding to telic/perfective and atelic/imperfective aspect, respectively), in both cases the aspectual reading is determined solely on the basis of the properties of the verb + object complex, and depends specifically on the cardinality of the direct object. So, temporal adverbs which are indifferent to aspect are not hypothesized to contribute to the aspectual interpretation of the utterance and should combine easily with any type of aspectual input. Therefore, we would not predict any processing load variation to be observed at, or subsequent to, the introduction of such adverbials despite the diverging interpretations ultimately required by such sentences.

Thus, this experiment examines the processing cost associated with coercion toward an iterative interpretation of a telic verb + singular object predicate triggered by the presence of a durative temporal adverbial. Two separate non-coerced, control conditions are employed. First, processing load for the hypothesized coerced sentences is compared with processing load for sentences in which there is a telic verb + bare plural object predicate followed by a durative adverbial. In this control, the bare plural independently signals an iterative event reading so there is no aspectual conflict between the predicate and the temporal adverbial. Secondly, sentences containing durative adverbials are compared to parallel sentences containing non-durative temporal modifiers. This control should allow us to distinguish any potential effects of coercion, as we have described it here, from effects that might instead arise directly out of the singular vs. bare plural object contrast, independent of the coercion operation.

## Method

**Participants** Twenty-four right-handed, native English-speaking undergraduates with no history of language deficits enrolled at the Johns Hopkins University participated in the experiment for course credit or compensation.

**Materials** Thirty-six transitive aspectual achievement verbs were used to construct the experimental sentences. Each verb was used to create two VP predicates which varied on the cardinality of the direct object (singular indefinite vs. bare plural) so the resulting predicates differed only in iterativity. For each predicate, adverbial modifiers (durative vs. non-durative/aspectually neutral) were selected to allow equally (ultimately) plausible readings in all conditions. Thus, the experiment consists of a 2×2 design crossing **Cardinality** (*singular* vs. *plural*) and **Modifier Type** (*durative* vs. *non-durative*). As can be seen from the stimulus example from Table 1, with the exception of the cardinality of the direct object and the specific temporal adverbial, the lexical content of the sentences was identical across the four conditions.

Table 1: 2×2 experimental design crossing factors of Cardinality and Modifier Type.

	Durative modifier	Non-durative modifier
Singular object	Even though Howard sent <i>a large check</i> to his daughter <u>for many years</u> , she refused to accept his money <b>A</b>	Even though Howard sent <i>a large check</i> to his daughter <u>last year</u> , she refused to accept his money <b>C</b>
Plural object	Even though Howard sent <i>large checks</i> to his daughter <u>for many years</u> , she refused to accept his money <b>B</b>	Even though Howard sent <i>large checks</i> to his daughter <u>last year</u> , she refused to accept his money <b>D</b>

Condition A reflects the hypothesized coerced context: the aspectual properties of the predicate and the modifier are mismatched and we expect that any processing costs associated with the coercion operation should be observed in this condition. In contrast, no effects of coercion should be observed in the other conditions.

Experimental sentences were constructed on a bi-clausal frame, in which the critical adverbial phrase always occurred in the initial clause. Table 2 shows each sentence subdivided into presentation regions (roughly corresponding to phrases), with the temporal adverbial always occupying Region V. With the exception of the critical alternation, the lexical material within particular regions of a given item was identical. Thus, we expect that any processing costs associated with the coercion operation will be observed at or immediately downstream of region V.<sup>3</sup>

Table 2: Sentential frame for experimental sentences with type of material by region.

Region :	II	III	IV	V	VI-IX
I					
Although Because Even though	Subject + verb	Direct object	Preposi- tional phrase	Tem- poral ad- verb	2nd clause

Critical items were distributed into 4 lists such that each list included one token of each of the 36 critical items and nine items from each of the 4 treatment conditions. The 4 sets of experimental stimuli were each embedded into a list of 70 filler sentences. Filler items, which were also subdivided into roughly phrasal presentation regions, ranging from 5 to 9 regions in length, varied in syntactic structure as well as syntactic and semantic complexity. Since the experimental paradigm employed in this study allowed for the collection of sensibility judgment data, 30 of the filler items were designed not to make sense. Nonsense filler items were

<sup>3</sup>There is some evidence that processing of semantic information follows a slower time-course than syntactic processing (Boland 1997). This suggests that a coercion effect is likely to occur later than the actual presentation of coercing material, namely in region VI.

incongruous based on some grammatical violation (e.g., subject/verb agreement), conceptual/pragmatic constraints (e.g., implausible event), or both. Thus, the experiment consisted of 4 separate list conditions containing 106 items each. Individual lists were pseudorandomized for presentation order such that one or more filler items intervened between every pair of target items. Presentation lists orders were randomized independently to avoid item-ordering effects.

**Overview of Task & Procedure** This experiment employed a self-paced, makes-sense judgment task in which participants evaluated sentences presented region-by-region in the center of a computer screen. Participants were instructed to quickly read each region and indicate whether that text region “made sense” with respect to the previously presented material from that trial. Two types of data were recorded for each participant for each text region: reading/judgment times and regional rejection rates. The rate of text presentation was controlled by individual participants in that new text material replaced the previous material as soon as a participant indicated a judgment (via a button press). At the end of each trial, participants were asked to provide make-sense judgments for the entire sentence. Thus, the data collected in this task allows us to examine processing load effects via reading/judgment latencies for specific regions. In addition, by recording sensibility judgments we can examine by-region rejection rates to test our intuitions regarding the aspectual infelicity in the coerced condition. Furthermore, we can confirm that ultimately participants do arrive at a meaningful interpretation in all sentence conditions. Finally, the make-sense judgment task has the added advantage of discouraging fast readers from buffering text material or postponing their interpretations until sentence-final regions are reached. Although no judgment feedback was given on critical trials, participants were encouraged to actively engage in regional make-sense judgments by receiving negative feedback when their make-sense judgments conflicted with those of the experimenters on filler trials.

**Results**

In sentence-final judgments, participants rejected 19% of the sentences in Condition A (the coerced condition), but only 7% in Condition B (the non-coerced, bare plural condition):  $\chi^2 = 14.73$ ,  $df = 1$ ,  $p < .001$ . Rejection rates in Conditions C & D were 8% and 9%, respectively. Sentences which were judged to be nonsensical *overall* were excluded from further analysis. However, items for which participants indicated that one or more regions were nonsensical, but judged the sentence to be acceptable overall were included in the analysis.

**Reading time data** Analyses on the reading/judgment latencies were computed separately using subjects and items as random factors. Analyses of latencies for text regions preceding the temporal adverbial (I-IV) and in regions VII and VIII were not different across treatment conditions ( $F_s < 1$ ). Subsequent analyses focus on differences observed in Regions V (the adverbial modifier) and VI (immediately following the coercion region). In the full analysis evaluating Modifier Type and Cardinality, no main effect of Modifier

Type emerged (all  $p_s > .25$ ), although a main effect trend toward Cardinality emerged (Region V:  $F_1(1, 23) = 3.71$ ;  $p < 0.06$ ;  $F_2(1, 35) = 1.8$ ;  $p < 0.19$ .; Region VI:  $F_1(1, 23) = 4.97$ ;  $p < 0.05$ ;  $F_2(1, 35) = 2.88$ ;  $p < 0.09$ ). The interaction between these two variables was nearly significant at Region V ( $F_1(1, 23) = 5.40$ ;  $p < 0.02$ ;  $F_2(1, 35) = 2.60$ ;  $p < 0.11$ ) and significant at Region VI ( $F_1(1, 23) = 17.6$ ;  $p < 0.005$ ;  $F_2(1, 35) = 5.97$ ;  $p < 0.05$ ). This is not surprising since, here, the operation of coercion occurs only within certain factor combinations.

The crucial comparisons contrasted the effects of Cardinality within the Modifier Type alternation. As can be seen in Figure 1, response latencies for Regions V & VI in Condition A, the Singular+Durative, coerced iterative items, were significantly longer than those in Condition B, the Plural+Durative non-coerced, iterative items (Region V:  $F_1(1, 23) = 7.34$ ;  $p < 0.05$ ;  $F_2(1, 35) = 4.66$ ;  $p < 0.05$ ; Region VI:  $F_1(1, 23) = 24.51$ ;  $p < 0.0001$ ;  $F_2(1, 35) = 9.27$ ;  $p < 0.005$ ). In contrast, as is shown in Figure 2, no effects of Cardinality emerge in the critical text regions of sentences modified by Non-Durative adverbials (All  $F_s < 1$ ).

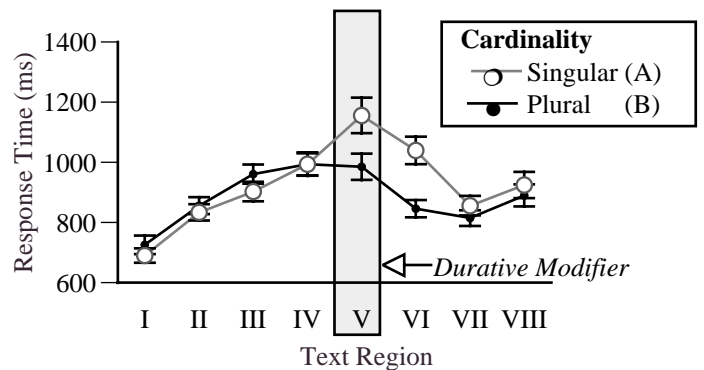


Figure 1: Response latency by text region for **Duratives** by Cardinality of Object

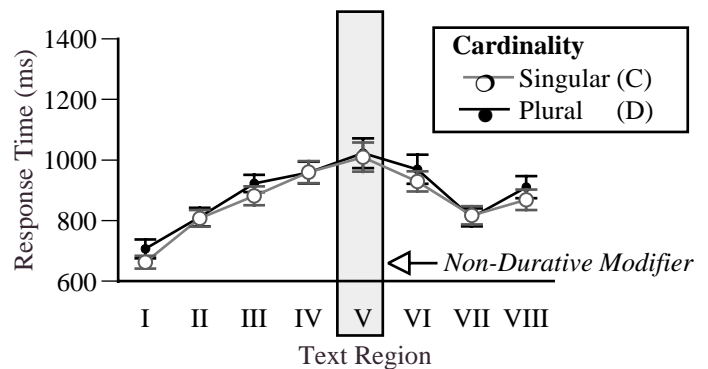


Figure 2: Response latency by text region for **Non-Duratives** by Cardinality of Object

**Make-sense judgment data** As can be seen in Figure 3, even in sentences which were ultimately judged to be acceptable, subjects indicated that the temporal adverbial Region V was difficult to integrate more frequently in the coerced singular durative condition than in the other treatment conditions. Chi-squared analysis reflects that this difference also holds for the Cardinality contrast within the Durative condition ( $\chi^2 = 9.40$  (df=1);  $p < .005$ ), but not between the non-durative conditions ( $\chi^2 = .07$ )

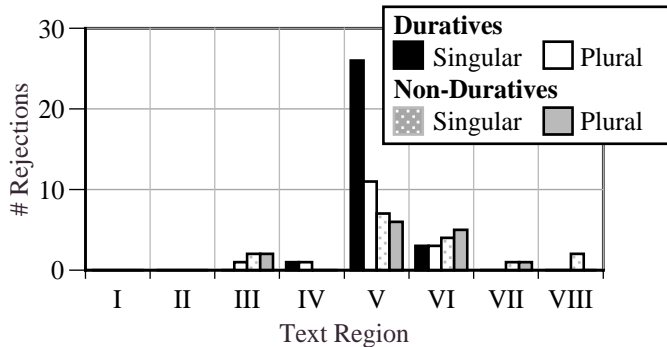


Figure 3: Regional make-sense judgments.

## General Discussion

This study investigated the comprehension of sentences where an aspectually incompatible predicate - modifier combination is interpreted with the aid of the semantic operation of aspectual coercion. Aspectual coercion operates by altering the aspectual specifications of the predicate in a direction matching the input specifications of the adverbial modifier. In the specific case studied here, the semantic consequence of coercion is an obligatory iterative (atelic) interpretation of predicates involving punctual eventive (telic) verbs when these predicates are modified with adverbs of duration. Iterativity arises in this particular situation as the only temporally unbounded analysis applicable to eventive verbs lacking a durational component in their conceptual structure. The goal of our study was to establish whether the coercion operation has any disruptive consequences for sentence comprehension.

We find that participants are significantly delayed when reading a durative adverbial modifier that follows an aspectually incompatible predicate (punctual verb + singular object), as compared to reading the same adverbial following an aspectually compatible modifier (punctual verb + bare plural object). No such difficulties arise when the same predicates are modified by aspectually neutral adverbials, which do not trigger coercion. We hypothesize that the observed latency is indicative of an increase in processing cost associated with the need to undergo coercion in order to form a coherent representation of the utterance. On the other hand, within the predicate, objects of specified vs. unspecified cardinality were read with a comparable degree of ease: this suggests that decisions about utterance aspectuality are made after both the verb and its arguments have been encountered.

While we take our results to indicate that coercion is, indeed, a costly operation, they are compatible with several hypotheses as to *why* this should be the case. On one hand,

it is possible that the difficulty in the comprehension of coerced sentences reflects a price associated with some reanalysis of the current representation of the utterance. That is, it could be that the combination of a telic verb and singular object leads to an early decision of a telic aspectual value for the utterance under construction; and subsequent modification of that value is undesirable (costly). If this is the case, we would expect to observe the same degree of processing difficulty to occur in sentences where iteration is introduced by means of an overt lexical item, e.g. *Howard sent a large check to his daughter every year*.

Alternatively, the difficulty in interpreting coerced iterative sentences may stem from the fact that the existing representation has to be updated through the mediation of an iterative operator that is not morpho-syntactically expressed. To make this point clearer: a durative adverbial must attach to input which has some atelic properties. When this input is a process (*write*) or contains a process-like subcomponent (*write a book*), combination with a durative modifier is unproblematic. However, if the input does not have a continuous interpretation (*send a check*), an attempt to combine it directly with a durative adverbial will lead to an incoherent conceptual representation. The strategy of introducing an iterative operator - which has the effect of creating a novel, atelic event as input to the modifier - can then be regarded as a form of repair. It is possible that the observed processing delay reflects an attempt at the combination of predicate and modifier without the mediation of an iterative operator with the concomitant failure to form a sensible interpretation of the whole. If this is the case, we would expect the coercion effect to disappear in cases where an overt iterative element makes the interpretation domain of the modifier explicit, e.g. again in *Howard sent a large check to his daughter every year*. We plan to address these issues in further research.

## Acknowledgments

We would like to thank Géraldine Legendre and Paul Hagstrom for helpful comments on an earlier draft of this paper.

## References

- Boland, J. (1997). The relationship between syntactic and semantic processing in sentence comprehension. Ms.
- Dowty, D. (1978). *Word meaning and Montague grammar*. Dordrecht: Reidel.
- Moens, M., & Steedman, M. (1988). Temporal ontology and temporal reference. *Computational Linguistics*, 14, 15-28.
- Mourelatos, A. (1981). Events, processes, and states. *Linguistics and Philosophy*, 2, 415-434.
- Piñango, M, Zurif, E., & Jackendoff, R. (1999). Real-time processing implications of enriched composition at the syntax-semantics interface. *Journal of Psycholinguistic Research*, 28, 395-414.
- Vendler, Z. (1967). *Linguistics in Philosophy*. Ithaca, NY: Cornell University Press.
- Verkuyl, H. (1993). *A Theory of Aspectuality*. Cambridge, UK: Cambridge University Press.

# Talking Through Graphics: An Empirical Study of the Sequential Integration of Modalities

Ichiro Umata (umata@mic.atr.co.jp)<sup>1</sup>

ATR Media Integration & Communications Research Laboratories;  
Seika Soraku Kyoto, 619-0288 Japan

Atsushi Shimojima (ashimoji@jaist.ac.jp)

Japan Advanced Institute of Science and Technology;  
1-1 Asahi Tatsunokuchi Nomi Ishikawa 923-1292 Japan

Yasuhiro Katagiri (katagiri@mic.atr.co.jp)

ATR Media Integration & Communications Research Laboratories;  
Seika Soraku Kyoto, 619-0288 Japan

## Abstract

An empirical investigation was conducted on the characteristics of language use in graphics communication settings. Graphics communications, such as dialogues using maps, drawings, or pictures, provide people with two independent systems of representation, spoken language and graphics. Drawing on our dialogue data, we will show that the presence of a graphical representation significantly changes the way the spoken language is used, extending its expressive capacity in most cases. As two remarkable uses of language affected in this way, we will report the phenomena of *mediated reference* and *dual description*, illustrating them with actual examples from our data. Finally, a quantitative analysis of our data will show that these special uses of language are indeed as common as conventional uses of language in the presence of graphical representations.

## Introduction

Conversational exchanges that involve external graphical representations are quite common in our daily lives. People often give and ask directions by referring to maps, or they might draw a floor plan in discussing where to place furniture in a living room. Now, linguistic expressions denote objects and relations in the world. This denotation relation is governed by conventions inside the language. An utterance of a linguistic expression carries unique information about the world through these conventions. This is what the standard view of the semantics of language tells us. However, when we look at speech in conversational exchanges involving graphics, regardless of the language used, we will immediately notice utterances that do not conform to this standard picture.

This paper is a detailed examination of the impact of graphics on the use of language. Our data consist of actual two-party dialogues where participants draw or consult a map during verbal exchanges. We will focus on two remarkable uses of language, called “mediated reference” and “dual description,” that we found through an examination of our data. Both phenomena are clearly specific to dialogues involving some graphical representation, or at least, some external representation other than speech.

Briefly, *mediated reference* is a case where a linguistic expression reaches its “final” referent due to the fact that its “immediate” referent has a referential connection to this final one in the system of graphics. For example, our subjects often use the indexical “kore” (this) to refer to a building or some other landmark, although its immediate referent is clearly an icon on the map; the icon refers to the landmark in the system

of map, and this fact somehow enables the indexical expression to do so too. We will discuss more examples of mediated reference later, and introduce three more varieties of the phenomenon.

*Dual description* is a case where a declarative sentence is used to describe a fact that holds in the graphic as well as the corresponding fact in the situation represented by the graphic. Suppose, when asked about the number of stations between two particular stations, one counts the number of icons on a railroad map and says, “There are three stations in-between them.” Is this report concerned with the map itself, or with the mapped railroad? Is it reporting that the railroad map has three station icons between two particular station icons, or that the railroad system has three stations between the two stations? Whichever the answer may be, it seems clear that the speaker has managed to describe both facts with this sentence. Note that, on the semantics associated with the railroad map, the first fact means the second fact, and this semantic relation somehow underwrites the duplicative use of the sentence.

Both uses of language are so natural and common in a dialogue involving a graphical representation that people may not even be aware of the phenomena. In fact, their frequent occurrence in such settings suggests that they are not a deviant but rather a perfectly legitimate use of language. Yet the empirical research on the integration of linguistic and graphical representations has focused on the issue of how speech is used to disambiguate a graphic (Neilson and Lee 1994) or how a graphic is used to disambiguate speech (Lee and Zeevat 1990). The linguistic-graphic integration has been also studied from a logical point of view, but the focus has been on how a graphic expresses what cannot be easily expressed by a linguistic representation (Barwise and Etchemendy 1996, Shimojima 1999). For both views, the fundamental form of linguistic-graphic integration is a *parallel* one, where each mode of representation expresses information in its own way, but since one mode of representation expresses what the other form does not, they may work complementarily to each other. In contrast, the two phenomena that we are highlighting in this paper point to a rather different form of integration, where the presence of one mode of representation extends the expressive capacity of the other by affecting the way it is used. Our goal is to draw due attention to this *sequential* form of graphic-linguistic integration by demonstrating that the instances of that type of integration are commonly observed in actual human dialogues, as opposed to mere logical possibilities.

In the next section, we will describe the methods through

<sup>1</sup>Also with Kobe University.



which we collected our dialogue data. The two subsequent sections are, respectively, qualitative descriptions of the phenomena of mediated reference and dual description, where we illustrate each phenomenon with examples of language use drawn from our dialogue data. In these sections, the phenomenon of mediated reference is classified into four different types, and the mechanism underlying dual description is analyzed. The final section is devoted to a quantitative description of the two phenomena, where we use “content phrasal unit” to quantify the frequency of mediated reference and dual description in the dialogue process. As it turns out, these graphic-oriented uses of language occur as frequently as standard uses in our dialogue data, indicating that the speakers are quite ready to exploit the graphical representations at hand to extend the expressive capacity of their language.

## Data

The conversational data analyzed in this paper were gathered from a series of graphical communication experiments, which were conducted for a larger research project investigating the interactions between cognitive/communicative factors and graphical representations<sup>2</sup>. Our data consist of 19 task-oriented dialogues, with a total length of 116 minutes. Of these dialogues, 14 involve “a map sketching task” while the other 5 involve “a GRE task.”

**Map Sketching Task** In this task, two subjects were asked to work together as partners to draw a map showing four landmarks in Nara (a local town) as accurately as possible. The subjects were seated in separate, soundproof rooms and worked together using a shared virtual whiteboard and a full duplex audio connection. All inputs to the screen were by stylus, and any writing or erasing by one participant would appear simultaneously on their partner’s screen. The subjects were video-taped during the task.

**GRE Task** In this task, two subjects were asked to solve a logical reasoning problem from the Graduate Record Examination (GRE). The problem was on possible route selections in a hypothetical truck delivery area. The subjects were again required to work on the problem together and it was suggested that drawing a diagram on the screen might help them to answer the question. All pairs drew diagrams and eight pairs among the nine drew map-like ones. The communication environment of the subjects was the same as that in the map sketching task. The GRE task differed from the Map Drawing task in that the subjects had to not only draw an accurate map, but also use it for problem solving.

## Mediated Reference

Studies on the use of multimodal information in reasoning and communication have mostly focused on the complementary or parallel form of integration and have not closely examined the sequential form of integration of multiple modalities. Particularly, little is known about the directionality and the systematic nature of such integration. In our data, two directions were observed in referring to entities through the

sequential integration: mediated references via a representation system (a sketch map, a diagram, etc.) to the world and mediated references via the world to a representation system. We will call the former a *forward* mediated reference, and the latter a *backward* mediated reference. Further, mediated references can be observed between not only individuals but also relations. We will examine, in this section, the phenomena of mediated references observed in our graphical communication experiments. Figure 1 shows four possibilities of mediated references.

### Individual Mediated Reference

Consider the following utterances:

- (1) (From the map data, pointing to a part of the map with the stylus)

*de, koko-ni-ne, tasika Deiri-Sutoa-ga-ne,*  
and, here-DAT, probably, Daily Store-NOM,  
*kono kado-ni atta.*  
this corner-DAT was.  
“And I think there’s Daily Store on this corner.”

- (2) (From the GRE data, pointing to a path on the map with the stylus)

*kore-ga 100 desuka?*  
this-NOM 100 is  
“This is 100km, isn’t it?”

In (1), the speaker was pointing to a part of the map, and the linguistic expression “*koko* (here)” and “*kono kado* (this corner) literally denoted a part of the map. However, there was just a blank space on this part of the map and there were no symbols that could be regarded as an icon of a store. If we assume that the speaker was talking about the map, this utterance would be regarded as simply meaningless or at most false. In this task setting, it is unlikely that the speaker was talking nonsense or lying. Therefore, this utterance was a statement not based on the map but on the real world. Here, the linguistic expressions “*koko*” and “*kono kado*” referred to some place in the world via the place on the map. Similarly, no signs showing the distance could be seen on the diagram in the case of utterance (2), and no suitable properties for the referent of the expression “100” could be found on the diagram. Consequently, this utterance was also a description of some situation regarding the delivery route, not a part on the diagram. In these cases, the reliable correspondence between the spatial configuration of a map and a place in the world enabled *forward* mediated references: references to places in the world through places on the map.

We can also find examples of *backward* mediated references in the data. Some of them are as follows:

- (3) (From the map data, pointing to the icon of Nara Park on the map)

*ja, kore, moo-tyotto koen okkiku suru?*  
So this a-little-more park big make  
“So, shall we make this park a little bigger?”

- (4) (From the map data, after realizing that they made a mistake)

<sup>2</sup>These experiments were designed by Patrick G.T. Healey, Nik Swoboda, Ichiro Umata and Yasuhiro Katagiri.

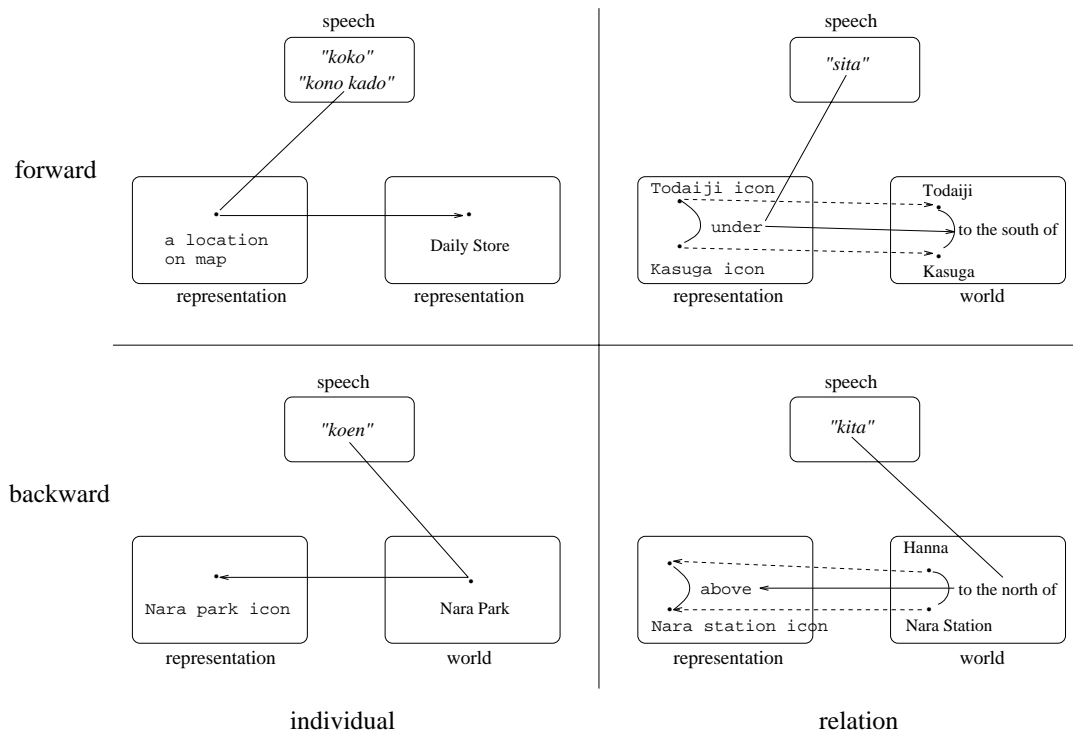


Figure 1: Four categories of mediated reference

*a jaa kesimasyoo Teramati Kita*  
 well so erase Teramati-town Kita-town  
 “Well, so, let’s erase the Teramati-Kita route.”

The linguistic expression “*koen* (park)” in utterance (3) literally denoted a real park. However, one cannot usually make a real park bigger, so this utterance cannot be regarded as a statement about the world. Rather, utterance (3) was a statement about some operation on the map, and “*koen*” referred not to the real Nara Park but rather to the icon on the map. Similarly, utterance (4) was a suggestion to erase the path on the diagram. They could not erase the delivery route itself because it was not assumed to be in the situation described by the GRE problem. In both statements, the objects in the world play an intermediate role, and the linguistic expressions refer indirectly to the icons on the map.

### Relation Mediated Reference

So far, we have concentrated on mediated references between individuals. However, there are also examples of mediated references between relations in our data. Consider the following examples:

- (5) (From the map data, pointing to a part of the map with the stylus)

*kokorahen-ni Toodaiji-ga aru*  
 around-here-DAT Todaiji-temple-NOM is  
*kara, kono sita-no hoo-kana?*  
 because this below-GEN direction-I wonder  
 “Because Todaiji is around here, it (Kasuga-shrine) is probably below this, isn’t it?”

- (6) (From the map data, looking at his partner’s drawing on the map)

*aa, soko zutto ue nobotteiku-to,*  
 Yes there straight up go  
*nyugakusiki-no toko dayo.*  
 entrance ceremony-GEN place is  
 “Yeah, if you go straight up there, you can find the place where we had the entrance ceremony.”

In dialogue (5), the speaker intended to show the listener where Kasuga-shrine was. However, since there were no Kasuga-shrine icons on the map, the subjects were talking about a real-world situation via the map. However, the linguistic expression “*sita*,” which roughly means *under* or *below*, cannot be regarded as referring directly to real world relations; in the real world, Kasuga shrine is not *under* Todaiji temple but *south* of it. Here, “*sita*” referred to the real world relation (i.e. *to-the-south-of*) indirectly via the relation on the map (*under*). This was based on the semantic correspondence established between the map and the world. Similarly, in the case of utterance (6), there were no icons for “*nyugakusiki-no toko*” (the place where we had the entrance ceremony) on the map. Therefore, this utterance was describing a state in the real world, and the expression “*ue*” referred to the spatial relation in the real world (i.e. *to the north*) via the relation on the map (i.e. *up*) in (6).

Excerpt (7) includes an example of a *backward* mediated reference to a relation, as well as examples of individual backward references.

- (7) (From the map data, revising the position of the Nara Station icon)

*Hanna Way-no yori kita-ni ittya*  
 Hanna Way-GEN than north-DAT go  
*akan?*  
 no good-Q

“So, we can’t draw it north of Hanna Way?”

Since one cannot change the place of the real train station, utterances in (7) were about the situation on the map. Thus, the linguistic expression “*Hanna Way*” in (7) referred backwardly to the icons on the map, not to a real world object. Similarly, the linguistic expression “*kita-ni*” (to the north of) makes a backward mediated reference to the relation of the places on the map (i.e. above) via the real world relation to the north of in this utterance<sup>3</sup>. Due to the preservation of the configurational constraints between the map system and the world, such mediated references to spatial relations are quite natural and are commonly found in conversations in which maps are used.

## Informational Duality

Thus, when a map accompanies a dialogue, speakers can make a wide variety of indirect references, either individual or relational, forward or backward, through the systematic semantic relations established between the map and the mapped area. From the speaker’s point of view, this means increased freedom of reference with a limited vocabulary, but from the listener’s point of view, this might mean an increased likelihood that an utterance will become ambiguous in regards to the map itself or the mapped region. For we cannot settle this issue simply by assessing the literal referents of the expressions used, due to the possibility of indirect reference. Purely semantic disambiguation of this sort is generally not applicable.

Fortunately, listeners can often rely on pragmatic cues to resolve such ambiguity, as we have seen in the cases of (1)–(7). Generally, listeners can reject an interpretation of a statement if, on that interpretation, an utterance is to perform a speech act that is not felicitous in that context, such as (i) describing or checking the current position of an object not on the map, or (ii) requesting or otherwise discussing an operation on the mapped region that is impossible to address. The examples (1), (2), (5), and (6) correspond to (i), and thus they were considered not to be about the map, while (3), (4), and (7) correspond to (ii), and they were considered not to be about the mapped region.

However, our data also contain a number of utterances *not* subject to even such disambiguations. In the following, we will consider some of those examples and analyze the informational and functional structures of such utterances.

### An Example

The following dialogue from our GRE data was conducted soon after the partners drew a graph-like map showing the routes connecting various towns, including Kawabata, Kitamati, and Hasimoto. The speakers are concerned with how many towns a truck driver can pass through in one day.

<sup>3</sup>The subjects erased the old icon and were just starting to draw a new one, so “*iku (itty)*” in this utterance expresses the movement of the icon on the map.

(8) A: *kazoemasukanee?*

(Shall we count?)

B: *soosuruto.*

(If we do so, then....)

A: *kazoeruto 3-tu kanaa. 4-tu-wa tyotto muridesuyonee.*  
(On my counting, it is three, I suppose. Four is not feasible, is it?)

B: *uun.*

(Hmm)

A: *Kawa toka dattara, kore moo sudeni 300 toka dakara, moo, Kita, kotti, Kita-ni*

(If this is Kawa or something, and if this is already 300 or so, well, Kita, here, to Kita [Mumbling indistinctively.]

*Kawa-kara Kita-ni itte Hasi-de, kore 3-tu desuyone.*

(Going from Kawa to Kita and then to Hasi, that’s three, isn’t it?)

The case in point is the last utterance of speaker A, which is underlined. On the one hand, one may well regard the names “*Kawa*,” “*Kita*,” and “*Hasi*” to denote the icons for Kawabata town, Kitamati town, and Hasimoto town. In this interpretation, the entire utterance concerns the map, and the speaker is reporting the following information:

(9) There are three town icons on the path: the Kawa icon, the Kita icon and the Hasi icon.

The icons for Kawabata town, Kitamati town, and Hasimoto town already exist on the map, connected by a particular path on the map. Thus, reporting the number of town icons on this path, such as reporting (9), is a speech act that the speaker may well perform at this point. In fact, the above excerpt shows that prior to this utterance, the subjects have explicitly agreed to do such counting. Thus, one cannot reject the interpretation of the utterance as concerned with the map for any obvious reason.

On the other hand, it is also natural to regard “*Kawa*,” “*Kita*,” and “*Hasi*” as indirectly denoting the real towns, and if so, the utterance conveys the following information about a traffic route in the mapped region:

(10) There are three towns on the route: Kawabata town, Kitamati town and Hasimoto town.

Recall that the present problem for the subjects is the maximum number of towns that a truck driver can pass through within one day. The reporting in (10) is directly relevant to the solution to this problem, and hence is a speech act quite likely to be performed at this point. Thus, the interpretation of this utterance as being about the mapped region cannot be rejected, either.

As we will see shortly, our data contain a number of examples of this type, where an utterance is as likely to be about the map as to be about the mapped region. Upon reflection, this type of utterance seems frequent in everyday dialogue involving some graphical representation—we do not always clearly explain this to listeners, or even to ourselves, whether our assertion or report is about the picture at hand or about the situation the picture depicts. So the question is: how can we understand such utterances? Are they *ambiguous* in the sense that: although such utterances are “really” concerned with either the map or the mapped region, they fail to provide sufficient cues to distinguish between them? If we frequently use such ambiguous utterances, how can we ever be successful in communication?

## Analysis

On these questions, we propose that such utterances are *not* concerned with either the picture or the depicted situation *exclusively*. Rather, they are concerned with *both*, and thus handle information about the picture as well as information about the pictured situation. Such utterances are therefore not ambiguous in the above sense. The problem of ambiguity occurs only when we assume that such an utterance is concerned with only one subject matter. Here we explicitly discard this assumption for the kind of utterances in question. For example, the subject matter of the underlined utterance in (8) is not single, but dual, and the utterance reports the number of the town icons on a particular path on the map, *as well as* the number of the towns on the corresponding route on the mapped region.

But how is it ever possible for a single utterance to have such dual informational contents? Briefly, this is possible because representation is a transitive relation. Due to this principle, whenever an utterance represents a picture having a property  $\alpha$ , and this property  $\alpha$  on the picture in turn represents the depicted object having property  $\beta$ , the original utterance will also represent this object having property  $\beta$ . In the present case, the underlined utterance in (8) represents the map having the structural property of (9), and due to the semantic convention associated with the maps, a map with the property (9) represents the mapped region as having the structural property (10). Thus, by transitivity, the utterance also represents the mapped region having the property of (10). This is how a single utterance carries two pieces of information: one about the map and the other about the mapped region.

This mechanism may be made clearer by using the analogy of a copy machine. Suppose you make a copy  $d'$  of a document  $d$ , and then make a copy  $d''$  of the copy  $d'$  that you just made. The copy  $d''$  being a copy of the copy  $d'$  accurately represents  $d'$  more or less, and carries information about  $d'$ . Notice that this copy  $d''$  also carries information about the original document  $d$ —we can look at the second copy  $d''$  and learn what the original document  $d$  is like. (In fact, this is usually the main use of the second copy: we look at it in order to get information about the original document, often forgetting that it also carries information about the first copy.) Thus, the second copy  $d''$  carries two pieces of information, one about the first copy, and the other about the original document. The second copy  $d''$  carries the latter *via* the first copy, thanks to the transitivity of representation.<sup>4</sup>

Our claim is that the same thing happens in the case of the utterance in (8), where the mapped region in the truck delivery area is the original document  $d$ , the map of it is the first copy  $d'$  of  $d$ , and the utterance is the copy  $d''$  of  $d'$ . The utterance in (8) carries information about the mapped region in the truck delivery area via the map, just as  $d''$  carries information about  $d$  via  $d'$ . The utterance carries depicted information (9) and (10) about the map and the mapped region, just as  $d''$  carries duplicated information about  $d'$  and  $d$ .

There are two major advantages to this claim. First, it avoids attributing ambiguity to utterances of this type that are

<sup>4</sup>The idea that the carrying of information is a transitive relation is called “the Xerox Principle” by Dretske (1981); this idea has been a focus of interest in situation theory (Barwise and Perry 1983) and was subsequently developed in qualitative information theory (Barwise and Seligman 1997).

found in spoken dialogues employing graphical representations. Therefore, such utterances are not particularly prone to misinterpretation, which is why these types of utterances can occur frequently without hindering smooth communications among speakers.

The second advantage of our analysis is that it gives us a way to capture a set of mechanisms provided by those “dual” utterances to facilitate problem solving processes involving the use of graphical representations. Recall, from our discussion on example (8), that a dual utterance occurs in a context where two different communicative acts are likely: one concerned with the picture and the other concerned with the depicted situation. In our reckoning, the speaker is considered to perform both acts with the utterance, without skipping or suppressing either act. A dual utterance effectively works as a bridge, and both the speaker and the hearer can engage in joint problem solving by matching and transferring information between the graphics domain and the problem domain. To clarify this point, let us suppose that utterance (8) was ambiguous and actually carried only one piece of information. If it was on the situation in the diagram, then the utterance itself would not convey the information about the world and would not directly lead to the answer to the question. If it was on the world situation, then it would show the answer but have no grounds for it. Under the assumption of dual information, the utterance provides both the answer and the basis for it at the same time: the information on the world based on the information on the graphics. Thus, in general, our proposal offers a more natural explanation of the use of such dual utterances in graphics communication, compared to theories that attribute a single informational content to it.

## Quantitative Analysis

We have demonstrated that a combination of graphical representation and linguistic representation in a graphical communication setting provides us with a novel sequential method for integrating of the linguistic and graphical modalities in the form of mediated and dual references. Our analysis so far has been concerned with classifications and functions of instances of these new types of references.

In order to further establish that the sequential integration actually provides us with a viable and effective mechanism for communication, we conducted a quantitative analysis on the relative frequencies of the “new” forms of references, both mediated and dual references; we performed comparison with “conventional” direct references within our data obtained in our Map and GRE experiment. Furthermore, the different characteristics of each task were expected to result in a different distribution of the final referents of linguistic phrases. The Map corpus was expected to have more instances referring solely to the object in the graphics domain, because the aim of the task was to complete a map. On the other hand, the GRE corpus was expected to include fewer of such instances, because the aim was to solve the problems of the world domain and the graphics simply assist in that purpose.

Our corpus consists of 14,011 words (9,179 for the Map and 4,832 for the GRE), and the number of content phrasal units<sup>5</sup> was 5,325 (3,394 for the Map and 1,931 for the GRE).

<sup>5</sup>A content phrasal unit is a minimum phrasal unit that has a con-

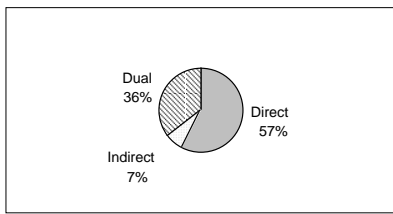


Figure 2: Relative usage frequencies of the direct, mediated and dual references)

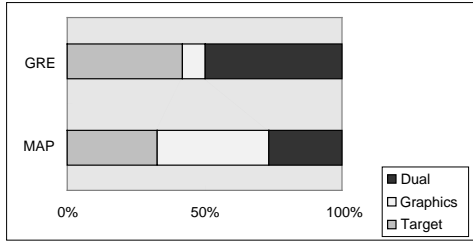


Figure 3: Distributions of final target domains of references for the Map and GRE data

Of them, 4,667 units were the ones describing the situations of the graphics and/or the world domain (2,875 for the Map and 1,792 for the GRE). We classified these units into the three categories shown above: direct, mediated and dual references.

Figure 2 shows the relative usage frequencies of the three types of references: direct, mediated and dual. Of all the reference occurrences, 57% were instances of direct references and 43% were instances of either mediated or dual references. This clearly shows that mediated and dual references are not mere theoretical possibilities or exceptional phenomena, but rather are mundane mechanisms routinely employed in actual communication.

Task characteristics of the Map sketching task and the GRE task can also be captured in quantitative terms. Figure 3 shows the distribution of final target domains of reference for the two tasks. A direct reference to the world and a forward mediated reference through the graphics to the world share the world as their final target domain of reference. Similarly, a direct reference to the graphics and a backward mediated reference through the world to the graphics eventually refer to the graphics as their final target domain. A dual reference is indeterminate as to its final target domain. The final target domains exhibit significantly different distributions between the two tasks ( $\chi^2_{(4)} = 595.60, p < .001$ ). More concretely, (1) the Map data had more instances of graphic-only references, (adjusted residual: Map = 23.75, GRE = -23.75); (2) the GRE data had more instances of world-only references, (adjusted residual: Map = -6.27, GRE = 6.27); (3) the GRE data had more instances of dual references (adjusted residual: Map = -16.02, GRE = 16.02). Thus the assumption that the GRE data would have more world referents and fewer graphic referents than the Map data was supported. Furthermore, it is likely that dual references are strongly related to inferences on graphics, given the abundant instances of dual references in the GRE data.

tent word as its head.

## Conclusion

Based on the data of spontaneous spoken dialogues involving graphic representations, we have analyzed the impacts of the presence of a graphic on the use of spoken language. We found (1) a pre-established semantic relation between a graphic and the situation depicted by it provides the speaker with rich possibilities of mediated references, including forward individual, backward individual, forward relational, and backward relational references; (2) the same semantic relation also lets the speaker use a declarative sentence to express dual pieces of information; (3) mediated reference and dual description are not exceptional but rather mundane mechanisms routinely employed in actual communication; (4) the characteristics of communicative contexts affect the distributions of the final referents of linguistic phrases. We further suggested that the use of dual descriptions is strongly related to inferences on graphics.

These findings indicate that in spontaneous human communications, spoken language and a graphic representation may be used in the *sequential* composition, where the latter affects the usage of the former to extend its expressive capacity. This is in stark contrast to the common view of the interaction between linguistic and graphic modalities, where the integration is made only at the level of multiple pieces of information expressed by the two modalities in individual manners. A *parallel* composition of this type is not the only form of the graphic-linguistic integration, and probably, not even a dominant form.

## References

- Barwise, J., and E. Etchemendy (1995). Heterogeneous Logic. In Glasgow, J. I., Narayanan, N. H., and B. Chandrasekaran (Eds.) *Diagrammatic Reasoning: Cognitive and Computational Perspectives*, 211–234 Cambridge, Mass.: MIT Press.
- Barwise, J., and J. Perry (1983). *Situations and Attitudes*. Cambridge, Mass.: MIT Press.
- Barwise, J., and J. Seligman (1997). *Information Flow: The Logic of Distributed Systems*. Cambridge, U.K.: Cambridge University Press.
- Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, Mass.: MIT Press
- Lee, J., and H. Zeevat (1990). Integrating Natural Language and Graphics in Dialogue. In Diaper, D., Gilmore, D., Cockton, G. and B. Schackel. (Eds.) *Human Computer Interaction—INTERACT'90*, 211–234 Amsterdam:
- Neilson, I., and J. Lee (1994). Conversations with Graphics: Implications for the Design of Natural Language/Graphics Interfaces *International Journal of Human-Computer Studies* 40, 509–541.
- Shimojima, A. (1999). The Graphic-Linguistic Distinction: Exploring Alternatives *Artificial Intelligence Review* 13, 313–335.

# The Dynamics of Simple Prediction: Judging Reachability

Iris van Rooij (irisvr@uvic.ca)

Cognitive Psychology, P.O. Box 3050, Victoria BC, V8W 3P5, Canada

Raoul M. Bongers (bongers@psych.kun.nl)

Developmental Psychology, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

W.F.G. (Pim) Haselager (pimh@nici.kun.nl)

NICI/Cognitive Science, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

## Abstract

This study addresses the dynamical nature of a ‘representation hungry’ cognitive task. Participants were asked to judge whether or not they thought they could reach a distant object with a hand-held rod. The dynamical effects observed in this study support a two-attractor model designed by Tuller, Case, Ding, & Kelso (1994). The results suggest that predictive judgments regarding the (im)possibility of an action may be better understood in terms of dynamically evolving basins of attraction rather than as depending on stable representational structures.

The ability to think about the outcome of a yet to be performed action seems to necessitate a representational explanation. How else to explain this ability except by assuming that the system constructs a model of the situation, represents the imagined action, and concludes on the basis of the ensuing representational structure whether the goal can be achieved by means of the action or not? In this paper we aim to question this representational presupposition by investigating the potential of dynamical systems theory (DST) to model simple prediction.

Within DST, the behavior of a system is analyzed as an emergent property of the interactions between its subsystems. During the last decade the tools of DST have proven to be valuable assets for understanding behavior emerging out of multiple interacting components (Beek, Peper, & Stegeman, 1995; Haken, Kelso, & Bunz, 1985; Schmidt, Carello, & Turvey, 1990; Vallacher & Nowak, 1994). However, most of the behavioral phenomena that are currently described with models developed in DST are not regarded as clear cases of *cognitive* behavior. DST has been challenged to try to deal with more ‘representation-hungry’ domains (Clark, 1997, p. 166-170; see also Clark & Toribio, 1994). One such domain, according to Clark, involves the class of cases that “include thoughts about temporally or spatially distant events and *thoughts about the potential outcome of imagined actions*” (Clark, 1997, p. 167; our emphasis). In the present paper we take a first, exploratory attempt towards answering this challenge by exploring whether participants’ verbal reports on the (im)possibility of an imagined action can be understood from within a DST framework. In our task participants have to indicate whether they think they can reach for an object on a distant table

with a rod. This task can be seen as a simple example of a situation in which one has to predict the possible outcome of an imagined action. By systematically manipulating rodlength we set out to study the dynamical aspects of this prediction behavior.

## Model description

Within the DST approach many different models have been developed to account for global patterns in behavior. Given that the task we studied involved discerning which rods enabled successful reaching and which did not, we used a dynamical model particularly designed to account for behavior with two attractor states. Tuller and colleagues (Tuller, Case, Ding, & Kelso, 1994; see also Case, Tuller, Ding, & Kelso, 1995) applied such a model to speech categorization phenomena. Following the example of Tuller et al. (1994) we use equation 1 to model our data.

$$V(x) = kx - \frac{1}{2}x^2 + \frac{1}{4}x^4 \quad (1)$$

$V(x)$  is a potential function with two minima which are assumed to correspond to two stable conceptual states, viz. ‘No’ (i.e., the participant indicates the belief or judgment that it is not possible to reach the object with the rod) and ‘Yes’ (the participant indicates the belief or judgment that it is possible to reach the object with the rod) respectively. The judgment regarding the imagined action is qualitatively denoted by  $x$  and  $k$  is the control parameter specifying the direction and the degree of tilt of the potential function (c.f. Tuller et al., 1994). As can be seen in Figure 1, for  $k = -1$  only one stable state exists in the system (i.e., ‘No’). Increasing  $k$  forces the function to tilt. Although the initial stable state persists, the attractor becomes more shallow. When the control parameter reaches the critical value  $-k_c$  an additional attractor appears (‘Yes’). From this point on, until  $k$  reaches the second critical value  $+k_c$ , the two stable states coexist (Both ‘No’ and ‘Yes’ are possible responses). At  $+k_c$ , however, the attractor corresponding to ‘No’ ceases to exist. Increasing  $k$  further only deepens the remaining attractor.

Figure 1 illustrates the tendency of dynamic systems to cling to the state they reside in. For each value of the control parameter the state in which the systems has settled is

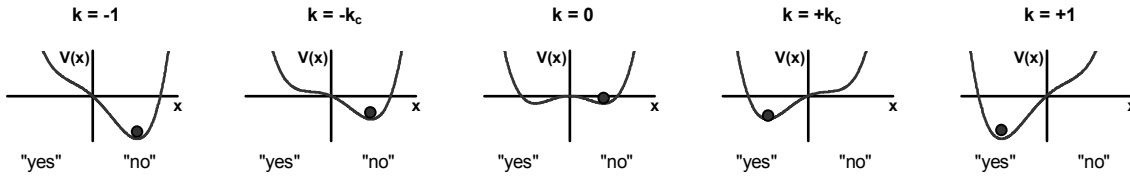


Figure 1: Potential landscape defined by equation 1 for different values of  $k$  (after Tuller, Case, Ding, & Kelso, 1994)

indicated by the black dot. Ideally, the black dot will remain in the attractor it is in for as long as the attractor is relatively stable. This means that when multistability exists the location of the black dot on the potential function depends on whether the control parameter increased from  $-1$  to  $+1$  or decreased from  $+1$  to  $-1$ . As can be seen in Figure 1 this can lead to an observable effect classically associated with dynamical system's behavior, namely *hysteresis*. That is, the switch from 'No' to 'Yes' occurs at a higher value of the control parameter than the switch back from 'Yes' to 'No'.

As was said, this holds for the ideal case, in which the system is not perturbed in any way. Switches between states within the multistable region can occur, however, as a consequence of random disturbances. In a cognitive task like the one we studied, random disturbances may be assumed to correspond to psychological factors, such as fatigue, attention, boredom, and so on (c.f., Tuller et al., 1994).

To capture participants' behavior in our task the relationship between the control parameter  $k$  and the independent variable has to be specified. Following Tuller et al. (1994) we assume that this relationship is not a one-to-one correspondence. Instead  $k$  is a function of (1) rodlength, (2) the number of repetitions of the categorical judgments<sup>1</sup>, and (3) perceptual and cognitive characteristics of the participant. The relationship between the control parameter and rodlength can be symbolized by the following equation,<sup>2</sup>

$$k = \lambda + (N_{no} - N_{yes})S, \quad (2)$$

in which  $k$  specifies the value of the control parameter,  $\lambda$  is linearly proportional to the length of the rod,  $N_{no}$  and  $N_{yes}$  are growing functions of the number of accumulative repetitions of 'No' and 'Yes' respectively,  $S \geq 0$  and represents relevant characteristics of the participant that may fluctuate during the time course of the experiment. Given that  $S$  represents uncontrolled factors influencing task behavior, we cannot know the exact value of  $S$ . Therefore, we take a qualitative approach to the combined influences of  $(N_{no} - N_{yes})$  and  $S$  on the dynamics of the behavior of the participants. In equation 2 if  $(N_{no} - N_{yes})S = 0$  then  $k = \lambda$ . So

when either  $N_{no} = N_{yes}$  or  $S = 0$ , then there is a one-to-one correspondence between  $k$  and  $\lambda$ . However, for  $S > 0$ ,  $k$  will be larger than  $\lambda$  when  $N_{no} > N_{yes}$  and  $k$  will be smaller than  $\lambda$  when  $N_{no} < N_{yes}$ . The rationale of Equation 2 is illustrated by Figure 2 for a coupled sequential run, in which rodlength first systematically increases and subsequently decreases (abbreviated ID-run). For the sake of clarity we hold  $S$  constant and only look at the effect of the accumulative repetitions of a response.

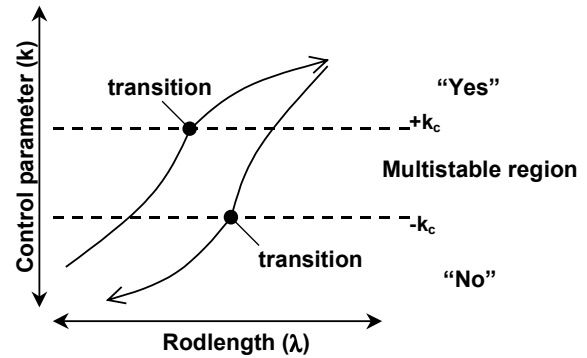


Figure 2: Illustration of the relationship between the control parameter and rodlength, for fixed  $S > 0$ , in a coupled sequential run in which rodlength first increases and subsequently decreases (see text for details).

In an ID-run the participant is presented at first with the smallest rod (bottom left in Figure 2). For short rods the participants start with no-responses and  $N_{no}$  will become increasingly larger than  $N_{yes}$  (which will remain zero) with every next trial. Due to the fact that  $N_{no}$  grows increasingly larger than  $N_{yes}$ ,  $k$  will increase faster than  $\lambda$  increases. When  $k$  reaches the value of  $+k_c$  a transition occurs and the participant switches to yes-responses. With every next trial  $N_{yes}$  will grow, whereas  $N_{no}$  will not. Hence the slope of the function  $k$  will decrease. Because the increase sequence is followed by a decrease sequence  $N_{yes}$  will start to outnumber  $N_{no}$ . This will cause  $k$  to decrease faster than  $\lambda$ . When  $-k_c$  is reached a transition occurs and the participant will switch to yo-responses. Figure 2 thus illustrates that for sufficiently large  $S$  the transition from 'Yes' to 'No' occurs at a larger rodlength than the transition from 'No' to 'Yes'. This is an example of the *enhanced contrast* effect. One can imagine that for a certain settings of the parameters one may find that the first and second transition occur at exactly the same rodlength, i.e. a *critical boundary*. However, the number of parameter settings that result in critical boundary

<sup>1</sup> See also Parducci's (1965; Parducci & Wedell, 1986) range-frequency theory and Helson's (1964) adaptation-level theory.

<sup>2</sup> See Tuller et al. (1994) for the original, more explicitly specified relationship between the control parameter  $k$  and the experimental variable  $\lambda$ . The simplification in the form of equation 2 is sufficient for our purposes.

is much smaller than the number of settings that result in either hysteresis or enhanced contrast.

The interrelationship between Equation 1 and 2 as described above leads to the following predictions for our experiment: (1) There is a tendency in the dynamic system to remain in the state it resides in. This means that participants will tend to give the same response as on preceding trials. (2) Accumulative repetitions of 'yes' will cause the multistable region to shift towards the upper end of the rodlength continuum. Conversely, accumulative repetitions of 'No' will cause the multistable region to shift towards the lower end of the rodlength continuum. (3) The higher the number of repetitions of 'Yes' in a run where rodlength increases and subsequently decreases the greater the chance of observing enhanced contrast and the smaller the chance of observing hysteresis. Conversely, the higher the number of repetitions of 'No' in a run where rodlength decreases and subsequently increases the greater the chance of observing enhanced contrast and the smaller the chance of observing hysteresis. Observations of critical boundary will overall be very limited. (4) Within the multistable region switches in perception can occur as a consequence of random disturbances. The narrower the multistable region (e.g., due to repetitions of a certain response – see figure 2) the smaller the chance of observing perceptual switches.

## Method

### Participants

Fourteen participants, 5 male and 9 female, participated in the experiment. All but two female participants were right-handed. The age of thirteen participants ranged from 22 to 28 years. One male participant was significantly older than the rest, viz. 56 years of age. The height of participants ranged from 1.56 to 1.88 meters, with an average of 1.76 meters. All participants, except one who volunteered, were paid for their participation or participated as a means to fulfilling a course requirement.

### Material

Rods with a diameter of 1.25 cm were used, ranging in length from 57.0 to 91.5 cm, in 1.5 cm increments.<sup>3</sup> The twenty-four rods were constructed from wood (density 0.67 g/cm<sup>3</sup>). Attached to each rod was a handle of identical material with the length of 11.5 cm and a diameter of 1.25 cm. A small disc divided the handle from the rod.

A PVC cylinder (diameter 5 cm, height 6 cm) was placed on a table (25x25 cm). The height of the table was adjusted

to the participant's wrist height with the arm at the side. The back of the cylinder was placed against a barrier of 12.5 cm height and the front of the cylinder was aligned with the front edge of the table.

### Procedure

A participant was asked to bend forward, with his/her preferred arm stretched as far as possible (i.e., bending forward while maintaining enough balance to stay flat on the feet). The distance between the feet and the hand in this position was measured. This measure was used to determine the distance to the table at which each participant was to be positioned during the experimental session (i.e., maximum distance reachable without rod + 75 cm<sup>4</sup>). Participants were subsequently asked to take this position and stayed there during the entire experiment. While standing at this distance it was explained to the participant that the goal was to displace the cylinder positioned on the table. The participant was subsequently handed a rod and was instructed to hold the rod so that it made an angle of approximately 45 degrees upwards with the horizontal. The participant stood upright with the rod in one hand and judged whether he was able to reach the cylinder with the rod from that position while keeping the two feet flat on the floor. After a participant had given his categorical judgment he returned the rod to the experimenter and was handed a new rod for which the participant again made a judgment. No feedback regarding accuracy was given.

### Design

Each participant performed this judgment under several conditions. There were three kinds of *sequences* in which rods were given to the participant, namely (1) increase sequences (I): rodlength increased from minimum to maximum in 1.5 cm increments; (2) decrease sequences (D): rodlength decreased from maximum to minimum in 1.5 cm increments; (3) random sequences (R): the rods ranging in length from the minimum to maximum were randomly assigned to the task. The two sequential condition I and D were always coupled, resulting in two kinds of coupled sequential runs: increase-decrease (ID) runs and decrease-increase (DI) runs. Coupled sequential runs were always followed by a random sequence, resulting in two possible blocks of runs, namely increase-decrease-random (IDR) blocks and decrease-increase-random (DIR) blocks. The random sequence served as a kind of buffer between the coupled sequential run preceding it and the coupled sequential run of the next block, and as a control condition in the analyses of the data.

Two different *ranges* were used in the experiment, namely range1 of 57.0 - 85.5 cm and range2 of 63.0 - 91.5 cm.

<sup>3</sup> Psychophysics studies (e.g. Morgan & Watt, 1989; Watt, 1984) suggest that the Weber fraction ( $\Delta I/I$ ) for length discrimination is approximately 0.05. In our experiment the fraction between the increment and rodlength ranged from 0.026 to 0.016. This means that in increase and decrease sequences the direction of change in rodlength is not perceivable for participants from one trial to another. On most occasions the fact that the hand-held rod is longer or shorter than a preceding rod does not become apparent within less than three or more trials.

<sup>4</sup> We added 75 cm to the personal maximum reaching distance, because in the range of rodlength used in this experiment 12 rods  $\leq$  75 cm and 12 rods  $>$  75 cm. Hence, for all participants exactly half of the rods used in the experiment would enable reaching, and half would not.



Thus, there were two possible minima and maxima for the three sequences described above. Within a given block the minimum and maximum for the three constitutive sequences (I, D and R) were the same. The four possible combinations of block and range in the experiment were thus, in shorthand, IDR-1, IDR-2, DIR-1 and DIR-2. Each of these combinations occurred twice in one experimental session, resulting in a total of 480 trials (2 ranges x 2 blocks x 3 sequences x 20 rods x 2 repeated measures) per participant. The block-range combinations were randomized within an experimental session, with the constraint that each block-range combination appeared as often in the first half of a session as in the second half.

## Results

Most participants showed a transition in judged possibility in all sequences (increase-, decrease- and random-sequences). Two of the fourteen participants, however, overestimated the distance reachable so much that the lower-end of the range (57.0 cm) was still too high to evoke a perceptual transition. For this reason these two participants were excluded from the analyses. Plotting the average response against rodlength for the remaining twelve participants for the three types of sequences resulted in the cumulative distributions as depicted in Figure 3.

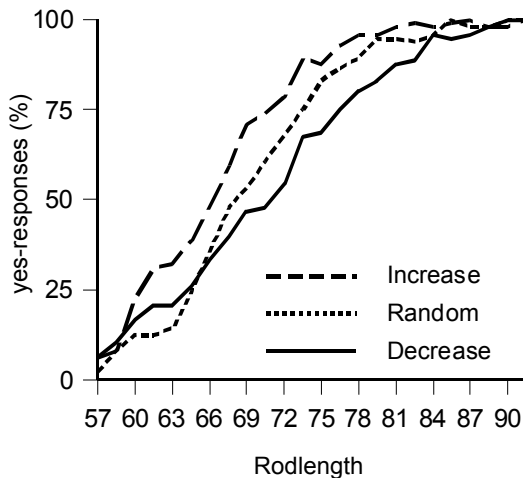


Figure 3: Percentage of 'Yes' responses, averaged over subjects, per rodlength for increase, random and decrease sequences separately.

On average participants in this experiment tended to *overestimate* their reaching distance. Given the individually defined distance to table (personal maximum reaching distance without rod + 75 cm) the *expected* 50% category boundary would be about 75 cm for all participants. The *observed* 50% category boundaries as depicted in Figure 3 are all lower than this. The finding that participants tended to overestimate the distance reachable is in correspondence with findings in other experiments on judging reachability (Heft, 1993; Rochat, & Wraga, 1997).

To test the effect of sequence, suggested in Figure 3, a measure was required for the transition point in each sequence independently. Because in 36 sequences multiple transitions were observed across the rodlength continuum, the data of these sequences were transformed so that a single 'average' transition-point resulted<sup>5</sup>. For the other 192 sequences the real transition-point was simply used as average transition-point. On the average transition-points a 3 x 2 x 2 repeated-measures ANOVA was performed with Sequence (random-, increase- and decrease-run), Block (IDR, DIR), and Range (range1, range2). A main effect of Sequence was found,  $F(2, 22) = 14.68, p < .001$ . A difference contrast, comparing the average transition-points in increase- and decrease-sequences, revealed a contrastive effect, viz. the average transition-point was significantly lower in increase sequences (66.91 cm) than in the decrease sequences (70.73 cm),  $F(1, 11) = 17.19, p = .002$ . The average transition-point in random sequences was 69.23 cm. Further, the main effect of Range was significant,  $F(1, 11) = 12.39, p = .005$ . The average transition-point was smaller for range1 (68.20 cm) than for range2 (69.70 cm). None of the other effects was significant.

To see whether local contrastive or assimilative effects were present in random sequences the conditional probability of judging each rod as belonging to the same category as the preceding rod was investigated. We found that in random sequences participants tended to give the same response as given on the previous trial,  $\chi^2(1) = 58.54, p < .001$ .

Within the multistable region perturbing influences can make one percept change into the other and vice versa. Outside the multistable region only one perceptual form is possible. Taking these theoretical assumptions into consideration the boundary of the multistable region was estimated by the *last* transition-point<sup>6</sup> within a given sequence. Each coupled sequential run (ID-runs and DI-runs) was coded for the type of response pattern it showed, i.e. either hysteresis, critical boundary or enhanced contrast. In 66 of the 91 coupled sequential runs<sup>7</sup> (72.8%) an enhanced contrast effect occurred and in 20 runs (20.7%) a hysteresis effect. Critical boundary occurred in only 6 coupled sequential runs (6.5%), and no more than once per participant.

<sup>5</sup> This transformation involved re-ordering of the no- and yes-responses within a given sequence so that a single transition-point resulted. The total number of no-responses was projected onto the lower part of the rodlength continuum and the total number of yes-responses onto the upper part. The average transition-point was taken to be exactly between the rod receiving the last no-response and first yes-response in the transformed data.

<sup>6</sup> In an increase sequence this last transition-point was defined as being in between the longest rod receiving a no-response and its subsequent rod. Conversely, in a decrease sequence the last transition-point would be in between the shortest rod receiving a yes-response and its subsequent rod.

<sup>7</sup> Five coupled sequential runs were excluded from the analyses because no perceptual transition occurred.

Because participants overestimated the distance reachable the number of accumulative repetitions of ‘Yes’ in a ID-run were, on average, larger than the accumulative repetitions of ‘No’ in a DI-run. Hence, the dynamical model predicts that the chance of observing enhanced contrast is greater, and the chance of hysteresis is smaller, in ID-runs than in DI-runs. An analysis of the frequencies of the two response patterns confirmed this prediction. A Pearson Chi-Square test with SequenceCoupling (ID-, DI-runs), and ResponsePattern (enhanced contrast, hysteresis) indicated a significant association between SequenceCoupling and ResponsePattern,  $\chi^2(1) = 4.44, p = .035$ . Enhanced contrast occurred more frequently in ID-runs (37 times) than in DI-runs (29 times). Hysteresis on the other hand occurred more frequently in DI-runs (14 times) than in ID-runs (6 times). Critical boundary occurred as often in ID-runs (3 times) as in DI-runs (3 times).

Additional switches (i.e., alternating yes- and no-responses on successive trials preceding the last transition-point) occurred on 88 of the 3574 trials<sup>8</sup> in sequential runs. We found that more additional switches occurred in range1 (61 times) than in range2 (27 times;  $\chi^2=13.70, p < .001$ ). Interestingly, this effect of range was observable for both increase and decrease sequences (see Figure 4).

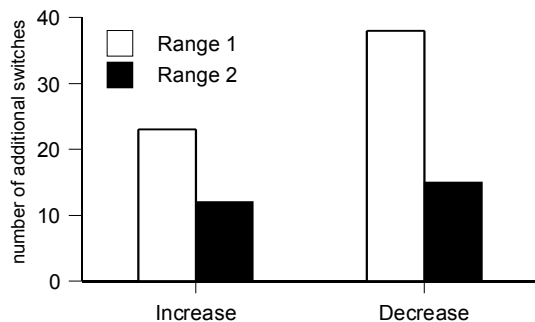


Figure 4: Number of additional switches observed in range 1 and range 2 for increase and decrease sequences separately.

The effect of range in decrease sequences can be understood as being due to the relatively large number of repetitions of ‘Yes’ in decrease sequences within range2 as compared to range1. The two-attractor model can also account for the effect of range in increase runs. As can be seen in Figure 3, even the shortest rods used in the experiment were occasionally judged to enable successful reaching. Further, the shortest rod in range2 (i.e., 63,0 cm) was judged as enabling successful reaching once by two participants, and as enabling successful reaching even 50% of the time by three participants. This means that the left boundary of the multistable region was not only on average

<sup>8</sup> This number indicates the total number of trials in the sequential runs of the experiment, i.e. (12 participants) x (40 trials) x (8 coupled sequential runs) = 3840, *minus* the first trial of each coupled sequential run and *minus* the trials in which a last transition-point was observed.

closer to the left end of the range of rodlengths, but even outside range2 for a considerable number of subjects. Thus, for these participants, the increase sequences in range2 started well within the multistable region. Consequently there were simply fewer opportunities for switching in increase sequences in range2 as compared to range1, which explains the low frequency of additional switches in range 2 for increase sequences.

## Discussion

Clark (1997) challenged DST to explain behavioral phenomena that are considered to be ‘representation hungry’ cases of cognition. We focused on the ability to predict the outcome of a to be performed action. Participants had to judge whether a rod afforded displacing an object from a certain distance. In our interpretation of Clark (1997) such behavior can be classified as ‘representation-hungry’, that is, the task seems to require a model of the situation, a representation of the imagined action, and computations based on those representations to determine whether the action will satisfy the goal.

In the present study we explored whether the judging behavior of our participants could be explained with a dynamical model. The results are in close agreement with the predictions derived from a two-attractor model (c.f., Case et al., 1995; Tuller et al., 1994). First, it was found that in random sequences participants tended to give the same categorical judgment as on preceding trials. This *assimilative* effect is in accordance with the notion that a dynamical system tends to cling to the state it resides in (*Prediction 1*).

Further, we observed that on average the transition from ‘No’ to ‘Yes’ in increase runs occurred at a shorter rodlength than the transition from ‘Yes’ to ‘No’ in decrease runs. Also we found that on average the transition occurred at a shorter rodlength in range 1 than in range2 (independent of the order of presentation of the rods). Both these effects can be interpreted as being due to the influence of accumulated repetitions of a certain response causing the multistable region to shift closer to one of the ends of the rodlength continuum (*Prediction 2*).

In coupled sequential runs (ID- and DI-runs) we observed all three effects that are predicted by the model, viz. hysteresis, critical boundary, and enhanced contrast. As expected critical boundary was the rarest of the three. Because participants overestimated their reaching distance to a large degree many more accumulative repetitions of ‘Yes’ responses occurred in coupled sequential runs in which rodlength first increased and subsequently decreased (ID-runs) than ‘No’ responses occurred in runs in which rodlength first decreased and then increased (DI-runs). As predicted, enhanced contrast occurred more often, and conversely hysteresis less often, in ID-runs as compared to DI-runs (*Prediction 3*).

Finally, more additional switches (alternating ‘No’ and ‘Yes’ responses) were observed when the multistable region

was expected to be relatively large, than when it was expected to be relatively small (*Prediction 4*).

Dynamic systems models typically describe behavior on the level of the whole system. On this account behavior is seen as a self-organized pattern, emerging from the interaction between subsystems. Such a pattern is called the collective variable or order parameter, which in turn can 'enslave' the behavior of the components (cf. Haken & Wunderlin, 1990, p. 7; Kelso, 1995, pp. 8-9). Despite the great complexity at the level of the interacting components the behavior of the system as a whole can be described and understood in terms of the lower-dimensional order parameter dynamics.

According to Clark (1997) an explanation of cognitive capacity in representational terms is valuable if the representations are distinguishable as entities serving a role as information-carriers for behavior. But what if, as DST would have it, a behavioral pattern is best understood as an emergent property of the overall activity of the system? Clark argues that in "such cases (if there are any), the overall system would rightly be said to represent its world—but it would not do so by trading in anything we could usefully treat as internal representations" (Clark, 1997, p. 168). We submit that the effects we observed can be fruitfully interpreted as a consequence of the inter-relationship between control parameter  $k$  and the collective variable  $V(x)$  governing the system. In all, these findings suggest that predictions regarding the possible outcome of an imagined reach are better understood in terms of dynamically evolving basins of attraction rather than as depending on stable representational structures.

## References

- Beek, P. J., Peper, C. E., & Stegeman, D. F. (1995). Dynamical models of movement coordination. *Human Movement Science, 14*, 573-608.
- Case, P., Tuller, B., Ding, M., & Kelso, J. A. (1995). Evaluation of a dynamical model of speech perception. *Perception & Psychophysics, 57*(7), 977-988.
- Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese, 101*, 401-431.
- Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, Massachusetts: MIT-Press.
- Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics, 51*, 347-356.
- Haken, H., & Wunderlin, A. (1990). Synergetics and its paradigm of self-organization in biological systems. In H. T. A. Whiting, O. G. Meijer & P. C. W. van Wieringen (Eds.), *The natural-physical approach to movement control*. Amsterdam: University Press.
- Heft, H. (1993). A methodological note on overestimates of reaching distance: Distinguishing between perceptual and analytical judgements. *Ecological Psychology, 5*, 255-271.
- Helson, H. (1964). *Adaptation-level theory: An experimental and systematic approach to behavior*. New York: Harper & Row.
- Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, Massachusetts: MIT-Press.
- Morgan, M. J., & Watt, R. J. (1989). The Weber relation for position is not an artefact of eccentricity. *Vision-Research, 29*(10), 1457-1462.
- Parducci, A. (1965). Category judgment: range-frequency model. *Psychological Review, 72*, 407-418.
- Parducci, A., & Wedell, D. H. (1986). The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 496-519.
- Rochat, P., & Wraga, M. (1997). An account of the systematic error in judging what is reachable. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 199-212.
- Schmidt, R. C., Carello, C., & Turvey, M. T. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance, 16*(2), 227-247.
- Tuller, B., Case, P., Ding, M., & Kelso, J. A. (1994). The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology: Human Perception and Performance, 20*(1), 3-16.
- Vallacher, R. R., & Nowak, A. (1994). *Dynamical systems in social psychology*. San Diego: Academic Press.
- Watt, R. J. (1984). Towards a general theory of the visual acuities for shape and spatial arrangement. *Vision-Research, 24*(10), 1377-1386.

# Goal Specificity and Learning with a Multimedia Program

**Regina Vollmeyer** (vollmeyer@rz.uni-potsdam.de)  
Institut für Psychologie, Universität Potsdam, Postfach 601553  
14415 Potsdam, Germany

**Bruce D. Burns** (burnsbr@pilot.msu.edu)  
Department of Psychology, Michigan State University  
East Lansing, MI 48824-1117

**Falko Rheinberg** (rheinberg@rz.uni-potsdam.de)  
Institut für Psychologie, Universität Potsdam, Postfach 601553  
14415 Potsdam, Germany

## Abstract

Previous research has found that nonspecific goals (NSG) lead to better learning than a specific goal (SG). We studied this effect with a multimedia program in which participants had to learn about the outbreak of World War 1 either with the goal to find twenty dates (i.e., SG) or with the goal to explain the reasons for the war (i.e., NSG). As expected, the NSG-group better remembered facts about the text during the task and knew more at the end than the SG-group. The NSG-group may also better transfer what they had learnt to a new situation. To try to explain this effect, a number of process variables (strategy systematicity, motivation, number of pages read) were measured. SG- and NSG-group differed in terms of which variable best predicted learning: As expected, for the NSG-group challenge was the best predictor of performance, but probability of success was the best for the SG-group.

## Introduction

Effects of goal specificity on problem solving have been found in a number of recent studies (Geddes & Stevenson 1997; Miller, Lehman, & Koedinger, 1999; Sweller, 1988; Vollmeyer, Burns, & Holyoak, 1996). All these studies have found that giving problem solvers a specific goal state to reach led to poorer learning of the task than if they were given a nonspecific goal, such as to explore. However a difficulty has arisen when trying to form general conclusions about this work and its scope: different researchers have used very different tasks and have given different instantiations to the concept of goal specificity. Further the question arises of what relationship does the work on goal specificity in problem solving have to other goal specificity in organization psychology (see Locke & Latham, 1990), and possibly related research, such as that into explanation effects (Chi, Bassok, Lewis, Reimann, & Glaser, 1989).

To deal with this issue, in this paper we propose a conceptualization that generalizes what goal specificity is and applies that concept to the development of a new task for investigating goal specificity effects. This paper is a preliminary study that aimed to demonstrate goal specificity effects with a new task, and to show that this task has the potential to increase our knowledge of why such effects occur. We chose to use a multimedia learning program as

our task, both because of the explosion of interest in such programs (Issing & Klimsa, 1997) as cheap computer technology becomes widely available, and because it is very different to the problem solving tasks with which goal specificity effects have previously been found.

## Defining Goal Specificity

In Vollmeyer et al. (1996) we proposed that goal specificity effects could be explained in terms of dual-space theories of problem solving (Klahr & Dunbar, 1988; Simon & Lea, 1974). Specific goals (SG) could be seen as encouraging search of an instance or experiment space. Such a space corresponds to what is usually meant when we refer to problem solving as search; that is, we set specific subgoals that are part of that space and reach the goal via those subgoals. A specific goal is a state in such a problem space, which is why specific goals encourage a focus on this space. In contrast, a nonspecific goal (NSG) could be seen as encouraging search of rule or hypothesis space. Such a search space contains the possible rules or hypotheses that may govern the task, but testing such rules requires a coordinated search of instance and rule space.

If we assume that dual space theories provide an appropriate way to characterize specific and nonspecific goals, then it provides a definition of goal specificity. Specific goals are goals that promote reaching set states, nonspecific goals are those that encourage discovery of the nature of the task. By this definition, goal specificity qualitatively affects how people learn, not just how much.

**Self-explanation Theory.** Chi et al. (1989) found that students who explained math problems to themselves did better than students who only worked out examples. Although they do not use the term *goal specificity*, their characterization of explanations seems to fit to our definition of a nonspecific goal. Explaining math problems requires understanding the principles they are based on, whereas not requiring explanation allows problem solvers to focus on the solution alone.

However, self-explaining students took more time for the same task, thus, it could have been that time on task was a moderating variable. In several studies of the self-

explanation effect, Renkl (1997; for an overview, see Renkl, 1999) pointed out that time-on-task could be a moderator of performance. Self-explaining students take more time if given the opportunity, so in his experiments Renkl controlled for time and manipulated self-explanation in asking the self-explanation group to learn in a way they could explain the task to another person. Although both groups worked for the same amount of time, the self-explanation group had better learning outcomes than a group who only solved the problems. Therefore time-on-task appears to be important in these types of tasks.

**Goal Specificity in Organizational Psychology.** In organizational psychology, what has been known as goal specificity has been studied extensively in terms of goal setting (see Locke & Latham, 1990). In this literature specific goals are a form of target (e.g., “Make ten widgets”) whereas nonspecific goals are general admonitions to do well (e.g., “Make as many widgets as possible”). Tubbs’ meta-analysis (1986) showed that specific goals help performance ( $d = .82$ ), although some studies reported exactly the opposite. However these goals are specific and nonspecific in a different way to what has been meant in problem solving research. These goals are forms of targets, rather than states of the problem. Using our definition of goal specificity, we can see that the question of whether the two types of specificity are related depends on how these target goals might affect which space learners focus on. In this paper we will not address the possible connections between these two types of goal specificity.

**Goal Specificity in Multimedia.** In the literature on learning from multimedia texts, *open tasks* have been found to lead to better understanding than *closed tasks*. In their meta-analysis, Chen and Rada (1996) found consistently strong evidence that learners given an open task were more effective than learners given a closed task. However, there is a huge variety in what is subsumed under the closed or open manipulation. In general, closed tasks can be seen as those presenting learners with specific goals, for example to find a particular piece of information, whereas open tasks have very general goals, for example to learn for a test. Thus this distinction can be seen as roughly fitting to our definition of goal specificity: closed and open tasks are those with specific and nonspecific goals, respectively. Learning with multimedia is a particularly interesting domain for examining goal specificity as how people gather information can be explicitly coded by looking at the sequence of actions they perform on the computer. So it provides an opportunity for us to try to measure strategies.

**Goal Specificity and Motivation.** Kanfer and Ackerman (1989) proposed that goal specificity might affect motivation. To examine this issue we applied Vollmeyer and Rheinberg’s model (1998). They proposed a model that assumes that initial motivation affects learning through the mediating variables of motivational state during learning and strategies used for learning. The initial motivation contains four factors: (1) *probability of success*, which is the learners’ level of certainty about whether they will succeed

in performing the task; (2) *fear*, which is how anxious learners are about failing in the task; (3) *challenge*, which is the extent to which learners perceive this task as requiring competence; and, (4) *interest*, which is how much learners like the topic of the program.

Schiefele (1996) reported that in several experiments interest had a positive effect on text learning, especially on understanding texts as opposed to learning facts. Given that interest and challenge are highly correlated (Rheinberg & Vollmeyer, in press), we assumed that challenge would have the same effect. For facts, we assumed that good learning depends more on the learner’s expectancy of receiving a good result. Our nonspecific goal definition can be seen as equivalent to what Schiefele called *understanding-oriented learning*, whereas our specific goal definition corresponds to *fact learning*. Therefore we expected that in a nonspecific goal condition initial interest and challenge would predict learning but in the specific goal condition probability of success and fear (which are indicators of expectancy) should be better predictors.

### Study Aims

This paper reports our first exploration of a multimedia program we developed to study goal specificity. To develop this task we had to determine how to apply the concept of goal specificity in problem solving to a very different task.

**Multimedia Program.** The topic of the program was the outbreak of World War 1. The computer could present up to 51 different pages describing the events leading up to the start of the war. Each page had links to other pages and could have links to videos, sound files, or text boxes containing additional information. Most of the pages were arranged into five sequences, each describing the events occurring in one of the five critical countries (Austria-Hungary, England, France, Germany, and Russia). In addition, topics such as nationalism and imperialism were covered. As an event could concern two or three countries (e.g., declaration of war), learners sometimes saw the same page in the sequence for multiple countries. After every page learners could decide whether to continue reading pages about the same country or to switch to another country or topic. Thus the program provided us with a way to examine the learners’ strategies. In Vollmeyer et al. (1996) we found that strategy systematicity was critical to performance but the task was so different that we could not use this operationalization. Therefore in the multimedia program we operationalized systematicity as the extent to which learners followed sequences, instead of jumping from one topic to another.

To operationalize goal specificity, Chi et al.’s (1989) idea of explanation seemed most applicable to this task given our definition of goal specificity. So we had NSG learners go through the program with the goal of explaining the outbreak of World War 1 to someone else. However unlike Chi et al. we did not have participants give explanations during the task. To increase the contrast between NSG and SG learners, the SG-learners were given a list of 20 specific events and asked to fill in the dates for those events.

**Predictions.** We expected the NSG group to learn more than the SG group, and to be better able to transfer the lessons of the outbreak of World War 1 to another situation. We expected this greater learning to be a result of NSG learners being more systematic. As the NSG learners are more systematic, that is they put more effort into understanding the content of a page they should spend more time per page than SG participants.

We did not expect goal specificity effects on initial motivation. However, the motivational state during learning may change during the task in different ways for the two groups as they react to their perceived success or failure in attaining their goal. In particular we tested Schiefele's (1996) proposal that challenge and interest would relate more strongly to performance for NSG than for SG learners.

## Experiment

### Method

**Participants.** Forty-five students at the University of Potsdam participated in the study and received DM 10.00 (~US\$5) or course credit.

**Design.** There were two levels of goal specificity. The SG group consisted of 24 participants who received instructions to look for dates in the history multimedia program. The NSG group consisted of 21 participants who were told to understand the problem as if they would have to explain it to another person.

**Procedure.** Before the participants started working with the multimedia program they read that they would learn about the outbreak of World War 1. They were informed that they would work with the program for about 25 minutes and then answer a questionnaire. We set a fixed time span as we felt it was important to control for time in order to remove any possibility of time-on-task being used as a factor to explain goal effects. We also told participants that they would be interrupted at various times so that we could ask them what they thought about the task. These interruptions were necessary in order to measure participants' motivational states and to sample their knowledge. The instructions also contained the goal specificity manipulation. The NSG participants were asked to "...work with the program so that you could tell another person about the reasons for the outbreak of World War 1." The SG participants were asked to "... work with the program so that you can fill out correctly the following time-line." The time-line consisted of twenty events, such as the assassination in Sarajevo, for which the learners had to find the dates in the program.

After reading the instruction participants answered the QCM (*Questionnaire of Current Motivation*, by Vollmeyer & Rheinberg, 1998). This questionnaire measured their initial motivation on the four factors *probability of success* (example items: "I think I am up to the difficulty of the task", "I probably won't manage to do this task"), *fear* (example items: "It would be embarrassing to fail at this task", "I feel petrified by the demands of this task"), *interest* (example items: "After having read the instruction the task seems to be very interesting to me", "For tasks like this I

don't need a reward, they are lots of fun anyhow."), and *challenge* (example items: "This task is a real challenge for me", "If I can do this task, I will feel proud of myself").

When working with the multimedia program participants were interrupted every seven minutes for a total of three times. During each of the three interruptions they were asked to answer two types of questions: a motivational state questionnaire, and one factual question about each of the last three pages the participants had seen in the program.

**Process Variables.** Three process variables were measured while learning.

(1) *Motivational state.* Every seven minutes participants answered ten questions (example items: "The task is fun", "I'm sure I will find the correct solution") on a seven-point scale. A composite score was calculated to represent motivational state. Responses were averaged together.

(2) *Strategy systematicity.* Our aim was to find indicators for how systematically a learner works through the program. As this was our first use of this multimedia program, it was not clear what the best measure was. We chose to measure how often learners read a page that followed from the previous one, as opposed to jumping to a new topic. This variable was called sequence. For this we counted the pages that followed logically from the previous one. We then divided this count by the total number of pages participants looked at. An example of following a sequence would be if after the first page for Germany the second page for Germany was looked at. Switching to the first page of France would have been counted as not following the sequence.

(3) *Number of pages.* We counted the number of pages that participants looked at for more than five seconds. Pages looked at for 5s or less were probably mistakes, or arose because the learner realized they had already read the page. Given that each learner was given about the same amount of time to work with the program, we expected that looking at fewer pages would be an indicator of going into the contents of the pages in greater depth.

**Outcome Variables.** To measure knowledge we used a pilot study to develop a questionnaire. For every page that was part of a sequence for a country or side topic (34 in all), we formulated a multiple-choice question with five options. As we had the hypothesis that NSG-learners would read more carefully than SG-learners we formulated factual as well as inference questions, similar to the suggestion of Royer, Carlo, Dufresne and Mestre (1996). So 24 pages had factual questions, ten had inference questions, and one general question was asked.

There were three outcome variables, two were indicators of knowledge and one of transfer.

(1) *Sampled knowledge.* The relevant item from the questionnaire was asked about each of the last three pages learners had seen.

(2) *Accumulated knowledge.* After participants had worked with the multimedia program for approximately twenty-five minutes, they were given the whole questionnaire.

(3) *Transfer.* If NSG-participants understood the program on a deeper level then they should have an advantage in

understanding a similar situation. So we asked them to imagine a scenario in which four tribes were deciding whether to form alliances. Analogous to World War 1, between some tribes there were permanent conflicts over resources and between some there were no fundamental problems. Participants took the role of one tribe's leader and decided whether to form an alliance, then justified their answer. The arguments of participants agreeing to make an alliance were classified into two categories: *security* ("It's more secure to have partners."), and *nationalism* ("Having a partner gives my tribe more power."). Participants who disagreed with making an alliance had their arguments classified as either *war avoidance* ("My tribe has to help the partner in a conflict even if it is not our conflict."), or *egoism* ("I don't want to fight for others."). The participants' statements could be assigned to categories with an inter-rater reliability of Cohen's (1960)  $\kappa = .94$ . Nationalism and egoism explanations were considered less interesting as they were only seldom mentioned in an unpublished pilot study.

## Results

**Preliminary Analyses.** Our intention was to examine the process of learning by measuring the variables motivational state and sampled knowledge, so we interrupted learning three times. However, motivational state during learning stayed constant (motivational state at time point 1:  $M = 5.42$ ,  $SD = 0.76$ ; at time point 2:  $M = 5.49$ ,  $SD = 0.82$ ; at time point 3:  $M = 5.52$ ,  $SD = 0.90$ ;  $F[2, 86] = 0.84$ ,  $p = 0.44$ ) and at a high level (scale from 1 = low motivational state to 7 = high motivational state). There was no feedback given to participants, but this result suggests that learners did not experience success or failure while working with the program. If they had done so, then motivation would probably have increased or decreased, as Vollmeyer, Rheinberg, and Burns (1998) showed.

The second variable was sampled knowledge. As a consequence of the design this measure should not change over time as participants always received questions about the last three pages they saw. As the pages were not cumulative, answers to these questions cannot reflect accumulation of knowledge. Instead they sample how well the participants were learning as they worked through the program. For each set of three questions learners scored between 0 and 3. We also tested whether knowledge was constant during learning. Because we expected no difference between knowledge sampled at different points, we averaged the knowledge scores as well as the motivational states.

**Effects of Goal Specificity.** The first aim of our study was to test the hypothesis that learners with a specific goal acquired less knowledge than participants with a nonspecific goal. Achieving a worse outcome may be a result of poor strategies and/or motivation, which were measured as mediating variables.

The most important measures were knowledge during learning and accumulated knowledge, as these should demonstrate that giving participants specific goals decreased

their learning performance. For knowledge during learning, the possible range for correct answers was 0 to 9 (three times questions about three pages). As expected, the SG-group could not answer correctly as many questions as the NSG-group,  $t(43) = 3.34$ ,  $p = .002$ , as the means in Table 1 show. The same effect was found on the multiple-choice questions after learning with the program. The answers were adjusted in that answers were only analyzed to questions about pages participants had actually seen. Thus the means in Table 1 show the proportion of correct answers to pages seen. For the factual questions, the NSG-group answered 44%, compared to 29% for the SG-participants,  $t(43) = 4.18$ ,  $p < .001$ . NSG-participants were also better on inference questions,  $t(43) = 3.20$ ,  $p = .003$ . Even if we had not applied the adjustment, the effect for goal specificity on knowledge for facts holds (SG:  $M = 8.96$ ,  $SD = 3.62$ ; NSG:  $M = 12.67$ ,  $SD = 3.34$ ,  $t[43] = 3.56$ ,  $p = .001$ ) as well as for inferences (SG:  $M = 3.08$ ,  $SD = 1.21$ ; NSG:  $M = 4.52$ ,  $SD = 1.86$ ,  $t[43] = 3.11$ ,  $p = .003$ ).

Table 1: Descriptive statistics for the SG-group ( $n = 24$ ) and the NSG-group ( $n = 21$ ) on process and dependent variables.

	Groups	M	SD
sampled knowledge	SG	3.79	1.74
	NSG	5.52	1.72
knowledge (facts)	SG	0.29	0.12
	NSG	0.44	0.13
knowledge (inference)	SG	0.23	0.12
	NSG	0.37	0.16
number of pages	SG	46.58	13.72
	NSG	39.43	9.83
Interest	SG	4.93	1.32
	NSG	4.93	0.99
Challenge	SG	4.96	1.00
	NSG	4.58	0.89
probability of success	SG	5.47	0.85
	NSG	5.45	1.04
Fear	SG	2.68	1.10
	NSG	3.06	1.38
motivation during learning	SG	5.47	0.70
	NSG	5.49	0.86
sequence	SG	0.71	0.01
	NSG	0.73	0.01

We found a clear effect of goal specificity on learning outcomes, so we tested whether our goal specificity manipulation affected process variables. As Table 1 shows, SG-participants looked at more pages,  $t(43) = 1.98$ ,  $p = 0.054$ . Although this effect was not quite significant, we analyzed the process and dependent variables by relating them to the number of pages looked at.

We did not expect any effect of goal specificity on initial motivation. As the means in Table 1 indicate, none of the four factors of initial motivation (interest, challenge, probability of success, fear) differed,  $p$ 's  $< 0.15$ . As the items were answered on a scale from 1 to 7, the means

demonstrate that all participants regarded the task as easy (probability of success:  $M = 5.26$ ,  $SD = 0.93$ ), interesting (interest:  $M = 4.93$ ,  $SD = 1.17$ ), challenging (challenge:  $M = 4.78$ ,  $SD = 0.95$ ), and that it aroused few fears (fear:  $M = 2.85$ ,  $SD = 1.24$ ).

As process variables, motivational state and strategy systematicity were measured. The motivational state measure (averaged over three time points) did not differentiate the SG-group from the NSG-group,  $t(43) = 0.11$ ,  $p = 0.92$ . The indicator for how systematically learners work with the multimedia program was the proportion of pages that followed logically from the previous page. Although we had expected that the SG-group would jump around more in the program, we could not find a difference,  $t(43) = 0.83$ ,  $p = .41$ . Across both groups, participants chose the next logical page 73% of the time; that is, they followed the sequence for the country they were reading about.

We expected that NSG participants faced with our transfer task would be less likely to agree to an alliance that risked a wider war. Therefore, we tested whether the groups differed in who would agree to an alliance and what argument they would use. As Table 2 shows, there was a tendency for more SG-participants than NSG-participants to agree to an alliance,  $\chi^2(1) = 3.27$ ,  $p = .071$ .

Table 2: Number of participants in SG- and NSG-group agreeing to an alliance.

	agreement to alliance	
	alliance	no alliance
SG	12	12
NSG	5	16

When we categorized the arguments used to justify whether to enter an alliance, we dropped the categories of *nationalism* (one in each group) and *egoism* (7 in SG, 6 in NSG) from the analysis. These two categories are unlikely to be influenced by the program. Of the remaining two categories, we found that war avoidance (the theme our participants would be likely to use if they understood the program) was used as a justification more often by NSG participants than SG participants, who instead were more likely to use security as a justification,  $\chi^2(1) = 4.82$ ,  $p = .028$  (see Table 3). Thinking in terms of security might be a more surface reaction to the event described in the program.

Table 3: Distribution of the SG- and NSG-group which arguments were given for or against alliance.

	argument pro/con alliance	
	war avoidance	security
SG	5	11
NSG	10	4

**Cognitive-motivational Model.** The aim of the study was to explore whether the concept of goal specificity could be applied to a more realistic multimedia task. Previously we have looked at how motivation may affect the learning process in different ways depending on the goal (Vollmeyer,

et al., 1998). However, the sample size in this preliminary study is not large enough for the type of path analyses required to examine this issue. Therefore, a first step was to look for differences between groups in term of which variables affected accumulated knowledge. As knowledge for facts and inferences were correlated,  $r = 0.66$ ,  $p < 0.001$ , and there were fewer inference questions than fact questions, we will only report knowledge for facts. Number of pages also did not relate to motivation. Correlations of final knowledge with initial motivation and process variables are presented in Table 4.

Independent of the manipulation, sampled knowledge and strategy systematicity (i.e., *sequence*) should be positively correlated with accumulated knowledge. Whereas for sampled knowledge a relationship to accumulated knowledge was found, this was not the case for *sequence*. The latter result makes doubtful whether *sequence* is a valid indicator for strategy systematicity.

We examined Schiefele's (1996) proposal that interest and challenge could play a more important role for NSG learners. Table 4 shows that interest and challenge correlate with accumulated knowledge for the NSG-group but not for the SG-group. However, the difference in correlations is only significant for challenge ( $z = 2.59$ ,  $p = 0.014$ ).

To analyze which variable was the best predictor for accumulated knowledge, we calculated a regression on the four initial motivational factors, motivation during learning and sampled knowledge. As the correlations had shown different patterns in the experimental groups, they were analyzed separately. For the SG-group the best predictor for knowledge is probability of success,  $\beta = 0.46$ ,  $t = 2.56$ ,  $p = 0.018$ ,  $R^2 = .23$ , whereas for the NSG-group challenge,  $\beta = 0.61$ ,  $t = 3.42$ ,  $p = 0.003$ ,  $R^2 = .27$ , is the best.

Table 4: Correlations  $r(p)$  of accumulated knowledge with initial motivation and process variables separated in SG- ( $n = 24$ ) and NSG-groups ( $n = 21$ ).

	Correlations with accumulated knowledge	
	SG	NSG
sampled knowledge	<b>0.35</b> (0.10)	<b>0.46</b> (0.03)
interest	0.02 (0.92)	<b>0.51</b> (0.02)
challenge	-0.23 (0.28)	<b>0.51</b> (0.02)
probability of success	<b>0.48</b> (0.02)	0.31 (0.17)
fear	<b>-0.45</b> (0.03)	-0.31 (0.17)
motivation during learning	0.19 (0.37)	<b>0.39</b> (0.08)
sequence	-0.11 (0.62)	-0.14 (0.55)

## Discussion

The major result of our study was that we could demonstrate an effect of goal specificity on a more realistic task than those on which this type of research has previously been based. Switching from a problem-solving task to multimedia learning in the domain of history presented two difficulties: the first was that goal specificity had to be re-interpreted, the second was that the theoretical constructs needed new operationalizations.



**Goal Specificity.** With a new task and a different operationalization of goal specificity we could replicate the effect that an NSG group would learn more than a SG group, even when the task was to learn about the causes that led to the outbreak of World War 1. Therefore we met our aim of generalizing the concept of goal specificity from problem solving to multimedia learning.

**Goals and Motivation.** Compared to our research on problem solving, motivation during the task seemed to play a different role when working with a multimedia program. As our previous task was difficult to solve, motivation changed while working with that task. When working with this shorter and easier program no changes of motivation during learning were observed, but initial motivation was more predictive than was previously found. However, what motivational factors were most influential varied with goal specificity. For NSG interest and challenge were most important, but for SG fear and probability of success were. This provides further evidence that the way motivation affects performance varies with the type of goal.

**Operationalizations of Theoretical Constructs.** In a problem solving task every input can be categorized on a continuum of systematicity, but with a multimedia program it is not clear how to operationalize this variable. Even if learners can choose after every page what to see next, it is unclear whether choosing the next page in the sequence is a systematic strategy as we defined it. As our measure does not correlate with performance it is still open as to whether this is a valid measure. However, we have shown that this is a useful task for examining goal specificity, so it should be worthwhile in future research to try to develop and validate better measures of strategy.

### Acknowledgements

This research was supported by the German-American Academic Council to Regina Vollmeyer and Bruce Burns.

### References

- Chen, C. & Rada, R. (1996). Interacting with hypertext: A meta-analysis of experimental studies. *Human-Computer Interaction, 11*, 125-156.
- Chi, M.T.H., Bassok, M., Lewis, M.W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 18*, 439-477.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.
- Geddes, B.W. & Stevenson, R.J. (1997). Explicit learning of a dynamic system with a non-salient pattern. *The Quarterly Journal of Experimental Psychology, 50A*, 742-765.
- Issing, L.J. & Klimsa, P. (1997). *Information und Lernen mit Multimedia* [Information and learning with multimedia] (2nd ed.). Weinheim: Psychologie Verlags Union.
- Kanfer, R. & Ackerman, P.L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74*, 657-690.
- Klahr, D. & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1-55.
- Locke, E.A. & Latham, G.P. (1990). *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice Hall.
- Miller, C.S., Lehman, J.F., & Koedinger, K.R. (1999). Goals and learning in microworlds. *Cognitive Science, 23*, 305-336.
- Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21*, 1-29.
- Renkl, A. (1999). Learning mathematics from worked-out examples: Analyzing and fostering self-explanations. *European Journal of Psychology of Education, 14*, 477-488.
- Rheinberg, F. & Vollmeyer, R. (in press). Sachinteresse und leistungsthematische Herausforderung [Topic interest and achievement thematic challenge]. In U. Schiefele & K.P. Wild (Eds.). *Lernumgebung, Lernmotivation und Lernverhalten*. Münster: Waxmann.
- Royer, J.M., Carlo, M.S., Dufresne, R., & Mestre, J. (1996). The assessment of levels of domain expertise while reading. *Cognition and Instruction, 14*, 373-408.
- Schiefele, U. (1996). *Motivation und Lernen mit Texten* [Motivation and text learning]. Göttingen: Hogrefe.
- Simon, H.A. & Lea, G. (1974). Problem solving and rule induction: A unified view. In L.W. Gregg (Ed.), *Knowledge and cognition* (pp. 105-127). Hillsdale, NJ: Erlbaum.
- Sweller, J. (1988). Cognitive load during problem solving: Effects of learning. *Cognitive Science, 12*, 257-285.
- Tubbs, M.E. (1986). Goal setting: A meta-analytic examination of the empirical evidence. *Journal of Applied Psychology, 71*, 474-483.
- Vollmeyer, R., Burns, B.D., & Holyoak, K.J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20*, 75-100.
- Vollmeyer, R. & Rheinberg, F. (1998). Motivationale Einflüsse auf Erwerb und Anwendung von Wissen in einem computersimulierten System [Motivational influences on the acquisition and application of knowledge in a simulated system]. *Zeitschrift für Pädagogische Psychologie, 12*, 11-23.
- Vollmeyer, R., Rheinberg, F., & Burns, B.D. (1998). Goals, strategies, and motivation. In M.A. Gernsbacher & S.J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 1090-1095). Hillsdale, NJ: Erlbaum.

# Human Belief Revision and the Order Effect

Hongbin Wang (Hongbin.Wang@uth.tmc.edu)

Jiajie Zhang (Jiajie.Zhang@uth.tmc.edu)

Todd R. Johnson (Todd.R.Johnson@uth.tmc.edu)

Department of Health Informatics<sup>1</sup>

University of Texas - Houston Health Science Center

7000 Fannin, Suite 600

Houston, TX 77030

## Abstract

The order effect, a phenomenon in which the final belief is significantly affected by the temporal order of information presentation, is a robust empirical finding in human belief revision. This paper investigates how order effects occur, on the basis that human belief has a coherence foundation and a probability/confidence distinction. Both the experimental results and the UEcho modeling suggest that confidence plays an important role in human belief revision. Order effects in human belief revision occur where confidence is low and disappear when confidence increases. UEcho provides a computational model of human belief revision and order effects

## Introduction

It is generally agreed that one constantly conducts *belief revision* – a process in which one revises one’s beliefs in the light of new information, with a goal to maintain a reasonably consistent and up-to-date belief system. It is of great philosophical and psychological interest to investigate whether one is able to achieve such a goal and what the underlying regularities are.

Psychological investigations of human belief revision have revealed an important finding – *the order effect* (e.g., Hogarth & Einhorn, 1992; Schlottmann & Anderson, 1995; Zhang, Johnson, & Wang, 1997). Generally speaking, the order effect refers to the phenomenon that the temporal order in which information is presented affects the final judgment of an event. Undoubtedly, the temporal order of incoming evidence often carries important information about the true meaning of an event. However, robust order effects have been found even in situations where the temporal order of incoming evidence seems not meaningful. It is these cases that make the order effect a very interesting phenomenon.

This paper aims to investigate how order effects occur in human belief revision, both empirically and computationally. It consists of four sections. In the first section, some previous studies on human belief and belief revision, uncertainty, and order effects are briefly reviewed. Then a psychological experiment and its UEcho modeling (see Wang, Johnson, Zhang; 1998; Wang, 1998) are reported in the next

two sections. The final section provides general discussions and conclusions.

## Human Belief Revision and Uncertainty

There are two main views regarding how an unconvinced belief could be justified (e.g., Gardenfors, 1990). According to the *foundations* approach, a rational individual derives beliefs from reasons for these beliefs. In other words, a belief is justified if and only if it possesses some satisfactory and “hard” underlying reasons. The *coherence* approach, in contrast, maintains that a belief may be held independent of its supporting reasons. An individual holds a belief as long as it logically coheres with the individual’s other beliefs. Therefore, coherent beliefs can mutually justify each other, and no belief is more fundamental than another.

How beliefs are justified has a direct implication on how beliefs should be revised when new information becomes available. Based on the foundations view, one should simply give up those beliefs that lose their underlying reasons and accept new beliefs that become well supported. An example is the *Truth Maintenance System* developed by Doyle (1979). In contrast, the coherence view emphasizes consistency and conservatism. Therefore, in belief revision one should retain as many of one’s beliefs as possible while accommodating any new evidence. In other words, as long as the coherence of the resulting state is maintained, a belief can survive without solid reasons. The so-called AGM theory of belief revision (Alchourron, Gardenfors and Makinson, 1985; Gardenfors, 1990) is one well-known example that adopts the coherence approach.

The coherence approach to human belief revision is generally preferred (see Gardenfors, 1990; Thagard, 1989). It has been argued that the foundational approach involves excessive computational cost. It is intellectually very costly to keep track of the reasons of beliefs. Moreover, it has been shown that the foundational approach conflicts with observed human behavior. For example, the *belief preservation effect* (e.g., Ross & Lepper, 1980) suggests that people are reluctant to give up some beliefs even when the original evidential bases of these beliefs are completely destroyed.

---

<sup>1</sup> Portions of this research were conducted when the authors were at The Ohio State University, Departments of Psychology (Drs. Wang and Zhang) and Pathology (Dr. Johnson).

Uncertainty is the ultimate reason for human belief and its revision. It is well agreed that there are two general types of uncertainty (see Walley, 1991). First, when the truth of a proposition is unknown but the average proportion of that proposition being true in the long run can be precisely specified, the indeterminacy involved in this case is called *uncertainty*. An example is tossing a fair coin. Second, in some cases, one can neither completely determine the truth of a proposition nor precisely specify the average proportion of that proposition being true in the long run. This type of uncertainty – the indeterminacy of the average behavior – is usually called *imprecision*.

The distinction between imprecision and uncertainty is so fundamental that it has caused a “holy war” in the field of uncertainty management. On the one hand, *probability theory* (along with Bayes’ Theorem for belief revision), the best-established formal method for uncertainty management, has long been criticized for its difficulty in handling imprecision. It has been suggested that while a probability number is sufficient to summarize the uncertainty dimension, a confidence measure is needed to handle the imprecision dimension, with a high confidence measure representing precise belief and a low confidence measure representing imprecise belief (see Almond, 1995). On the other hand, *fuzzy sets* and the *possibility theory* (see Zadeh, 1978) often deal with imprecision but not uncertainty. The *theory of belief functions* (see Shafer, 1976) deals with both imprecision and uncertainty. Along with Dempster’s rule for evidence combination, it thus provides a more complete picture of formal belief management.

### The Order Effect

A large number of empirical studies on human reasoning have demonstrated that people often systematically deviate from normative postulates. With the assumption that these normative postulates prescribe how a reasonable individual should behave, these systematic deviations are often labeled as *cognitive illusions, biases, or fallacies* (e.g., Kahneman, Slovic, & Tversky, 1982). Several well-known biases include *base rate fallacy, conjunction fallacy, and overconfidence* (see Kahneman & Tversky, 1996 for a review).

The order effect in human belief revision is yet another robust empirical finding (e.g., Hogarth & Einhorn, 1992). By a similar standard, the order effect should also be called a bias since the normative postulates, in particular Bayes’ Theorem, have no room for it – it simply violates commutativity. However, as many researchers have already pointed out, calling it a bias is nothing more than giving it a label, which provides no help to understand how and why the order effect occurs.

Miller and Campbell (1959) argue that order effects in belief revision represent order effects in memory. Specifically, due to memory decay, previous evidence items get weighted less as time goes by. Later studies showed that this view is problematic since direct comparisons suggest that beliefs are largely independent of recall of evidence items (e.g., Anderson & Hubert, 1963).

The serial integration model (e.g., Schlotzmann & Anderson, 1995), proposed in the framework of information integration theory (Anderson, 1981), claims that people pay less

attention to successive items of evidence due to attention decrement. Attention decrement results in different weights being assigned to different evidence items, which in turn results in order effects. Unfortunately, this model fails to specify what factors affect the attention decrement.

Hogarth and Einhorn (1992) proposed an anchoring and adjustment model to explain order effects. According to this model, belief revision is a sequential anchoring-and-adjustment process in which people adjust the current belief (the anchor) on the basis of how strongly new information confirms or disconfirms this belief. In addition, the adjustment weight is a function of both the anchor and the new evidence. More specifically, when the impact of the new evidence is smaller than the reference point, the adjustment weight is proportional to the anchor. And when the impact of the evidence is larger than the reference point, the adjustment weight is inversely proportional to the anchor. It is this kind of contrast effect that results in order effects. The model further adopts two parameters ( $\alpha$  and  $\beta$ ) to regulate this weight assignment process. It claims that the two parameters represent people’s sensitivity toward negative and positive evidence, respectively. In particular, the model argues that some individuals tend to view negative (or positive) evidence more seriously than others. Therefore, in terms of the underlying factors that regulate the weight assignment, the model actually points to unidentifiable individual differences.

### Summary

The above review reveals two important findings in the area of human belief and its revision. First, human belief has a coherence foundation. A belief can survive without solid foundational evidence. Beliefs hold each other as a coherent system. Second, human belief has a multi-component structure. The probability/confidence distinction suggests that a single probability number cannot capture all the important aspects of a belief. A confidence component is necessary.

Previous theories of order effects hardly take these findings into consideration. They often attempt to explain order effects by a weight assignment mechanism that weighs members of the evidence sequence differentially. However, they encounter great difficulties in fully explaining why weights have to be assigned in a particular way at a particular time. Consequently, in some cases, one or more task characteristics are particularly emphasized (e.g., memory decay, or attention decrement), which of course often only account for a fragment of the order effect. In some other cases, arbitrary parameters are adopted in the weight assignment to summarize unidentifiable sources.

The probability/confidence distinction suggests that the impact of the new evidence cannot be fully understood without the nature of the current beliefs being sufficiently appreciated. More specifically, the confidence component of a belief, mainly determined by the amount of previous experience, represents how easily this belief can be revised. A belief with no previous experience has very low confidence and is easiest to change. And a belief established by significant previous experience is committed with a high confidence level and thus is hard to change. In the context

of order effects, this analysis implies that the order effect pattern may change with different levels of experience. The rationale is as follows. As one keeps interacting with the environment, one gains more and more experience. As a result, beliefs are gradually tuned to the statistical structure of the environment (see Anderson, 1990). In addition, confidence increases as more experience is gained. Both factors will make one react to any new evidence more realistically rather than over-react or under-react. Since over-reacting and under-reacting are the fundamental causes of the order effect, then when one gains more and more experience about the environment, the order effect in belief revision should tend to diminish and disappear.

The experiment reported in the next section is designed to test this hypothesis.

## Experiment

### Design and Procedure

A modified version of the CIC (Combat Information Center, Towne, 1995; see also Wang, Johnson, & Zhang, 1998; Zhang, Johnson, & Wang, 1997) simulation was used as the task domain. In the CIC task used for this experiment, the goal of the participant, acting as a commanding officer of a naval ship, was to collect two pieces of information sequentially about an aircraft in the radar area and accurately identify its intention.

One piece of information was about the route (R), which indicates the target is either on or off a commercial air route. The other piece of information was self-identification (SelfID), which indicates the target's response after being warned. In a typical trial, the participant was shown a target and had to report the degree of belief (on a 0-100 scale) that the target is friendly before any evidence (i.e., initial belief) and after each piece of evidence (i.e., sequential belief revision). Finally, the participant was forced to make a two-alternative (i.e., friendly or hostile) judgment about the identity of the target. After the decision was made, the participant could request the true identity of the target if available. Whether this true identity information is available or not depends on the type of the trial, as explained later.

The experiment adopted a 3x(4) factorial design. The between-subject independent variable was the ratio of total friendly targets to total hostile targets in the training samples. The ratio was 1:1 (equal number of friendly and hostile targets), 3:1 (friendly targets are three times as frequent as hostile targets), or 1:3 (hostile targets are three times as frequent as friendly targets). The purpose of this factor was to create environments with different statistical structures and test if participants could gradually tune their beliefs to capture these structures.

The experiment attempted to investigate how the patterns of order effects changed with training. The training was organized in four blocks, which is the within-subject variable (see Figure 1). Five evaluation blocks were inserted in the process to provide a way to easily evaluate the pattern changes of order effects. The major difference between training trials and evaluation trials is that no true identity feedback was provided at the end of each evaluation trial.

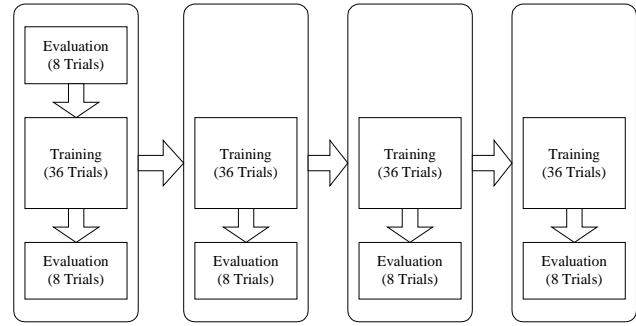


Figure 1. The experimental design

Each evaluation block had eight evaluation trials in it. The eight evaluation trials were constructed in the following way. There were two pieces of evidence (Route and SelfID), each of which had two possible values (“on” and “off” for Route, and “friendly response” and “no response” for SelfID), so there were 4 kinds of trials. Since each piece of evidence could be collected before the other, we had a total of eight different evaluation trials. Participants were instructed to summarize their training experience in order to perform these evaluation trials.

Each training block consisted of 36 trials. The trial distribution is dependent on the friendly-hostile ratio and is shown in Table 1. Since a value of “on” for Route and a value of “friendly response” for SelfID are regarded as positive evidence for a friendly target, they are represented by “+”s in Table 1. Similarly the opposite values are represented by “-”s.

Table 1. The trial distribution

Route	SelfID	1:1		3:1		1:3	
		F	H	F	H	F	H
+	+	8	2	12	1	4	3
+	-	4	4	6	2	2	6
-	+	4	4	6	2	2	6
-	-	2	8	3	4	1	12
Total		18	18	27	9	9	27

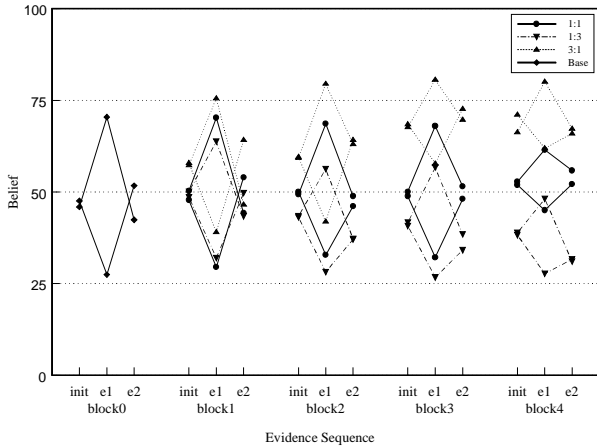
140 undergraduate students participated in the experiment. They were randomly assigned to the three friendly-hostile-ratio treatment groups. The trials in each block were completely randomized for each participant.

### Results

The five evaluation blocks, distributed in the critical positions in the training, are the focus of our analysis. In addition, for the purpose of easily examining order effects, only the data from the two critical evidence sequences (“+-“ and “-+”) are reported. The results are shown in Figure 2.

Three major findings are identified. First, the effect of the friendly-to-hostile ratio is evident. While the average initial belief judgment (i.e., before any evidence) tends to increase with training in the 3:1 group (56.7, 59.0, 67.1, 68.6, from block1 to block4, respectively), it tends to decrease with training in the 1:3 group (47.8, 43.0, 41.3, 40.2,

from block1 to block4, respectively). Note that it is largely unchanged with training in the 1:1 group (50.0, 50.0, 49.5, 53.3, from block1 to block4, respectively). This pattern of result suggests that the initial belief judgments were gradually tuned to more closely reflect the built-in friendly-to-hostile ratios.



**Figure 2.** The belief revision patterns in all three friendly-to-hostile ratio conditions. The initial evaluation block before any training is labeled as block0, which also combines data from all three ratio groups. The evaluation blocks after each training block are labeled as block1 to block4, respectively. In each block, belief evaluation (from 0 to 100) is plotted against the evidence sequence, from init (before any evidence is presented) to e1 (the first piece of evidence is presented) to e2 (the second piece of evidence is presented). Because in general positive evidence raises belief ratings and negative evidence lowers belief ratings, plotting opposite evidence sequences (“+−” and “−+”) together results in a diamond shape (e.g., block4). Importantly, when the final belief ratings after both pieces of evidence are different, the diamond shape becomes the fish-like shape (e.g., block0), which indicates a recency order effect.

Second, the belief revision patterns change significantly across the whole training session. A recency effect is evident in block0 (the final belief judgment is 41.9 for the “+−” sequence vs 50.8 for the “−+” sequence), and this recency effect tends to disappear in the later blocks. More specifically, recency effects appear in block1 and disappear in block3 and block4 in all three ratio groups. This pattern is consistent with our prediction that order effects diminish and disappear with training.

Finally, it is interesting to note that the areas inside the diamond-like order effect patterns tend to become systematically smaller as the training progresses. Since the pattern is approximately symmetric vertically, we could use the height of the diamond as a rough estimation of the size of the area. The result shows that the area size decrement is statistically significant in the 1:1 and 3:1 groups, though not in the 1:3 group. This pattern of area decrement indicates

that participants fluctuated less in their belief judgments as more experience was gained, which further suggests that participants tended to be less sensitive to new evidence as confidence goes up.

In summary, the experiment results reveal that the recency effect disappeared as more training trials were performed. The disappearance of the recency effect suggests that instead of over-reacting in the light of new evidence, participants made more proper and more realistic reactions. As suggested previously, as more experience was acquired during training, the statistical tuning led participants to make more confident belief judgments, which eliminated over-reaction.

UEcho, first proposed in Wang, Johnson, and Zhang (1998) as a model of belief evaluation in abduction, is further developed to model the experiment results.

### A UEcho Model

UEcho is based on Echo, which is a connectionist implementation of the Theory of Explanatory Coherence (TEC), proposed by Thagard (1989, 1992) as a model of human abductive reason. Different from other theories of belief revision such as Hogarth & Einhorn’s anchoring and adjustment model, Echo takes a coherence view of belief evaluation as its foundation. According to Echo, a belief should be accepted if it is coherent with other beliefs, and rejected if it is incoherent with other beliefs. By quantitatively defining (explanatory) coherence, an Echo system pursues highest coherence by considering all related beliefs in a holistic manner. When the system converges, the most believable hypothesis set will defeat any competitors and pop out.

Although Echo has gained much empirical support, they have serious limitations (e.g., Wang, Johnson, & Zhang 1998): (1) Echo does not handle sequential belief revision; (2) Echo does not learn from experience; and (3) Echo does not distinguish confidence and probability. All these limitations cast doubt on Echo as a general model of human belief revision.

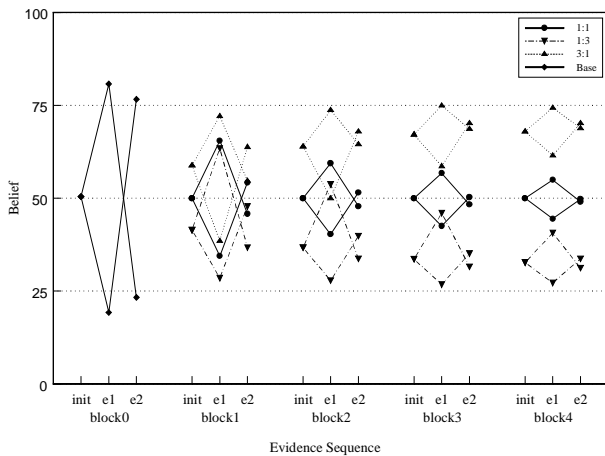
Wang, Johnson, and Zhang (1998) proposed UEcho (“U” for Uncertainty) as an extension of Echo to address the first two problems. They have shown that UEcho is able to model order effects. UEcho is further extended here to embed the probability/confidence distinction. By doing so, we expect that UEcho, as a coherence-based model of belief evaluation, provides an alternative model of human belief revision that is more plausible than the traditional weight-assignment-based integration models.

UEcho maintains that the activation of a node determines acceptability, thus representing the probability component of a belief. UEcho adopts three mechanisms to add a confidence dimension to the system.

All three mechanisms try to tune critical parameters based on previous experience. The first parameter is the parameter of skepticism  $\theta$ . In Echo,  $\theta$  represents the decay rate in the activation updating. The higher  $\theta$  is, the faster does the node activation decay. Confidence cures skepticism. Gradually tuning down  $\theta$ , based on experience, is a natural way to represent confidence. The second mechanism has to do with the parameter  $\alpha$  and  $\beta$  in the anchoring and adjustment model. As mentioned earlier,  $\alpha$  and  $\beta$  represents one’s

sensitivity toward negative and positive evidence, respectively. Although the anchoring and adjustment model attributes the sensitivity to some unidentifiable factors of personality, the two parameters are functionally closely related to confidence in the sense that as confidence goes up, the sensitivity to new evidence goes down. Incorporating and gradually tuning  $\alpha$  and  $\beta$  represents another aspect of confidence management in UEcho. Finally, UEcho extends Echo's parameter of *data excitation*. In Echo, data excitation is used to represent the assumption that observed data nodes have independent support of their own. The hypothesis nodes have no associated data excitation. By generalizing this parameter to hypothesis nodes, UEcho enables hypothesis nodes to learn and remember their activation values, thus to gradually gain self-support (or dis-support) of their own, based on past experience. For a detailed description of these tuning mechanisms, please see Wang (1998).

The exact same design and procedure was used to train a UEcho network, and the corresponding simulation results are shown in Figure 3.



**Figure 3.** The belief revision patterns in all three friendly-to-hostile ratio conditions, based on the UEcho simulation.

The modeling results match the experiment results remarkably well. First, the gradual separation of the curves of the three ratio groups nicely reflects the statistical tuning toward the built-in environmental friendly-to-hostile ratios. From block1 to block4, the average initial belief judgment is 50.0, 50.0, 50.0, 50.0 for the 1:1 group, 58.8, 63.9, 67.2, 69.0 for the 3:1 group, and 41.4, 36.7, 33.6, 32.7 for the 1:3 group, respectively. Second, the order effect pattern change is evident. A recency effect is significant in block0 (23.3 for “+−” vs 76.7 for “−+”). The magnitude of the recency effect, measured as the difference between the final judgment in “+−” and the final judgment in “−+”, decreases significantly from block1 to block4. More specifically, they are 8.4, 3.6, 2.0, 0.8 for the 1:1 condition, 9.1, 3.5, 1.6, 1.3 for the 3:1 condition, and 11.1, 6.1, 3.6, 2.5 for the 1:3 condition, respectively. Finally, the areas inside the diamond

shapes become systematically smaller with training as well, indicating the fluctuation in belief revision tends to be smaller as the training progresses.

In summary, by embedding the probability/confidence distinction, UEcho is capable of capturing the changes of order effect patterns at different experience levels. The close match between the simulation results and the experimental results in the decrement and disappearance of order effects with the increase of experience supports UEcho as a model of coherence-based and complex human belief revision.

## Discussions and Conclusions

Human belief and human belief revision are ubiquitous in everyday life and scientific discovery. The order effect, a phenomenon in which the final belief is significantly affected by the temporal order of evidence is a robust empirical finding in human belief revision. The order effect is generally regarded as a manifestation of human biases and an indication of human irrationality. It is the goal of this paper to study how the order effect occurs.

Previous research leads to the conclusion that human belief has a coherence foundation and consists of multiple components. Such a conclusion motivates and guides both the experimental study and the computational modeling work described in the paper. Both the experimental results and the UEcho modeling results show that order effects in belief revision exist at the early stage of training when the confidence level is low and they tend to diminish and disappear later when the confidence increases.

It is interesting to further speculate how the UEcho modeling results could tell us the possible rational basis of order effects. First of all, the fact that UEcho, which is based on rational postulates and intended to prescribe what people should do, naturally shows order effects (when the confidence level is low) convincingly “debiases” order effects. Second, the existence of order effects has ecological implications. UEcho reveals that order effects appear when the relevant experience is scarce, and order effects disappear when the relevant experience becomes rich. When the relevant experience is rich, one has confident expectations, which eliminate the need to over-react. When the relevant experience is scarce, one has to sufficiently appreciate every single piece of information since its relevance cannot be easily and accurately determined in the first place. In this sense, both the existence and the disappearance of the order effect are rational.

It should be noted that this study involves only the recency effect. It would be of great importance to explore how it can be tuned or extended to model the primacy effect. Whether it can model the full range of order effects using the same mechanism is a strong test for UEcho as a general model of human belief revision.

What does the current study say about human rationality in general? For a long time, the order effect, along with various other heuristics and biases (Tversky & Kahneman, 1974), has been taken as a demonstration that people systematically deviate from rationality. This view has been greatly challenged recently. Beyond philosophical debates, systematic investigations have been carried out to determine

the conditions under which the biases appear or disappear. For example, Gigerenzer (1991, 1994, 1996), among others, has shown that while people perform poorly in assessing subjective probability they assess relative frequencies reasonably well. Since using/reporting subjective probability is not something people are equipped with, “biases are not biases” (Gigerenzer, 1991, page 86), and heuristics are meant to explain something that does not exist. It has been demonstrated that all the biases, including the base rate fallacy, conjunction fallacy, and overconfidence, disappear or are significantly reduced when information is presented to participants in frequency format (e.g., 10 out of 100) instead of single-event subjective probability format (e.g., 10%) (Gigerenzer & Hoffrage, 1995). Noting that normative postulates often assume a stationary and discrete environment, many researchers have argued that the environment is neither stationary nor discrete. People may appear biased or deficient according to those normative postulates, but they are in fact very functional and optimal when a continuous and dynamically changing environment is assumed (e.g., Jungermann, 1983). The current study provides another example to show that this might be the case.

### Acknowledgements

This work is funded by Office of Naval Research Grant No. N00014-95-1-0241.

### References

- Alchourron, C., Gardenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction functions and their associated revision functions. *Journal of Symbolic Logic*, 50, 510-530.
- Almond, R. G. (1995). *Graphical belief modeling*. London: Chapman & Hall.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Press.
- Anderson, N. H. (1981). *Foundations of information integration theory*. New York, NY: Academic Press.
- Anderson, N. H., & Hubert, S. (1963). Effects of concomitant verbal recall on order effects in personality impression formation. *Journal of Verbal Learning and Verbal Behavior*, 2, 379-391.
- Doyle, J. (1979). A truth maintenance system. *Artificial Intelligence*, 12, 231-272.
- Gardenfors, P. (1990). The dynamics of belief systems: Foundations vs coherence theories. *Revue Internationale de Philosophie*, 172, 24-46.
- Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond “heuristics and biases”. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology*, Vol 2. Chichester, England: Wiley.
- Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is relevant for psychology and vice versa. In G. Wright & P. Ayton (Eds.), *Subjective probability*. New York, NY: Wiley.
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A rebuttal to Kahneman and Tversky (1996). *Psychological Review*, 103(3), 592-596.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency format. *Psychological Review*, 102(4), 684-704.
- Hogarth, R.M. & Einhorn, H.J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24, 1-55.
- Jungermann, H. (1983). The two camps of rationality. In R. W. Scholz (Ed.), *Decision making under uncertainty*. Amsterdam: North-Holland.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and Biases*. Cambridge, NY: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, 103(3), 582-591.
- Miller, N., & Campbell, D. T. (1959). Recency and primacy in persuasion as a function of the timing of speeches and measurement. *Journal of Abnormal and Social Psychology*, 59, 1-9.
- Quine, W. V., & Ullian, J. S. (1978). *The web of belief*. New York, NY: Random House.
- Ross, L., & Lepper, M. R. (1980). The perseverance of beliefs: Empirical and normative considerations. In R. A. Shweder (Ed.), *Fallible judgment in behavioral research: New directions for methodology of social and behavioral science*, Vol. 4. San Francisco, CA: Jossey-Bass.
- Schlottmann A., & Anderson, N. H. (1995). Belief revision in children: Serial judgment in social cognition and decision-making domains. *Journal of Experimental Psychology: LMC*, 21(5), 1349-1364.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.
- Thagard, P. (1989). Explanatory Coherence. *Behavioral and Brain Sciences*, 12(3), 435-502.
- Thagard, P. (1992). *Conceptual revolutions*. Princeton, NJ: Princeton University Press.
- Towne, D. (1995). CIC: Tactical Decision Making (Version 2.0). Unpublished manuscript. Behavioral Technology Laboratories, University of Southern California.
- Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. New York, NY: Chapman and Hall.
- Wang, H. (1998). *Order effects in human belief revision*. Ph.D. Dissertation, The Ohio State University.
- Wang, H., Johnson, T. R., & Zhang, J., (1998). UECHO: A model of uncertainty management in human abductive reasoning. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1), 3-28.
- Zhang, J., Johnson, T.R., & Wang, H. (1997). The relation between order effects and frequency learning in tactical decision making. *Thinking and Reasoning*, 4(2), 123-145.

# Situating GOMS Models Within Complex, Sociotechnical Systems

**Robert L. West** (robert\_west@carleton.ca)

Department of Psychology; Carleton University  
Ottawa, Canada

**Gabriella Nagy** (gnagy@chat.carleton.ca)

Department of Psychology; Carleton University  
Ottawa, Canada

## Abstract

In this paper we present a methodology for situating GOMS models in complex sociotechnical work domains. The methodology is presented within a larger theoretical framework that relates GOMS modeling to other modeling systems according to principled and systematic guidelines.

Increasingly, computers play critical roles in the running of complex systems such as telecommunications networks and nuclear power plants. However, the role of human agents in these systems is also critical. As computer and software technology improve we see a decrease in the number of technical errors caused by computers, but there is also evidence of a corresponding rise in errors attributable to humans (e.g., Bennett, 1998). No doubt, this is due to the increasing complexity of computers and network systems.

In this paper, we consider the role of GOMS (Card, Moran, & Newell, 1983) in designing systems situated within complex, sociotechnical systems, that is, systems with multiple humans and multiple computers all interacting (see Vicente, 1999 for a more complete definition). GOMS is a method for modeling tasks according to a human agent's goals, operators, methods and selection rules (John, 1995). But in complex sociotechnical systems the task is often a small part of a larger, distributed task. The design problem is analogous to designing a complex operating system. Individual programmers design different components of the system, but each time a new component is added it is unclear if it will create a conflict in the system. Similarly, changing the way an individual performs a task within a complex sociotechnical system can have unforeseen consequences (for a discussion of this point and some interesting examples, see Hutchins, 1995). To deal with this problem operating systems are beta tested. Unfortunately, changes in a sociotechnical system cannot be beta tested and then fixed the next day. In fact, such changes are usually costly and time consuming, especially if people need training. Thus, we need a means to evaluate changes before they are implemented.

## Task Analysis

A task analysis is important for understanding the sort of knowledge driven tasks common in technical areas and large organizations. By knowledge driven we mean that the agent knows, implicitly and/or explicitly, the steps that must be completed. The need for a task analysis presupposes that the

process for completing the task quickly and without error is not common knowledge. Many studies have found that experts often have specialized knowledge that is not expressed in any manual, but is nevertheless crucial for completing the task in an acceptable manner (Mayer, 1997). This is particularly true of tasks situated in sociotechnical environments, which often involve a considerable amount of undocumented knowledge concerning how the various agents, computers, and artifacts involved are coordinated to complete the task.

The result of a task analysis is a model, which is then used to simulate changes in the system. The level of detail of the model will depend on the modeler's goals, and the representation of the model can range from a mental representation, to a paper and pencil representation, to a computerized representation. Furthermore, the goal may be to represent the whole task or only the major components, relationships, and/or functions that characterize the task. The important point is that this process allows some level of foresight into the effect of the proposed changes.

In this paper, we will be concerned with "*modeling systems*". This term is further defined below but for now we can say that a modeling system tells the researcher what types of behaviors to observe and how to organize the data into a functioning model. Thus, a modeling system both guides the task analysis and produces the model. A modeling system could be quite formal (e.g., NGOMSL, Kieras, 1988) or very informal, based on common sense notions about what is important in the task (in this case the researcher may be unaware they are using a modeling system). However, both our formal and informal modeling systems have difficulty coping with complex, distributed systems. One reason for this is that it is easier to think in terms of tasks performed by single agents than tasks distributed across multiple agents, especially when the distributed system is not under some sort of centralized control. When agents act locally and organize themselves, multiple different ways of completing the task can emerge. This results in several different levels of analysis, including the following: (1) the knowledge level, the steps that must be taken to complete the task, (2) variations on a theme, the different ways the task can be done given the constraints of the knowledge level, and (3) the different ways that agents can organize themselves to accomplish different steps of the task. To cope with this, a modeling system must be able to represent the



task at different levels and also be capable of integrating factors involved in completing the task with the factors involved with organizing and sustaining cooperation between the agents. In this paper, we describe how GOMS can be used to cope with this type of system, and the relative advantages of using GOMS under these conditions.

## GOMS

GOMS is a family of relatively formal modeling systems, but we would argue that it has a special status amongst modeling systems. In this regard, it is useful to consider Anderson's (1993) distinction between frameworks, theories and models of human cognition. According to this scheme, frameworks are "bold, general claims about cognition," (p. 2). Theories are created by adding specific assumptions as to how frameworks could be applied to the relevant class of behaviors, and models are created by adding assumptions as to how a theory could be applied to a specific situation or task. The idea that cognition can be understood in terms of production rules (i.e., if..then statements) is therefore a framework, and systems embodying assumptions as to how to use production rules are theories. However, rather than theory, we will use the term *modeling system*, because we are focusing on the process of model building, rather than on testing theories. So, to be clear, we will define a modeling system as a system that allows us to create a model within a specified framework.

The general idea behind GOMS is that well learned human behavior can be modeled using goals, operators, methods, and selection rules (e.g., John, 1995). This claim places GOMS clearly within the production rule framework. Using selection rules to choose between different methods for accomplishing a task essentially embodies the idea of the production rule (i.e., if this, then use this method). Also, operators are necessary for any production system to specify how the system retrieves information from the world and generates behaviors in the world (although operators are sometimes not explicitly represented in production system models of cognition, they are always assumed to exist). The idea that people have goals, or more specifically the idea that people create sub goals to bring them closer to their end goals, is the only element of GOMS that is not directly tied to implementing production systems. For example, the first attempts at implementing production systems (e.g., SOAR, ACT) did not contain any mechanisms for managing goals (Anderson & Lebiere, 1998). However, as Anderson and Lebiere (1998) point out, all of the current production system architectures have a structure for keeping track of goals. Thus, the idea that we use goals to organize cognition can be considered another framework (i.e., it is a bold, general claim about cognition). Therefore, GOMS can be interpreted as asserting that well learned behaviors can be captured using the combined frameworks of production rules and goal structures. At this level, GOMS itself is a general claim about a class of behaviors and remains at the framework level (it is also not possible to falsify this claim without adding further assumptions, another hallmark of the framework level, see Anderson, 1993).

Cognitive architectures can be considered as relatively complete modeling systems (Anderson, 1993). Unlike these

systems, GOMS has no mechanisms for constructing or searching the problem space, it presupposes that the agent has already learned how to get to the end goal. *The key insight, on which GOMS was founded, is that once a path through the problem space has been learned, the complexity of the modeling task is hugely reduced.* This makes moving from the framework level to the modeling system level easier. In fact, the simplest possible GOMS modeling system can be created by merely assuming the appropriate operators exist and structuring goals by connecting them serially, essentially creating a flow chart of goals with branching paths gated by production rules. This type of GOMS modeling system is frequently used, often to sketch out the task structure before creating a more fully functional model. Since this system has no name we will refer to it as Minimal-GOMS.

Other, specific GOMS modeling systems, such as NGOMSL (Kieras, 1988) and CPM-GOMS (Gray, John, & Atwood, 1993), have more detailed assumptions that are contained in the human information-processor (Card, et al., 1983). In this sense, GOMS can be considered a general outline for moving from the dual production rule/goal framework to a specific modeling system by adding assumptions concerning the human information-processor. Following from this, any model that (1) is explicitly or implicitly based on the dual production rule/goal framework, (2) refers only to knowledge driven behaviors (i.e., no learning or problem solving), and (3) makes assumptions concerning the behavior of the human agents involved, can be interpreted as a type of GOMS model. For lack of a better term, we will refer to models that fall into this category, but have not been explicitly created and labeled as GOMS models, as GOMS-like.

Since we are currently interested in modeling errors within complex sociotechnical systems, we searched the literature for error modeling systems and found over 50. However, comparing GOMS to these modeling systems it is clear that they are not on the same conceptual level. In fact, the product of many of the modeling systems we reviewed would be a GOMS-like model. This issue is often the source of confusion and contention between designers that favor GOMS and designers that do not. It is not uncommon to hear people say that modeling system X is a better approach than GOMS, when in actuality modeling system X is a system that produces GOMS-like models.

Part of the problem seems to have arisen from the association of GOMS with models of how long it takes to perform isolated tasks described at the level of individual mouse clicks and button presses. GOMS is particularly good at describing low level activities because the operators are relatively simple and can be described with a reasonable accuracy in the human information-processor (Card, et al., 1983). Since a lot of research, explicitly represented as GOMS research, was done at this level, there is a strong tendency for people to view GOMS as synonymous with the use of low level operators. In actuality, the grain size of the operators should depend on the goals of the researcher (West, Wong, and Vera, 1998).

In terms of complex sociotechnical systems, it is unlikely that GOMS could produce very accurate time estimates as it

is often not possible to assign very precise times to high level social operators (e.g. how long does it take arrange a lunch meeting with a colleague), although, it should still be possible to get good time estimates for well-defined sub tasks. However, the value of GOMS in a multi-agent system is that it allows us to examine the goals and methods of individual agents, and how these relate to the overall task. For example, multi-agent tasks are often described using a critical path analysis. In the case of a centrally controlled task the critical path represents the plan of the central controller. However, when the task is not centrally controlled (i.e., a complex system) the critical path is an emergent property of the interactions between the agents. A multi-agent GOMS model can allow us to examine these interactions for inefficiencies, goal conflicts, and sources of error.

### Complex Sociotechnical Systems

One of the most influential modeling systems in terms of modeling complex systems has been Rasmussen's decision ladder model (1980). As Vicente (1999) points out, the step ladder model is not really a model, but rather a *template* for creating models. Essentially, it is a generic model of information processing that can guide the modeler in terms of the general form a model should take (Vicente, 1999). We believe that the template approach is important for modeling complex, sociotechnical systems, and, more specifically, that it can be used to effectively situate GOMS models within such systems. As Vicente (1999) points out, work within a sociotechnical system cannot be fully captured by GOMS or GOMS-like models because this type of work involves ongoing learning and problem solving, which these models cannot handle. However, as John (1995) points out, GOMS can be very useful for elucidating the components of a task that are amenable to GOMS modeling. In other words, GOMS doesn't have to be the whole solution, but can be part of the solution.

Another important aspect of sociotechnical modeling systems is that they need to be multifaceted in focus. For example, Vicente's modeling system is actually a collection of modeling systems for examining various aspects of the sociotechnical environment, including: the work domain, control tasks, strategies, social organization and cooperation, and worker competencies. Likewise, a modeling system advocated by a well known consulting firm in this area involves a work flow model, a cultural work model, a sequence work model, an artifact model, and a physical environment model (this system is adopted from Beyer & Holzblatt, 1997). What GOMS adds is the potential to integrate knowledge gained in these different domains into a unified model of the knowledge driven portions of the process. Our approach to this is to use a template that (1) allows the task to be described at different levels of complexity and (2) describes how people situate knowledge driven tasks within a complex environment involving ongoing learning and problem solving.

### The Basic Model

Our modeling system is closely related to Norman's (1986) seven-stage model of user activities. However, similar to Rasmussen's decision ladder model, we intend our model to be a generic template for information processing in general, rather than a specific model of human cognition. The framework, which is described in Figure 1, revolves around the goal, *create-plan*. This goal is meant to deal with learning and problem solving, so overall it lies outside the reach of GOMS. One approach to modeling this component would be to use a production/goal based cognitive architecture (e.g., ACT-R, SOAR). This would tie in nicely with the GOMS aspects of the model since they share a common framework, however, any approach can be used, including treating *create-plan* as a black box.

In our current work on telecommunications network maintenance and management we are using Vicente's (1999) work domain analysis to provide the underpinnings for the create-plan component. This involves understanding the constraints imposed by the sociotechnical system and, rather than specifying what a worker should do, specifying what a worker should not do. For example, the main constraint that we have identified is that of the working path (the path through the network carrying live traffic) and the protection path (a path to which the traffic could be shifted). This constraint is critical because whenever work needs to be done or a problem occurs the traffic must be rerouted along a protection path. We are also using Hutchins' (1995) concept of organizational learning to look at how workers pick up on this constraint. GOMS modeling, based on the Figure 1 template, provides the means for describing and evaluating how knowledge driven, procedural tasks fit into the picture. The use of GOMS is very important since this type of sociotechnical system involves many knowledge driven components.

From the perspective of the rest of the model, the function of the create-plan component is to output a knowledge driven plan. The plan may be complete and well thought out but in many cases this will not be the case. Essentially, the cycle embodied by the template is to continue with a plan until it is evaluated as inappropriate or is completed. To further structure this process we need to invoke another GOMS concept, the *unit task* (see Card, et al., 1983). In theory, a plan could be of any size, but we conceptualize plans as unit tasks in the sense that they should correspond to actions that the agent believes can be accomplished without a terminal interruption. Thus, the size of the plan is determined by the nature of the task. For example, the results of Kvan, West, and Vera (1998) indicate that architects in the process of collaborating over a shared whiteboard use very short plans, whereas maintenance procedures on network hardware can involve lengthy procedures that must be completed once started.

Another important function of the create-plan component is to integrate technical, environmental, and social aspects of the task. Thus, in addition to technical procedures a plan should include how to deal with issues arising from the physical environment the task is situated in, as well as the social issues involved in getting cooperation from other agents. As West, et al. (1998) argued, in many cases there are routine ways of dealing with these issues if they

represent routine occurrences. However, in other cases these issues may be dealt with in unique, creative ways. Either way, the model is capturing valuable information (i.e., routine solutions or different case based solutions). Note, though that we are not saying that plans are always complete in this sense. In many cases, plans fail because they do not include ways to deal with problems arising from the physical or social environment. In this case, the system returns to the create-plan process to fix the plan or come up with another.

The other components of the template are described below:

1. Retrieve-next-action – This is meant to reflect the fact that the representation of the plan may be distributed. It is often the case that workers do not have all the knowledge necessary for the task, but they know where to get it (e.g., memory, personal notes, manuals, colleagues).
2. Execute-action – This step refers to firing of operators. As is normally the case in GOMS models, operators can be either physical (e.g., move the mouse), perceptual (e.g., search the screen), or cognitive (e.g., add two numbers). Operators can also vary in grain size and represent complex tasks. For example, an architect might use an operator, *make-aesthetic-judgement*. Such an operator could be represented in terms of the % chance that such a judgement will be positive, or may merely represent the fact that the judgement takes place at a certain point in the model. The grain size and function of the operators will depend on the modeler's goals (see West, et al., 1998 for a discussion about high level operators). This step is also where communication is initiated between agents by using an operator to place messages in the environment (e.g., voice, email, etc.).
3. Update-situation-knowledge – After having acted in some way this section refers to updating the task knowledge to reflect these changes and any other relevant changes that may have occurred during that time (including messages from other agents). For isolated, low level actions this step could be assumed to occur as the actions are being executed. For complex, interactive actions the process of checking may be quite extensive, and may also involve retrieving knowledge from various sources. In this case adding a box above it entitled, *retrieve-data*, might be a good idea.
4. Evaluate – Like create-plan, this box may involve actions that step outside of GOMS. If the situation has changed in an unexpected way there must be a judgement as to whether or not the plan is still appropriate. By definition, unexpected changes will not be part of the plan (Vicente, 1999), so there is a need to step outside of the plan into problem solving or creative thinking to make this judgement. However, it is possible to handle expected or common problems within a plan. Another issue that is important here is the agent's evaluation of risk. Human agents will often engage in risky behaviors, especially if they are under time pressure. Often workers will have heuristics for evaluating risk that can be captured by GOMS.

5. Execute-Patch – if there is a known and immediate fix for a problem the agent goes to execute-patch where the patch with highest probability of success is executed. These known patches can be considered to be implicitly part of the plan. If there is no immediate fix the agent goes to create-plan where a known fix is inserted into the plan to be executed later, or the plan is recreated to cope with the problem.
6. Parallel External Monitoring (PEM) – This module operates in parallel, monitoring the environment for alarms. Creating the model for PEM involves understanding the extent to which the agent can pay attention to the task and also to the general environment. For example, an alarm siren could be assumed to be always picked up, whereas an alarm on a screen could only be picked up when the agent is looking at the screen. The other aspect of the PEM model is that it contains rules for when to interrupt the system and go directly to update-situation-knowledge, and when to store the information in memory until the update-situation-knowledge process comes up. The goal of this model is to capture expert knowledge about monitoring and interruptions.

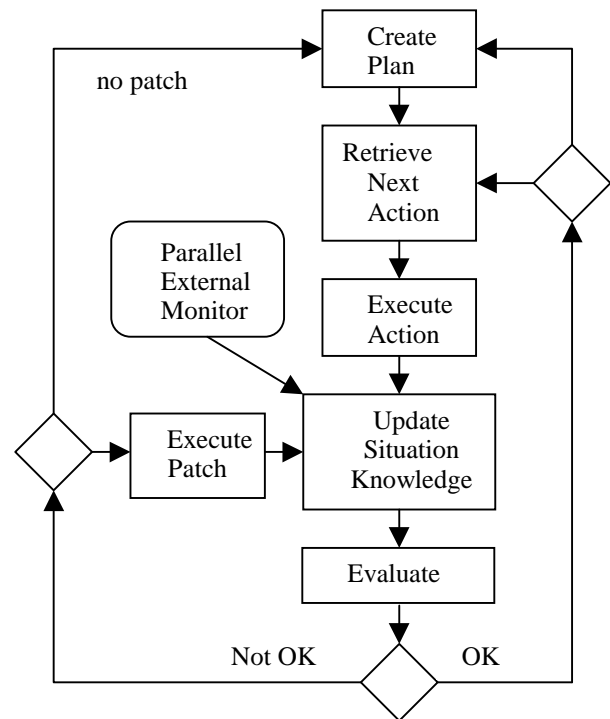


Figure 1. A generic Minimal-GOMS template

### Multiple Agents

So far we have dealt only with modeling an individual. In fact, the original version of this template was developed in an attempt to make sense of data gathered from pairs of architects working collaboratively over a shared whiteboard. As reported in Kvan, West, and Vera (1998), the architects never developed a plan for the collaboration, instead they

dealt with issues and organized themselves as they went along. This resulted in very different organizational structures, all of which were difficult to model. To simplify things a version of the Figure 1 template was developed to first understand the behavior of the individual architects. To create a model of two agents working together you just simply add another template. No lines of communication need to be drawn between the two templates. Instead what is needed is a simple model of the environment that the agents can act on by altering the physical components of the task and by creating messages (e.g., voice, notes, email, etc.). Since the agents are modular they can be added or deleted without too much trouble, so it is possible to have more than two agents.

Using this approach, it was obvious that architects generate very small plans with regard to the task (e.g., draw box at location X) that serve a constantly evolving creative vision. Thus, low level GOMS models of the task components, as defined by the plan size, would be appropriate. However, in addition to creating objects the architects also needed to understand the objects that their partner was creating. This caused a problem in one condition of the experiment in which the architects used a chat line to communicate. To attach a message to an object (e.g., "what is that?") they would either have to describe the object in the message or tell the other person to watch the their whiteboard pointer while they pointed (the white board could get quite complex in terms of the number of objects on it). A solution that would involve fewer steps would be to attach a text box to the pointer to combine the activities of message passing and pointing. This particular solution is not complex, but recognizing the need for it was facilitated by integrating the collaborative elements of the task into the model. Also, notice that although the pointing/messaging solutions the architects came up with were the result of online problem solving, once created they could be treated and evaluated as GOMS type methods.

### Distributed Agents

Although the template is useful for organizing models in which individuals interact, only a relatively small number of agents can be included before the model gets unwieldy. In contrast, complex sociotechnical systems often involve a considerable number of agents. However, we have found that the template is scalable to what we call *distributed agents*. The central premise of distributed cognition is that cognitive agents can organize themselves to form larger, distributed cognitive systems (Hutchins, 1995). Our approach is to treat these distributed cognitive systems as individual agents and apply the same template. This is not to say that there are no differences between brain based cognition, distributed cognition occurring across small groups, or distributed cognition occurring across large groups. There are important differences between these types of structures. However, our argument is that the template captures something basic about the way cognitive systems, in general, deal with interactive, knowledge driven tasks.

We tried this approach at Oxfam, Hong Kong, for modeling the process of deciding how to deliver aid to flood victims in China and found that it simplified the process

considerably (West & Yeun, 1999). It also brought to our attention the distinction between distributed agents and official groups defined within the management structure (i.e., specific departments and their subdivisions). There is a strong tendency for organizations to understand themselves in terms of their official subdivisions, and this information should be part of a complex systems task analysis. However, the goal of GOMS is to build task models, not organizational models. Therefore, a distributed agent is meant to map onto agents that function together to complete a particular task, and will not necessarily map onto a particular department or section. This also means that a person may be a part of different distributed agents depending on the task they are working on. One benefit of this type of analysis is that it can provide insight into the relationship between the task and the management structure.

Following this approach, it is possible to create a higher level model describing the interaction between distributed agents. As with individual human agents, the approach is to represent each distributed agent using the template structure, with communication occurring by placing messages in the environment. Also note that it is possible to combine distributed agents into higher level distributed agents or to break them up into lower level distributed agents, depending on the level of the analysis. It is also possible to mix agents representing individuals with distributed agents. This allows the model to focus in on an individual without representing every other individual connected to the task.

Currently, we are using this modeling system to model tasks involved in telecommunications network maintenance and management. We have found that using this system greatly simplifies the modeling process and also allows the flexibility to address a wide variety of questions.

### References

- Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Bennett, J. (1998). FCC-Reportable Service Outages (3Q92-4Q98) with Procedural errors as Root Cause. Telcodia White Paper.
- Beyer, H., & Holtzblatt, K. (1997). *Contextual design: A customer-centered approach to systems design*. Morgan Kaufmann Publishers.
- Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Gray, W. D., John, B. E., & Atwood, M. E. (1993) Project Ernestine: A validation of GOMS for prediction and explanation of real-world task performance. *Human-computer interaction*, 8 (3), 237-309.
- Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: The MIT Press.
- John, B. E. (1995). Why GOMS? *Interactions*. 2 (10), 80-89.
- Kieras, E. E. (1988). Towards a practical GOMS methodology for user interface design. In M. Helander

- (Ed.), *The handbook of human computer interaction* (pp. 135-138). Amsterdam: North-Holland.
- Kvan, T., West, R. L., & Vera, A. H. (1998). Tools for a virtual design community: Modeling the effects of different tools on design communication. *International Journal of Virtual Reality*, 3 (3), 21-33.
- Mayer, R. E. (1997). From novice to expert. In M. Helander, T.K. Landauer, and P. Prabhu (Eds.), *Handbook of human-computer Interaction* (pp. 781-795). Amsterdam: Elsevier Science.
- Norman, D. A. (1986). Cognitive engineering. In D. A. Norman and S. W. Draper (Eds.), *User centered system design: New perspectives on human-computer* (pp. 31-61). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Rasmussen, J. (1980). The human as a systems component. In H. T. Smith and T. R. G. Green (Eds.), *Human interaction with computers* (pp. 67-96). London: Academic Press.
- Vicente, K. J. (1999). Cognitive work analysis: Toward safe, productive, and healthy computer-based work. Mahwah, NJ: Lawrence Erlbaum Associates.
- West, R. L., Wong, A., & Vera A. H. (1998). GOMS, Distributed Cognition, And the Knowledge Structures Of Organizations. *Proceedings of Cognitive Science 1998*.
- West, R. L., & Yuen, K. L. (1999). *A framework for incorporating social context into GOMS models*. Poster presented at Cognitive Science 1999, Vancouver, B.C.

# Motivation in Insight versus Incremental Problem Solving

Mareike Wieth (wiethmar@msu.edu)

Bruce D. Burns (burnsbr@msu.edu)

Department of Psychology; Michigan State University  
East Lansing, MI 48824-1117 USA

## Abstract

Previous studies have shown a correlation between initial motivation and subsequent performance (e.g. Vollmeyer, Rheinberg, & Burns, 1998). However, it is possible that this relationship is due to a third factor such as general ability. To address this issue, our participants completed insight as well as incremental problems. These two types of problems have been shown to differ both theoretically and empirically due to differential underlying processes (e.g., Metcalfe & Wiebe, 1987). Results showed that motivation (in particular, *interest*) correlated with incremental problem solving but not with insight problem solving. The results were replicated with two different sets of problems solved by different groups of participants. Motivation was measured before solving the problems, so the difference between these two types of problems provides us with evidence that motivation is causal in producing better problem solving performance. Further, it suggests that when processes differ, motivational effect on performance will differ.

## Introduction

It has been difficult to demonstrate conclusively an effect of motivation on problem solving. This is partly because the difficulty of manipulating motivation reliably has forced research designed to examine this issue to rely on correlational studies. Studies such as Vollmeyer and Rheinberg (1998) and Vollmeyer, Rheinberg, & Burns, (1998) have shown a correlation between initial motivation and performance in a complex problem solving task. Although motivation is predictive of performance in these studies, it could still be argued that the correlation is due to a third factor. It is plausible that people with higher general ability at problem solving may not only be better at this task, but also be more highly motivated when faced with such a task. So motivation may not be a causal factor.

In order to learn more from correlational studies of motivation, a slightly different methodology is required. If we give problem solvers qualitatively different problems to solve and find that motivation has a different relationship to performance on these different types of problems, then we would have good evidence that it is not a general ability factor that accounts for any relationship found between motivation and problem solving performance. Two types of

problems that can have a similar form, but have been shown to be qualitatively different, are insight and incremental problems (e.g., Metcalfe & Wiebe, 1987). This makes them good candidates for a methodology looking for qualitatively different motivational influences on problem solving. So in this study we compared the effect of motivation on insight and incremental problem solving.

## Motivation and Problem Solving

It has long been acknowledged that motivation is important, for example, Simon (1967) emphasized the importance of motivational and emotional influence on cognition. However, for the most part motivation and its relationship to cognitive processes has been largely ignored by cognitive scientists. Investigating this influence has been seen as unnecessary because differences in motivation have been treated as background noise that that can be ignored when investigating specific cognitive processes. Even though the operation of Anderson's (1993) ACT-R depends crucially on the goal of the actor and how likely they think an action will be successful, he specifically rules out having to consider the more general goals of the actor. Although Anderson acknowledges the importance of wider goals, he takes the stance that once the actor is committed to doing something in a situation, the actor's more general motivation is irrelevant.

Whether it is sustainable to routinely ignore motivation and emotion when studying cognition is something that has come into question. For example, Kuhl and Kazén (1999) have shown that one of the most well-known of cognitive phenomena – the Stroop effect – can be wiped out by manipulating emotion. Recent research has also started to address the relationship between motivation and cognition (e.g., Kanfer & Ackerman, 1989; Lord & Levy, 1994; Pokay & Blumenfeld, 1990; Vollmeyer & Rheinberg, 1998; Vollmeyer, et al., 1998). With respect to problem solving, Vollmeyer and Rheinberg (1998) fitted their cognitive-motivational process model to a complex problem solving task called *Biology-lab*. The cognitive-motivational process model proposes an interaction between motivation and cognition such that initial motivation affects the motivational state during learning which in turn influences strategy use and acquisition of knowledge. In *Biology-lab* participants have to learn how to manipulate a complex learning environment by controlling several inputs and output variables. In particular, Vollmeyer & Rheinberg have shown that participants with higher motivation were

more likely to use a systematic strategy for acquiring knowledge and therefore performed better during the knowledge application phase. The results of this study and similar studies by Vollmeyer and colleagues using the *Biology-lab* simulation indicate that motivation can influence cognitive processes, such as strategy systematicity, and therefore lead to differential knowledge acquisition and performance.

Similarly, Pokay and Blumenfeld (1990) investigated the effect of motivation on learning strategies and performance on geometry proofs. In this study questionnaires assessing motivation and learning strategies were given to high school students in geometry classes at various points in the semester. The results of this study indicated that various motivational factors predicted strategy use, which in turn influenced performance on geometry tests (especially proofs) throughout the semester. This study provides further evidence for an interaction between motivation and cognition. These conclusions are consistent with other researchers such as Locke and Latham (1990) that have also argued that motivation affects performance via the processes used in a particular task.

### **Insight versus Incremental Problems**

Incremental problems require the solver to take a number of incremental steps in order to solve the problem. Incremental problems have also been referred to as *analytic* (Schooler & Melcher, 1995) or “grind-out-the-solution” problems since people can often solve these types of problems by persisting at the task. It might take time to reach the solution, but the solver has a good idea of how to get there. In contrast, insight problems are those in which the solver has a high probability of meeting an impasse, at which point the solver does not know what to do next (Schooler & Melcher, 1995). Insight problems are usually solved by a “flash of illuminance” (Metcalf & Wiebe, 1987), or by what has been referred to as an “Aha” experience where the solution is suddenly obtained (Schooler & Melcher, 1995).

Differences between these two types of problems have been demonstrated empirically by studies comparing performance on the two types. In a study by Metcalfe & Wiebe (1987) participants were asked to rate how close they thought they were to the solution every 15 seconds while solving incremental and insight problems. The rating results showed that problem solvers had a good idea when they were close to the solution for incremental problems, but were unable to perceive when they were close to a solution for insight problems. Solutions for insight problems came suddenly and with little awareness that the solution was about to be found. Additionally, it was discovered that participants were more successful at predicting which incremental problems they could solve than which insight problems they could solve. These results indicate that there are distinct difference between incremental and insight problems, which could be caused by qualitative differences in underlying processes used to solve these two problems (Metcalf & Wiebe, 1987). Weisberg (1992) has argued that the procedures used in the experiments by Metcalfe & Wiebe (1987) are questionable.

However, it appears that he agrees with the notion that there are different processes involved in solving insight and incremental problems (Weisberg, 1995). Further evidence that there are differences in the processes used to solve these two types of problems has been provided by studies that have had participants give verbal protocols while solving both incremental and insight problems (Schooler & Melcher, 1995; Schooler, Ohlsson, & Brooks, 1993). Schooler et al. found that participants asked to verbalize their problem solving strategies showed significantly impaired performance on insight problems but not on incremental problems. Additionally, it was found that participants paused more and tended to have a harder time articulating their thoughts while solving insight problems compared to incremental problems. Furthermore, the nature of the protocols also differed in that incremental problem protocols contained more logic or means-ends analysis statements than insight problems (Schooler & Melcher, 1995). These findings have been attributed to differences in the processes used to solve these two types of problems. Specifically, Schooler & Melcher and Schooler, et al. argued that the impairments during insight problem solving while verbalizing are due to the disruption of nonreportable processes that are critical to solving insight problems but are not necessary for solving incremental problems.

In an effort to better understand what these nonreportable processes might be, Schooler and Melcher (1995) cite unpublished data by Schooler, McCleod, Brooks, & Melcher (1993) that examined the correlation between a variety of ability measures (e.g., recognizing out of focus pictures, finding remote associates) and success at solving both incremental and insight problems. It was found that the measures predicting performance on the two types of problems were generally different. Anagrams and categorization tasks correlated with incremental problem solving, whereas the embedded figures and out of focus pictures tasks correlated with insight problem solving. These differential patterns of findings lend further support that the two problems draw on qualitatively different processes (Schooler et al.).

It has been suggested that these underlying differences arise from the way we solve insight and incremental problems. Insight problems require searching for an appropriate way to represent the problem and are often easily solved once the correct representation has been found. On the other hand, representation is not the focus for incremental problem solving, instead figuring out what steps to take to reach the solution is often the focus (Kaplan & Simon, 1990; Ohlsson 1984).

These empirical and theoretical differences in incremental and insight problems lend themselves very well to our aim: finding a differential influence of motivation on incremental and insight problems.

### **Motivation Effects on Both Problem Types**

How to get to the solution may be clear when solving incremental problems, but following the required strategy may require some effort and persistence. Vollmeyer and Rheinberg (1998) found that motivation influenced

performance via the use of a good strategy. They suggested that motivation keeps people persisting with the good strategy instead of trying to find some less effortful shortcut. Similarly Sweller (1988) has explained poor problem solving performance as due to the cognitive load imposed by effortful strategies such as means-ends analysis. Based on this we predicted that motivation would influence incremental problem solving. However, not all aspects of motivation may equally relate to performance. Schiefele (1996) argues that interest in the task should be particularly highly related to performance. Therefore, we focused particularly on this aspect of motivation.

In order to argue that any relationship that we may find between problem solving and motivation is not due to some third factor, we wanted to show that motivation does not just correlate with everything. From the above discussion, it appears that insight problems should provide this contrast. As mentioned before, the process of solving insight problems differs from that for incremental problem solving. It is process that Vollmeyer and Rheinberg (1998) and Locke and Latham (1990) focus on. In particular, Vollmeyer and Rheinberg believe motivation influences performance by encouraging participants to persist with a good strategy, yet for insight problems there is no good strategy to follow or to fail to persist with. Persistence may even be detrimental due to the creation of Einstellung effects (Luchins, 1942). Incubation, instead, has been found to be effective for solving insight problems. (Silveira, 1971). (Experienced problem solvers may learn heuristic strategies for insight problem solving, but our participants were not such experts.) The work reviewed above on insight problem solving suggests there is no conscious strategy to be followed in insight problem solving, so we predicted that there would be no relationship between motivation and insight problem solving.

The discovery of such a contrast between insight and incremental problem solving would argue that motivation plays a causal role in how well people solve problems, especially if motivation was measured before the task began. Evidence for this contrast would be finding a higher correlation between motivation and incremental problem solving than between motivation and insight problem solving. However neither of these correlations would be expected to be high given that ability rather than motivation should be the best predictor of problem solving performance.

## A Study

### Method

**Participants.** Two hundred and ninety-two Michigan State University students participated in this experiment for course credit.

**Materials.** Participant's initial motivation was assessed using the Questionnaire of Current Motivation (QCM, Vollmeyer & Rheinberg, 1998). This motivation questionnaire consists of 37 items which have been shown to measure four independent factors of motivation:

Challenge ("This task is a real challenge for me"), confidence in Success ("I think I am up to the difficulty of the task"), Fear of failure ("I am a little bit worried"), and Interest ("I would work on this task even in my free time"). The QCM is designed to measure motivation for a specific task (originally the *Biology-lab* task of Vollmeyer, Burns, & Holyoak, 1996), so some items had to be modified to fit this problem solving task. However, none of the items used to measure the four factors had to be modified and a check of the psychometric qualities of the questionnaire found the same factor structure (see Rheinberg, Vollmeyer, & Burns, under review).

Two separate sets of problems were created, each consisting of one insight problem and two incremental problems. The problems were randomly selected from problems previously studied by Metcalfe & Wiebe (1987) and Schooler, et al. (1993). The actual problems used in the two sets are presented in the Appendix.

Each problem also included a state motivation questionnaire consisting of five questions (see Vollmeyer & Rheinberg, 1998) in order to assess participants motivation towards solving each given problem.

**Procedure.** Participants solved their set of problems in the middle of a 45-50 minute group testing session composed of short unrelated tasks. Group size ranged from five to ten individuals. At the beginning of the session participants were asked to complete the QCM, which was then followed by one of the two sets of three problems (two incremental and one insight). Within each set, the three problems were given in a random order. When solving the set of problems participants were asked to first read the problem, then answer the five questions pertaining to the problem (which measured state motivation), re-read the problem, and then solve it. We did not restrict the time that participants were given to solve each problem, but they were aware that they would be given more tasks. Upon completion of the problem set participants went on to complete a series of unrelated tasks.

### Results

**Overall performance.** The incremental and insight problems were scored on a dichotomous right or wrong scale. To derive a subject's incremental score, the result for the two incremental problems was averaged together. Participants did more poorly than we expected on some of the six problems. For Problem Set One participants on average solved 1.52 of the 2 incremental problems correctly but only 16 percent solved the insight problem. For Problem Set Two participants on average solved 0.70 of the 2 incremental problems and 16 percent solved the insight problem. Note that unlike some other studies of insight problem solving (e.g., Schooler, et al., 1993) participants were not given another chance to attempt the problem if they handed in an incorrect solution.

We tested for any order effects on the problems as each set of three problems was presented in one of six possible orders. We found no evidence in either problem set that the order in which participants solved the problems affected



their performance: Problem Set One  $F(5, 146) = 1.35, p = .25$ ; Problem Set Two  $F(5, 133) = 0.98, p = .43$ .

Table 1: Correlations of the motivation factors with incremental and insight problem scores, and z-score test of the difference between the two correlations.

Problem Set One (n=152)	Incremental	Insight	z-score
Interest	.255 **	.000	2.16 $p = .030$
Challenge	.143 *	-.006	1.23 $p = .22$
Fear of Failure	.061	-.026	.076 $p = .96$
Success	.097	-.011	.89 $p = .37$
Problem Set Two (N=141)	Incremental	Insight	z-score
Interest	.240 **	-.011	2.26 $p = .024$
Challenge	.204 *	-.003	1.85 $p = .064$
Fear of Failure	-.039	.060	-.87 $p = .38$
Success	.169*	-.092	2.33 $p = .020$

\* $p < .05$  \*\* $p < .01$

**Initial Motivation.** As measured by the QCM, participants in both problem sets had reasonably high motivation. The means for the four motivation factors for participants in Problem Set One were as follows: Interest  $M = 4.02$  ( $SD = 1.07$ ), Challenge  $M = 4.41$  ( $SD = 0.88$ ), Success  $M = 5.19$  ( $SD = 0.92$ ), and Fear of failure  $M = 2.66$  ( $SD = 0.88$ ). For Problem Set Two the means were: Interest  $M = 4.02$  ( $SD = 1.05$ ), Challenge  $M = 4.36$  ( $SD = 0.93$ ), Success  $M = 5.36$  ( $SD = 0.83$ ), Fear of failure  $M = 2.65$  ( $SD = 1.02$ ). The two groups did not differ on any of these motivation scales (all  $p > .10$ ). Both groups thought the task moderately interesting and challenging, did not fear failure, and thought they would succeed.

The incremental problem solving scores and the insight problem solving scores were correlated with the four motivational factors of the QCM (Fear of failure, Challenge, Interest, and Success). These correlations are presented in Table 1. For both sets of problems it was found that both Interest and Challenge correlated significantly with incremental problem solving but not with insight problem solving. For each motivation factor we used a z-score conversion (see Olkin, 1967) to test the difference between the correlations of that factor with incremental and insight scores. Only for Interest was the correlation with incremental scores significantly higher than the correlation with insight scores, in both sets. Success correlated significantly with incremental problem solving only for set two, this finding was not replicated with set one. Fear of

failure did not significantly correlate with either incremental or insight problem solving for either set of problems. The difference between correlations for Success was significant for Problem Set Two, but we did not replicate this result with the other set. These findings show that motivation, in particular interest, consistently correlated with incremental problem solving but not insight problem solving.

**State Motivation.** In order to assess if participant's state motivation for each problem influenced problem solving, the three critical motivation questions presented on each of the problem sheets were averaged to create a state motivational score (see Vollmeyer & Rheinberg, 1998). The three critical items were: "This task will be fun", "I'm sure I will find the correct solution", and "It is clear to me how to proceed". Participants' motivation during the task was moderate (see Table 2). Each state motivation score was correlated with its corresponding problem. As Table 2 shows, only one correlation was found to be significant. In Problem Set One, one incremental problem correlated significantly with its state motivation,  $r(141) = .263, p < .01$ . All other correlations were found not to be significantly different from zero. Overall these findings indicate that motivation for each problem was not correlated with performance, regardless of problem type.

Table 2: Correlations of state motivation with each specific problem, together with the percent of subject correctly answering that problem and its mean state motivation (standard deviation in parenthesis).

Problem Set One	Percent correct	Mean (SD) of state motivation	State motivation correlation
Incremental Problem 1	63%	4.55 (1.23)	.263 **
Incremental Problem 2	89%	4.85 (1.18)	.098
Insight Problem	16%	4.15 (1.01)	.060
Problem Set Two	Percent correct	Mean (SD) of state motivation	State motivation correlation
Incremental Problem 1	14%	4.33 (1.23)	.057
Incremental Problem 2	55%	4.11 (1.37)	.058
Insight Problem	16%	4.00 (1.11)	.100

\*\* $p < .01$

## Conclusions

This study achieved our aim of demonstrating a differential relationship between motivation and different types of problems. We not only showed that motivation correlates with one type of problem solving, but that it does *not* correlate with another type. Thus we supported the claim that motivation affects problem solving, and made it hard to argue that such correlations are simply due to some general ability factor. Consistent with the suggestion of Vollmeyer and Rheinberg (1998) and Locke and Latham (1990), motivation only affected problems for which there was a process to be helped or to be disrupted.

The critical motivation factor was Interest, as predicted by previous research (Schiefele, 1996). Interest correlated significantly with incremental problem solving scores. This correlation was significantly higher than the correlation between motivation and insight problem solving scores. Note that although we only report a single study here, in effect the two groups represent a replication of this result. Given that the two groups of participants did different problems sets with differing degrees of success, there appears to be some generality to our findings.

Although the amount of variance in performance explained by motivation is statistically significant, it is small. This does not equate to saying that the influence of motivation on cognition must be correspondingly small. The measures we used were inherently noisy. The QCM is only a pencil-and-paper test of motivation and whether someone solves a particular problem on a particular day is a product of many factors. Further it should be noted that these sorts of problems are often considered to be stable tests of intelligence, so to find motivation influences on even these types of problems is particularly interesting. In future work we will measure ability so as to determine how much of the remaining variance in performance is accounted for by motivation once the variance due to ability has been removed.

We also measured participant's state motivation for each problem they solved. Out of the two problem sets we found only one problem, an incremental problem to be significantly correlated with its state motivation. Whereas this accords with our findings for initial motivation, it was not found consistently; therefore we cannot draw any conclusions. It is possible that the five-question state motivation measure was not sensitive enough or was just not appropriate for this type of task. The state motivation measure is also hard to interpret because it may not only anticipate the problem about to be solved, but also be a reaction to performance on the previous problems. Unlike initial motivation, which is measured before participants solve any of the problems, the direction of any causal arrow would be harder to determine for the state motivation questionnaire.

The main aim of this study was to find a differential effect of motivation on different types of problems, but as well as this we found a much more specific effect (on Interest) as we predicted we would. However, consistent with the idea that motivation affects performance via the mediating processes, we would have to concede that under appropriate conditions motivation may help insight

problem solving. Our conditions may have been particularly conducive to producing a motivation effect on incremental problem solving, but not insight problem solving. By not restricting time, we provided a way in which persistence could help incremental problem solving as the participant did not have to worry about how long it took to get to the solution. Not giving a time limit may have reduced the opportunity for persistence (due to motivation) to affect insight problem solving, as the participant could simply decide that there was no chance to get any further with the problem and just hand it to the experimenter. It is plausible, however, that giving a specific time period for solving problems might actually encourage motivated participants to persist with looking for a solution to insight problems. Therefore, in a situation such as this, motivation might correlate with insight problem solving and not necessarily with incremental problem solving. Note that this in no way weakens our primary aim, demonstrating a motivational effect on problem solving, as these arguments are predicated on the assumption that motivation affects the process of solving problems. The exact patterns of effects on problems solving under different conditions, is a matter for future research.

This was a preliminary study, so further research will be necessary to determine the exact nature of the different effects of motivation on insight and incremental problems. One potential problem with this study was that the insight problems used had a low rate of solution, therefore it would be useful to conduct future research on easier problems. Future research will also need to test our assumption that intellectual ability helps both insight and incremental problem solving. In this study, we assumed that ability affects insight and incremental problem solving equally. This assumption was critical to our argument that motivation helped problem solving rather than any relationship being due to a third-factor, such as ability.

**Implications.** The findings of this study, although somewhat preliminary, have several implications. They show that motivation can influence problem solving, and by extension other cognitive tasks. A practical implication of this is that cognitive scientists should be aware that different tasks might be influenced by motivation in different ways. These possible influences of motivation need, therefore, to be taken into account when designing studies and experiments, otherwise effects may be found simply due to influences of motivation. Most importantly the finding that insight and incremental problems are influenced differently by motivation can be used as a stepping stone to further disentangle motivation and its relationship to cognition.

## References

- Anderson J. R. (1993). *Rules of mind*. Hillsdale, NJ: LEA.
- Kanfer, R. & Ackerman, P.L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology*, 74, 657-690.
- Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive Psychology*, 22, 374-419.
- Kuhl, J., & Kazén, M. (1999). Volitional facilitation of

difficult intentions: Joint activation of intention memory and positive affect removes Stroop interference. *Journal of Experimental Psychology: General*, 128, 382-399.

Lord, R. G., & Levy, P. E. (1994). Moving from cognition to action: A control theory perspective. *Applied Psychology: An International Review*, 43, 335-398.

Locke, E.A. & Latham, G.P. (1990). *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice Hall.

Luchins, A. S. (1942). Mechanization in problem solving. *Psychological Monograph*, 54, No. 248.

Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition*, 15, 238-246.

Ohlsson, S. (1984). Restructuring revisited: I. *Scandinavian Journal of Psychology*, 24, 65-78.

Olkin, J. (1967). Correlation revisited. In J. C. Stanley (Ed.), *Improving experimental design and statistical analysis* (pp. 102-128). Chicago: Rand McNalley.

Pokay, P., & Blumenfeld, P.C. (1990). Predicting achievement early and late in the semester: The role of motivation and use of learning strategies. *Journal of Educational Psychology*, 82, 41-50.

Rheinberg, F., Vollmeyer, R. & Burns, B.D. (under review). *FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen* [QCM: A questionnaire to assess current motivation in learning situations].

Schiefele, U. (1996). *Motivation und Lernen mit Texten* [Motivation and text learning ]. Göttingen: Hogrefe.

Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychological Review*, 74, 29-39.

Silveira, J. (1971). *Incubation: The effect of interruption timing and length on problem solution and quality of problem processing*. Unpublished doctoral thesis.

Sweller, J. (1988). Cognitive load during problem solving: Effects of learning. *Cognitive Science*, 12, 257-285.

Schooler, J. W., & Melcher, J. (1995). The ineffability of insight. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.). *The Creative Cognition Approach*. Cambridge, Mass.: MIT Press.

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, 122, 166-183.

Vollmeyer, R., Burns, B.D., & Holyoak, K.J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science*, 20, 75-100.

Vollmeyer, R. & Rheinberg, F. (1998). Motivationale Einflüsse auf Erwerb und Anwendung von Wissen in einem computersimulierten System [Motivational influences on the acquisition and application of knowledge in a simulated system]. *Zeitschrift für Pädagogische Psychologie*, 12, 11-23.

Vollmeyer, R., Rheinberg, F., & Burns, B.D. (1998). Goals, strategies, and motivation. In M.A. Gernsbacher & S.J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 1090-1095). Hillsdale, NJ: Erlbaum.

Weisberg, R. W. (1992). Metacognition and insight during problem solving: Comment on Metcalfe. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 426-432.

Weisberg, R. W. (1995). Prolegomena to theories of insight in problem solving: A taxonomy of problems. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.). *The Creative Cognition Approach*. Cambridge, Mass.: MIT Press.

### Appendix

<b>Problem Set One</b> (order was random)
<i>Incremental Problem 1:</i> Three cards from an ordinary deck are lying on a table, face down. The following information is known about those three cards (all the information refers to the same three cards):  To the left of the Queen, there is a Jack. To the left of a Spade, there is a Diamond. To the right of a Heart, there is a King. To the right of a King, there is a Spade. <b>Can you assign the proper suit to each picture card?</b>
<i>Incremental Problem 2:</i> Next week I am going to have lunch with my friend, visit the new art gallery, go to the Social Security office, and have my teeth checked at the dentist. My friend cannot meet me on Wednesday; the Social Security office is closed weekends; the dentist has office hours only on Tuesday, Friday, and Saturday; the art gallery is closed Tuesday, Thursday, and weekends. On which day can I do everything I have planned?
<i>Insight Problem:</i> A woman has 4 pieces of chain. Each piece is made up of 3 links. She wants to join the pieces into a single closed loop of chain. To open a link costs 2 cents and to close a link costs 3 cents. She only has 15 cents. How does she do it?
<b>Problem Set Two</b> (order was random)
<i>Incremental Problem 1:</i> The police were convinced that Alan, Ben, Chris, or Dan had committed a crime. Each of the suspects made a statement, but only one of the statements was true:  Alan said, "I didn't do it." Ben said, "Alan is lying." Chris said, "Ben is lying." Dan said, "Ben did it." <b>Who is telling the truth? Who committed the crime?</b>
<i>Incremental Problem 2:</i> If the puzzle you solved before you solved this one was harder than the puzzle you solved after you solved the puzzle you solved before you solved this one, was the puzzle you solved before you solved this one harder than this one?
<i>Insight Problem:</i> A dealer in antique coins got an offer to buy a beautiful bronze coin. The coin had an emperor's head on one side and the date 544 B.C. stamped on the other. The dealer examined the coin, but instead of buying it, he called the police. Why?

# Making Inferences and Classifications Using Categories That Are Not Linearly Separable

**Takashi Yamauchi**  
Texas A&M University  
College Station, TX 77843

**Arthur B. Markman**  
University of Texas, Austin  
Austin, TX 78712  
markman@psy.utexas.edu

## Abstract

Previous research suggests that categories learned through classification focus on exemplar information, while categories learned by making predictive inferences focus on summary (i.e., prototype) information. To test this idea further, we demonstrated that it is more difficult to learn nonlinearly separable categories by making inferences than by classifying. This research also supports previous studies by indicating that different processes are likely to mediate inference and classification

In this paper, we examine the type of categorical information people assess in the process of obtaining inductive knowledge. Specifically, we investigate the extent to which abstract summary information about a category and specific information about individual exemplars of a category are used to make feature inferences.

Categories license inference in at least two ways. First, categories provide a summary representation of their members (e.g., a prototype). Given an unknown feature of a bird, for example, people may predict the value of that feature by referring to the bird prototype (Rips, 1975; Tversky & Kahneman, 1974; Yamauchi & Markman, 2000, in press). Another source of category-based induction comes from individual exemplars of a category. Many studies have shown that people classify items by retrieving information about specific exemplars from memory (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986). A similar process may be used to make feature inferences. In predicting an unknown feature of an item, people may predict characteristics of the new item based on exemplars stored in memory.

Studies investigating classification have shown that exemplar information plays a crucial role in making classification judgments. Research on inductive inference, however, reveals that category-level abstract feature information (e.g., prototypes) is crucial for inference. For example, Anderson and Fincham (1996) demonstrated that people are capable of predicting the value for one feature given the value of another, based on the overall correlation between the features in the study phase of the experiment, rather than on the basis of seeing those specific values during the study phase. Yamauchi and Markman (2000) further showed that varying the appearance of exemplars during learning disrupts classification, but not inference.

These findings suggest that, while classification and inference may be formally equivalent, they make use of

different kinds of information in practice (Yamauchi & Markman, 1998). In this paper, we extend this hypothesis and examine the idea that category-level summary information provides a basis for inference (e.g., prototypes), while exemplar information plays a major role in classification. In the following sections, we describe the inference and classification tasks that were employed in our experiments. Then, we examine the role of exemplar and prototype information in two experiments.

## Classification and Inference

In our experiments, classification is defined as a practice in which an item is placed into one of two groups based on its attributes. Inference is defined as a practice in which an attribute of an item is predicted given the category label of the item as well as information about its other attributes. For example, classification as we define it is akin to the prediction of a category to which a person belongs (e.g., Democrat) having observed his attributes (e.g., supports affirmative action and favors reduced defense spending). Inference is akin to predicting an attribute of a person (e.g., supports affirmative action) given a category to which he belongs, and other known attributes (e.g., is a Democrat and favors reducing defense spending). We further define the term *category label* as a symbol that represents category membership by denoting a particular group of exemplars, and the term *category feature* as a symbol that denotes a characteristic of an exemplar. Classification requires the prediction of the category label based on the features of the item; inference requires the prediction of a category feature based on the information about other features and the category label.

In our experiments, subjects learn two categories (Table 1a) through a classification task or an inference task. On a classification trial, subjects are presented with a stimulus depicting the values of the form, size, color and position of the geometric figure, and they predict the category label of that stimulus (see Figure 1a). On an inference trial, subjects are presented with the values of the size, shape and position of the geometric figure along with the category label to which the stimulus belongs (e.g., Set A), and they predict the value of the missing feature (e.g., the color) (Figure 1b). On different trials, subjects predict different features. In this manner, classification and inference are formally equivalent if a category label is regarded as simply another feature (Anderson, 1990; but see Yamauchi & Markman, 1998, in press, for further discussion).

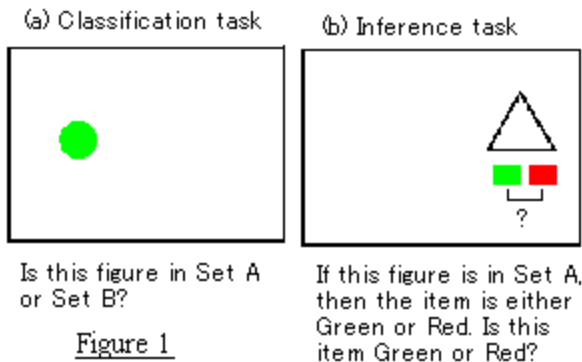


Table 1a The category structure Experiment 1

Learning	Category A					Category B			
	F	S	C	P		F	S	C	P
A 1	1	1	1	1	B 1	0	0	0	0
A 2	1	0	1	0	B 2	1	0	1	1
A 3	0	1	0	1	B 3	0	1	0	0
Transfer									
A 4	0	1	1	1	B 4	1	0	0	0
A 5	1	1	0	1	B 5	0	0	1	0
A 6	1	1	1	0	B 6	0	0	0	1
					B 7	0	0	1	1
					B 8	1	1	0	0

Table 1b Linearly Separable Categories

A 1	1	1	1	0	B 1	0	0	0	1
A 2	1	1	0	1	B 2	0	0	1	0
A 3	1	0	1	1	B 3	0	1	0	0
A 4	0	1	1	1	B 4	1	0	0	0
A 0	1	1	1	1	B 0	0	0	0	0

Figure 1a

In our previous studies (Yamauchi & Markman, 1998, 2000), we used linearly separable categories (see Table 1b) and found that these categories are easier to learn given an inference learning task than given a classification learning task. We reasoned that this result was obtained because inference relies on summary information about category members. The linearly separable categories have prototypes that summarize the feature values of the individual exemplars, although the prototype differs from all of the exemplars by a feature (e.g., A0&B0 in Table 1b). In this structure, additive combinations of feature values divide the two categories nicely; therefore, extracting prototypes from the two categories facilitates learning these categories.

Categories that are not linearly separable have a very different structure, and hence we expect a different pattern of performance on inference and classification tasks. A sample set of nonlinearly separable categories is shown in Table 1a. For these stimuli, subjects may find prototypes in the two categories (Category A=(1, 1, 1, 1) and Category B=(0, 0, 0, 0)). Nonetheless, this information is not useful for integrating category members in each category as no additive combination of feature values can predict category

coherence (Medin & Schaffer, 1978; Wattenmaker, Dewey, Murphy, & Medin, 1986). For example, the stimulus B2 differs minimally from the prototype in Category A but is included in Category B. In order to learn these categories, subjects need to remember the specific exemplars (see Medin & Schaffer, 1978). Because there are only 6 exemplars in the two categories, it is not difficult for subjects to store these exemplars in memory. It is difficult to learn to make feature inferences, however, because there is no abstract summary information that provides a good description of the categories. Thus, for nonlinearly separable categories, we expect a reversal in the ease of inference and classification relative to linearly separable categories, with the categories being difficult to learn and process through inference than through classification. We test this idea in Experiment 1.

## Experiment 1

We used geometric figures as stimuli. All the stimuli varied along four binary feature dimensions: size (large, small), form (circle, triangle), position (left, right) and color (red, green). This structure is shown in Table 1a. These stimuli and the categories are equivalent to those employed by Medin and Schaffer (1978).

In Experiment 1, the subjects learn these two categories in one of two conditions: (1) Classification or (2) Inference.<sup>1</sup> In the Classification Learning condition, the subjects respond to classification questions. In the Inference Learning condition, the subjects respond to inference questions. Initially, no information about the categories is given to subjects in our studies, so that they have to learn the two categories incrementally by trial and error, based on the feedback that they receive after their response. The learning phase continues until subjects reach a criterion of 90% accuracy in three consecutive blocks (18 trials) or until they complete 30 blocks (180 trials).

Following the learning phase, we test the nature of this category representation using transfer trials, which consist of classifications and inferences of old stimuli that appeared during learning and new stimuli that did not appear during learning. In the transfer phase, all the subjects receive the same trials. Transfer stimuli were designed to explore the distinction between inference and classification. For example, the transfer stimuli, A4-A6 and B4-B6, deviate equally from the prototype of each category. Thus, subjects in Inference Learning should be able to classify these stimuli equally well after learning. These stimuli differ in the extent to which they share features with individual exemplars. The stimuli B4-B6 are highly similar to one exemplar in Category A and one exemplar in Category B. In contrast, the stimuli

<sup>1</sup> In our original experiment, we also included a Mixed condition, in which half trials consisted of classification questions and the remaining half were inference questions. Most scores obtained from the Mixed condition fell approximately midway between the Classification condition and the Inference condition. In order to focus on the distinction between inference and classification, we will not report the results from the Mixed condition in this paper.

A4-A6 are highly similar to two exemplars in Category A, but are not similar to any of the exemplars in Category B (Medin & Schaffer, 1978, p. 218). Thus, subjects in Classification Learning (in contrast to those in Inference Learning) should classify the stimuli A4-A6 more accurately than the stimuli B4-B6. A similar prediction holds for the stimuli B7 (0, 0, 1, 1) and B8 (1, 1, 0, 0). These two stimuli are neutral with respect to the two prototypes. Both stimuli have two feature values consistent with Category A and two feature values consistent with Category B. However, they are highly similar to at least one of three exemplars of Category B (B7 is similar to B2, and B8 is similar to B3), but they are not similar to any of the exemplars of Category A. As a consequence, the stimuli B7 and B8 should be accurately classified into Category B as a function of exemplar storage during learning. Finally, because categories that are not linearly separable do not provide an accurate summary of category members, subjects in the two conditions should have difficulty making transfer inferences to new stimuli.

**Participants and Materials.** 49 subjects participated in this study. The data from 1 subject were lost due to an error in recording. In total, the data from 48 subjects (24 in each condition) were analyzed. Each category consisted of three exemplars that were shown during learning and transfer trials. In addition, there were eight new stimuli that were given only in the transfer phase. Two versions of the feature assignment were introduced in this experiment. In one version, the value of 0 was triangle and the value of 1 was circle. For color, the value of 0 was green and the value of 1 was red. For size, the value of 0 was small and the value of 1 was large. For position, the value of 0 was right and the value of 1 was left. In the other version, the values of form and size were reversed. Each stimulus was bounded by a 20.3 x 17.4 cm rectangular frame drawn with a solid black line on the computer screen.

**Procedure.** The experiment involved three phases — a learning phase, a filler phase and a transfer phase. In the learning phase, subjects were randomly assigned to one of two conditions — Classification and Inference. In the two conditions, subjects continued in the learning phase until they performed three consecutive blocks with a combined accuracy of 90% or until they completed 30 blocks (180 trials). A classification block consisted of presentations of six exemplars. One inference block consisted of one inference (along one of the four dimensions) for each of the six stimuli. In the two conditions, every exemplar appeared once in the feedback of each block. The order of stimulus presentation was determined randomly.

In Classification Learning, subjects saw one of the six stimuli and indicated the category to which it belonged by clicking a button with the mouse (Figure 1a). In Inference Learning, subjects inferred a value for one of the four feature dimensions while its category label and the remaining three feature values were depicted in the stimulus frame (Figure 1b). Different dimensions were predicted on different trials. Subjects responded by clicking one of two labeled buttons

with the mouse. For each stimulus, the location of the correct choice was randomly determined. Following each response, feedback and the correct stimulus were presented on the screen for three seconds. The stimuli presented during feedback were identical in both the classification and inference tasks.<sup>2</sup>

After the learning trials, there was a brief filler task, and then all subjects carried out the same transfer tasks. In the transfer phase, subjects were first given classification transfer followed by inference transfer. The transfer stimuli consisted of 6 old stimuli and 8 new stimuli (Table 1a). All of which were shown both in the classification transfer task and in the inference transfer task. The order of stimulus presentation for each task was determined randomly. All the feature inferences were given in Inference learning. No feedback was given during transfer.

## Results and Discussion

Overall, the basic results of Experiment 1 are consistent with our hypothesis (Table 2). With nonlinearly separable categories, inference was much more difficult than classification. This finding contrasts with previous research with linearly separable categories, where inference was easier than classification (Yamauchi & Markman, 1998).

In all, 17 subjects reached the learning criterion in the Inference Learning condition, and 22 subjects reached the criterion in the Classification Learning condition. Considering only those who reached the learning criterion, subjects in the Inference Learning condition ( $\bar{m}=15.8$ ) required significantly more blocks during the learning phase than did subjects in the Classification Learning condition ( $\bar{m}=10.5$ ),  $t(37)=3.32$ ,  $p<0.01$ , (Table 2).

For the classification transfer of old stimuli, subjects given Classification Learning ( $\bar{m}=0.92$ ) were significantly more accurate than subjects given Inference Learning ( $\bar{m}=0.69$ );  $t(37)=5.28$ ,  $p<0.01$ . As predicted classification, but not inference, involves comparisons to exemplars. Subjects given Classification Learning classified the stimuli A4-A6 ( $\bar{m}=0.76$ ) more accurately than the stimuli B4-B6 ( $\bar{m}=0.45$ ), although the two sets of stimuli deviate equally from the prototype of each category;  $t(42)=3.73$ ,  $p<0.01$ . In contrast, there was no statistical difference in classification accuracy for the stimuli A4-A6 ( $\bar{m}=0.63$ ) and the stimuli B4-B6 ( $\bar{m}=0.55$ ) in subjects given Inference Learning;  $t(32)=0.77$ ,  $p>0.1$ . Also as predicted, for the neutral stimuli B7 and B8, subjects in Classification Learning were more likely to classify these stimuli into Category B ( $\bar{m}=0.61$ ) than were subjects in

<sup>2</sup>The inference for the size of the stimuli B1 and B3 has two right answers. Given the inference question (0, ?, 0, 0), the response of the feature value 1 corresponds to the stimulus B3 and the response of the feature value 0 corresponds to the stimulus B1. We gave subjects a correct feedback irrespective of their responses for this question. This treatment should make inference learning faster, and thus functions against our hypothesis that inference learning requires more trials than classification learning for this category structure.

Inference Learning ( $\bar{m}=0.50$ ), but this difference was not statistically significant;  $t(40)=1.04$ ,  $p>0.10$ .

Table 2 The main results from Experiment 1

Classification Transfer					
	Old		New		Neutral
		Average	A4-A6	B4-B6	B7&B8
IL	0.69	0.59	0.63	0.55	0.44
CL	0.92	0.61	0.76	0.45	0.61
Inference Transfer					
	Old		New		Neutral
		Average	A4-A6	B4-B6	B7&B8
IL	0.79	0.46	0.40	0.51	0.36
CL	0.75	0.50	0.50	0.50	0.50

IL: Inference Learning, CL: Classification Learning

For the neutral stimuli B7&B8, we measured the proportion that subjects classified the two stimuli into Category B.

For the inference transfer, subjects in the two conditions were about equally accurate in making feature inferences for old stimuli; Inference Learning,  $\bar{m}=0.79$ , and Classification Learning,  $\bar{m}=0.75$ . Their performance declined sharply given the inference transfer of new stimuli; Inference Learning,  $\bar{m}=0.46$ , Classification Learning,  $\bar{m}=0.50$ . The performance exhibited by subjects in Classification Learning was no better than a chance level;  $t(21)=0.11$ ,  $p>0.1$  (one-tail). The performance exhibited by subjects in Inference Learning was actually significantly below chance;  $t(16)=-2.36$ ,  $p<0.05$  (one-tail). This poor performance contrasts with what we observed in classification transfer, where performance on new items was significantly above chance in both learning conditions. These results are consistent with the view that categories that are not linearly separable provide little support for predictive inference.

The results of Experiment 1 support our view that it is difficult to make inferences for nonlinearly separable categories. Furthermore, the results indicate that inference and classification, two of the main functions of categories, differ significantly in the category information they utilize. In Experiment 2, we investigate this hypothesis further by examining a factor that distinguishes inference and classification.

## Experiment 2

We have proposed that inference focuses on summary information about the category. In contrast, there is evidence that people who are trying to classify a set of items tend to focus on diagnostic information that reliably distinguishes between categories (Nosofsky, Palmeri, & Mckinley, 1994). For example, in sorting tasks people tend to divide the stimuli into groups on the basis of a single dimension, even when there is a clear family resemblance structure among the exemplars (Ahn & Medin, 1992; Medin, Wattenmaker, & Hampson, 1987). The hypothesis that classification tends to focus on diagnostic features and

inference tends to focus on summary information received indirect support in our previous studies (Yamauchi & Markman, 1998, 2000, in press). In Experiment 2, we will test this idea more directly and scrutinize the distinction between inference and classification.

Table 4 shows the structure of the two categories used in Experiment 2. The categories consist of 3 exemplars each. The stimulus configuration A0(1, 1, 1, 1) summarizes Category A, and the stimulus configuration B0(1, 1, 0, 0) summarizes Category B because these feature values are dominant in each feature dimension of the two categories. In this category structure, the first two dimensions (form and size in Table 4) of the two prototypes are the same, so that they are not useful for distinguishing between the two categories. In contrast, the last two dimensions (color and position in Table 4) are more informative for distinguishing between the categories. Thus, if classification promotes attention to the features that differentiate the two categories, subjects in Classification Learning should attend more to feature information about color and position than to information about form and size. In contrast, because inference is assumed to focus on relations among features within a category, subjects given inference learning should be equally sensitive to the four feature dimensions.

This category structure is also useful for distinguishing the extent to which subjects assess a summary of the category as opposed to individual exemplars. In particular, subjects in Inference Learning should have difficulty acquiring these two categories because the stimulus A2 is the prototype of Category B, but is actually a member of Category A. Subjects in Inference Learning should also have trouble inferring features that do not correspond to the prototype stimuli of the two categories (which we call *category-discordant features*). For example, subjects in Inference Learning should exhibit less accurate performance for feature values that do not correspond to the prototype (the value 0 of Category A, and the value 0 of form and size in Category B and the value 1 of color and position of Category B). These factors, however, should not influence subjects in Classification Learning, because this task should focus people selectively on diagnostic features and individual exemplars.

**Participants and Materials.** Subjects were 48 members of the Columbia University community. The materials used for this experiment were the same kind of four-dimensional stimuli used for Experiment 1, but they were organized into a different category structure (Table 4). Each exemplar of a given category had two feature values in common and one feature value different from the rest of the members of that category. The prototype of Set A was (1, 1, 1, 1), which was also a member of the category (exemplar A1 in Table 4). The prototype of Set B was (1, 1, 0, 0), which was actually a member of category A (exemplar A2 in Table 4). The six exemplars from Table 4 were used for Classification Learning and classification transfer. Inference Learning and inference

transfer consisted of inferences of all the feature dimensions of the six exemplars (in total 24 different questions).

Table 3 The category structure used in Experiment 2

	F	S	C	P		F	S	C	P
A 1	1	1	1	1	B 1	1	1	0	<i>I</i>
A 2	1	1	<i>0</i>	<i>0</i>	B 2	<i>0</i>	1	<i>I</i>	0
A 3	<i>0</i>	<i>0</i>	1	1	B 3	1	<i>0</i>	0	0
A 0	1	1	1	1	B 0	1	1	0	0

Category-inaccordance features are shown in italics.

A 0 is the prototype of category A and B 0 is the prototype

of category B. F=Form, S=Size, C=Color, P=Position

**Procedure.** The basic procedure of this experiment was identical to that described in Experiment 1.

## Results and Discussion

As predicted, learning these categories was particularly difficult for subjects given Inference Learning. All subjects (24) in Classification Learning, but only 8 subjects in Inference Learning reached the learning criterion. On average, subjects in Classification Learning spent 10.4 blocks, and subjects in Inference Learning spent 27.4 blocks in learning;  $t(46) > 10.0$ ,  $p < 0.01$ . Because the number of subjects who reached the criterion differed considerably between Classification Learning and Inference Learning, we analyzed the transfer data from each learning condition separately.

In Classification Learning, subjects exhibited accurate performance for classification transfer ( $\underline{m}=0.94$ ). Subjects' classification performance was generally high for all six stimuli. For the six transfer stimuli, the accuracy ranged from 88% to 96%. Subjects were also accurate in the classification of the stimulus A1 ( $\underline{m}=0.88$ ), which is the prototype of category A (and a member of category A) as well as stimulus A2 ( $\underline{m}=0.92$ ), which is the prototype of category B, but is actually a member of category A. During the transfer phase, subjects classified the stimulus A1 and the stimulus A2 equally well;  $Z=-0.02$ ,  $p > 0.1$  (Table 4).

Subjects in Classification Learning were also accurate in inference transfer ( $\underline{m}=0.83$ ). Consistent with our prediction, Classification Learning clearly led subjects to focus on the features that were useful for distinguishing between categories. Subjects in Classification Learning performed significantly better for the feature inferences of color and position ( $\underline{m}=0.86$ ) than for form and size ( $\underline{m}=0.80$ );  $t(23)=1.83$ ,  $p < 0.05$  (one-tailed).

In Inference Learning, we analyzed the data from all subjects, because only 8/24 subjects reached the learning criterion. First, the average performance for classification transfer by subjects in Inference Learning was  $\underline{m}=0.70$ . Unlike in Classification Learning, in Inference Learning there is a wide disparity between accuracy in classifying the stimulus A1 and the accuracy in classifying the stimulus A2. Subjects in Inference Learning accurately classified the prototype stimulus of Category A — A1(1, 1, 1, 1),  $\underline{m}=0.83$  —

but not the prototype stimulus of Category B — A2(1, 1, 0, 0),  $\underline{m}=0.46$ ;  $Z=2.41$ ,  $p < 0.01$ . This result suggests that subjects were focusing on information that summarized the categories rather than on information about specific exemplars.

Table 4 The main results of Experiment 2

	Classification Transfer			
	A1	A2	All exemplars	
IL	0.83	0.46	0.70	
CL	0.88	0.92	0.94	
	Inference Transfer			
	F	S	C	P
IL	0.72	0.68	0.73	0.68
CL	0.81	0.80	0.88	0.85

F=form, S=size, C=color, P=position

Consistent with our prediction, subjects in Inference Learning did not differ in the feature inferences of form and size, as compared to the feature inferences of color and position (form & size,  $\underline{m}=0.70$ , color & position,  $\underline{m}=0.70$ ). This result, combined with the results from Classification Learning, clearly indicates that inference and classification make use of different types of feature information.

Subjects in Inference Learning were not different in the inference transfer of Category-accordant features ( $\underline{m}=0.71$ ) (i.e., prediction of feature values that are the same as the value for the prototype of that category) and Category-discordant features ( $\underline{m}=0.69$ ) (i.e., prediction of features that have a different value than the prototype of the category);  $t(22)=0.62$ ,  $p > 0.10$ . A similar tendency appeared for subjects in Classification Learning; Category-accordance features ( $\underline{m}=0.84$ ) and Category-discordant features ( $\underline{m}=0.82$ );  $t(22)=0.69$ ,  $p > 0.1$ . We applied the same analysis to the learning performance of subjects in Inference Learning. The results revealed that subjects' learning performance was significantly more accurate for Category-accordant features ( $\underline{m}=0.63$ ) than for Category-discordant features ( $\underline{m}=0.56$ );  $t(22)=3.46$ ,  $p < 0.01$ . This analysis indicates that people find it difficult to make correct inferences for features that do not correspond to the category prototype during learning.

Taken together, The results of these studies support the hypothesis that nonlinearly separable categories are difficult to learn through inference. Our results also suggest that inference and classification promote a focus on different types of category information: The Classification Learning task guides subjects to focus on features that distinguish between categories; the Inference Learning task directs subjects to attend to the features that integrate the members within a category.

## General Discussion

These studies demonstrate that it is easier to learn categories through classification than through inference when the categories are not linearly separable. This finding



contrasts with earlier research with linearly separable categories, which found that inference learning was easier than classification learning. This finding reflects that summary category information is more important for inference than for classification. Our experiments, combined with the results from previous studies (Yamauchi & Markman, 1998, 2000, in press), suggest that the structure of a category is one of the major constraints on inductive inference. Unlike classification, inference requires feature information that relates the members of a category. Although some researchers argue that inference and classification are the same thing (e.g., Anderson, 1990), our results reveal that people exercise different strategies for the two tasks.

Why do people look for abstract summary information for inference, while they seek information about specific exemplars or diagnostic features for classification, even when they are given the same categories? This difference may follow from an intricate link between category representation and category functions. Classification is related to object identification and recognition (Nosofsky, 1986). Thus, it requires finding relationships between an individual exemplar and its category label. Once an object is identified, its overall feature information may become irrelevant except some features that are useful to distinguish between categories. In contrast, inference involves the prediction of missing feature values, and thus requires finding relationships between the category label and the features of the category (Gelman, 1986). In this case, the category identity of the object is known, and so information about the category features is needed to predict the value of missing features. Thus, differences in what is demanded in each task lead people to look for distinctions between groups given a classification task, and to seek commonalities within a group given an inference task.

### Acknowledgements

This work was supported by NSF grant SBR-9905013 given to the second author. The authors thank Brad Love for comments on this manuscript.

### References

- Ahn, W., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science, 16*, 81-121.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.
- Anderson, J. R., & Fincham, J. M. (1996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(2), 259-277.
- Gelman, S. (1986). Categories and induction in young children. *Cognition, 23*, 183-209.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification. *Psychological Review, 85* (3), 207 - 238.
- Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19*, 242-279.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.
- Nosofsky, R. M., Palmeri, T. J., & Mckinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53-97.
- Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14*, 665-681.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.
- Yamauchi, T., & Markman, A. B. (1998). Category-learning by inference and classification. *Journal of Memory and Language, 39*, 124-148.
- Yamauchi, T., & Markman, A. B. (2000). Learning categories composed of varying instances: The effect of classification, inference and structural alignment. *Memory and Cognition, 28* (1), 64-78.
- Yamauchi, T., & Markman, A. B. (in press). Inference using categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*.
- Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology, 18*, 158-194.

# Structure-Mapping Theory and Lexico-Semantic Information

Daniel Yarlett & Michael Ramscar  
{dany,michael}@cogsci.ed.ac.uk  
Division of Informatics  
University of Edinburgh  
2 Buccleuch Place  
Edinburgh EH8 9LW

## Abstract

In modelling analogy the Structure Mapping Engine (Gentner, 1983; Falkenhainer, Forbus and Gentner, 1989) can only map successfully on representations in a canonical form because it only permits mappings between relations with lexically-identical functors. We examine whether co-occurrence statistics can remedy this by providing an appropriate basis for modelling lexico-semantic relations. Using a co-occurrence model we reimplement SME to allow it to map between relations with functors that are lexically-distinct. Computational experiments are then reported which show that the resulting model, M-SME, maps successfully on representations which faithfully encode lexical properties, indicating that semantic constraints should only play a minimal role in the mapping process.

## The Structure-Mapping Theory

The structure-mapping theory was originally proposed as a set of constraints defining permissible mappings between a base and target domain in analogy (Gentner, 1983), and implemented in the Structure-Mapping Engine (Falkenhainer, Forbus and Gentner, 1989). Structure-mapping theory constructs analogical mappings between discrete domains (called ‘Dgroups’) of propositional statements, with the main focus being on mapping interconnected *relational* structure.

### The Lexical-Identicality Constraint

In detecting shared relational structure the structure-mapping theory only permits mappings to be made between relations if, and only if, they have lexically-identical functors and the same number of arguments. Thus there are two constraints on the formation of an initial *match hypothesis*. We call the first constraint on match hypothesis formation the *lexical-identicality constraint*, and it is important to observe that it carries a commitment to a canonical theory of representation because it requires that mappable relations are represented with identical names. For example, structure-mapping theory would not permit an alignment between the following two relations, even though it might be appropriate in a wider context:

(ORBITS PLANET SUN)

(REVOLVES\_AROUND ELECTRON ATOM)

The fact that ORBITS is not lexically-identical to REVOLVES\_AROUND also means that the corresponding

analogical mappings between the arguments of the relations (PLANET with ELECTRON, and SUN with ATOM) are not made. Holyoak and Thagard (1995) have argued that this constitutes a significant weakness in structure-mapping theory: “with its emphasis on structure to the exclusion of all other constraints, SME does not simply discourage mappings between non-identical but semantically similar items; it does not even permit them.”

Both the ACME (Holyoak and Thagard, 1989) and LISA (Hummel and Holyoak, 1997) models of analogy avoid this objection by postulating semantic links that hold between the names of relations. These links are hand-coded into the propositional representations on which the analogical mappings are generated. If a sufficiently strong semantic link is coded between two relations then a mapping can be countenanced between them. Thus, in the example above, ACME’s or LISA’s representations could incorporate a sufficiently strong semantic link between ORBITS and REVOLVES\_AROUND to enable a mapping to be generated from one relation to the other.

## The Canonical Representation Theory

Holyoak and Thagard’s criticism of the structure-mapping theory is not entirely fair, however, as it ignores SME’s commitment to a *canonical representation* (CR) theory. The CR theory claims that relations that are sufficiently similar in ‘meaning’ to facilitate mappings (e.g. ‘orbits’ and ‘revolves around’) are coded with identical tokens (in this case both might be coded as ‘orbits’). This extra assumption of the structure-mapping theory would allow the intuitively correct mapping to be made in the above case. However, since the postulation of semantic links and the CR theory rely on human-based coding decisions – and neither subscribe to a worked out model of semantics – both are ultimately equivalent in terms of their explanatory power.

The CR commitment of structure-mapping theory allows a *modular* approach to be taken to the cognitive modelling of analogy. By mapping across canonical representations questions of semantics are left outwith the scope of structure-mapping theory – SME thus remains noncommittal with respect to a theory of lexical semantics. In the experiments that follow we exploit SME’s modular approach to modelling by using the information provided by a co-occurrence model of lexical semantics to see if this allows SME to map successfully on non-canonical representations, and avoid the underspecifica-

tion inherent in the CR theory.

## Experimental Materials

**The ‘Karla the Hawk’ stories.** The Karla the Hawk materials were chosen as the test domain in this study (Gentner, Ratterman and Forbus, 1993). The materials consist of twenty sets of stories written in natural language. Each set consists of a base story, and four systematic variations of that story. Two factors are crossed over the four variant stories, as shown below.

	+ST	-ST
+SF	Literal Similarity	Surface Similarity
-SF	Analogical	First-Order Relations

Table 1: The commonalities each variant category shares with the corresponding base it is derived from.

The four story categories systematically vary the commonalities that are shared with the base-story from which they are derived. Each variant can either share or not share surface ( $\pm SF$ ) and structural ( $\pm ST$ ) commonalities with the corresponding base-story. Because analogy consists in two domains possessing a shared structure, this  $2 \times 2$  materials design allows for the controlled examination of SME’s performance. If SME is performing appropriately then we would expect a better mapping performance when mapping the base representations on to the LS and AN category materials, as they share structural commonalities.

**The Faithful Dgroups.** The standard representations that SME operates on, the *Original Dgroups*, encode relation names in canonical form in accordance with the CR theory. In order to test the performance of SME on representations that do not embody a commitment to the CR theory we developed our own representations that faithfully encode the relation names as used in the original natural language Karla stories. We call this set of representations the *Faithful Dgroups*, and they were produced by transferring the lexemes used to express relations in the original natural language Karla materials directly into the propositional form required by SME.

### Experiment 1A

This first experiment was conducted to test the performance of SME on the Original Dgroups, which are the original encodings of nine of the twenty Karla the Hawk story-sets (Forbus, Gentner and Law, 1994). This was in order to provide a base measure of SME’s performance.

**Method.** For each of the nine sets of Original Dgroups SME was used to map the base Dgroup onto its four variants. The Structural Evaluation Score (SES)<sup>1</sup> and number of match hypotheses formed for each mapping were then recorded.

<sup>1</sup>SES scores are automatically calculated by SME and provide a measure of the *quantity* of structure that has been mapped between two domains.

**Results.** The data for Experiment 1A can be seen in Table 2. The results of two-factor repeated-measure ANOVA testing are given below.

*SES scores:* the only significant effect was for  $\pm ST$  ( $F(1, 8) = 5.43, p < 0.05$ ). Both the  $\pm SF$  ( $F(1, 8) < 1$ ) and interaction ( $F(1, 8) = 1.24, p > 0.05$ ) factors produced nonsignificant effects.

*Match hypothesis formation:* the only significant effect was for  $\pm SF$  ( $F(1, 8) = 51.44, p < 0.01$ ). Both the  $\pm ST$  ( $F(1, 8) = 1.12, p > 0.05$ ) and interaction ( $F(1, 8) = 1.29, p > 0.05$ ) factors produced nonsignificant effects.

	LS	SS	AN	FOR
SES Category Mean	21.51	17.14	21.16	16.23
MH Category Mean	240.5	239.0	214.3	205.4

Table 2: The SES scores and number of match hypotheses formed with the SME model on the nine Original Dgroups.

**Discussion.** As expected, SME exhibits the required sensitivity to the structural commonalities of the Original Dgroups (witness the higher SES scores for the LS and AN mapping tasks). This is demonstrated by the fact that the only significant factor in the analysis of the SES scores was  $\pm ST$ . Interestingly, the number of match hypotheses formed for each category of match is sensitive to  $\pm SF$ . This reflects the fact that lexically-identical functors are more likely to occur in the Original Dgroups when there are shared surface features, and SME can only form match hypotheses between relations with lexically-identical functors.

### Experiment 1B

**Method.** The format of this experiment is the same as the previous one, except that this time SME was required to map across the Faithful Dgroups that faithfully encode the lexical properties of the original Karla representations.

**Results.** The results for Experiment 1B can be seen in Table 3. The details of repeated-measure ANOVA testing for two factors are given below.

*SES scores:* All three factors produced nonsignificant effects:  $\pm ST$  ( $F(1, 8) < 1$ );  $\pm SF$  ( $F(1, 8) = 4.72, p > 0.05$ ); and interaction effects ( $F(1, 8) < 1$ ).

*Match hypothesis formation:* Again, all three factors produced nonsignificant effects:  $\pm ST$  ( $F(1, 8) < 1$ );  $\pm SF$  ( $F(1, 8) = 3.21, p > 0.05$ ); and interaction effects ( $F(1, 8) < 1$ ).

Testing on both the SES scores ( $t = 11.37, df = 35, p < 0.01$ ) and the number of match hypotheses ( $t = 8.38, df = 35, p < 0.01$ ) revealed that there was a significant decrease in the the means of both from mapping on the Original Dgroups.

**Discussion.** As expected, SME does not exhibit the required sensitivity to  $\pm ST$  on the Faithful Dgroups, and the greatly reduced SES scores from its performance on the Original Dgroups show that it fails to map signifi-

	LS	SS	AN	FOR
SES Category Mean	1.62	1.21	1.47	0.94
MH Category Mean	92.1	84.7	93.1	78.6

Table 3: The SES scores and number of match hypotheses formed with the SME model on the Faithful Dgroups.

cant quantities of structure from one domain to another. Furthermore, the greatly reduced number of match hypotheses formed for each category of mapping (reduced from an overall mean of 224.8 in Experiment 1A to 87.13 in 1B) suggests a possible explanation of this failure: the constraints on the formation of match hypotheses are too strict to allow the appropriate local alignments to be made on the Faithful Dgroups (because there are an insufficient number of lexically-identical relations between different domains). This means that the raw material is not there for SME to combine to form the appropriate global mappings, and suggests that the process of match hypothesis formation needs to be altered if SME is to perform successfully on the Faithful Dgroups.

As noted above, the only point at which SME is committed to the CR theory is during the formation of match hypotheses. Therefore, if we are to remove SME’s commitment to the CR theory we need to do so by changing the constraints on the formation of match hypotheses to allow them to be formed between relations that are *sufficiently similar* instead of identical. This begs the question of what ‘sufficiently similar’ means.

### Co-occurrence Statistics

There is a growing body of evidence that the frequency with which different lexemes co-occur with one another (that is, are used together within a particular context, such as a paragraph or moving-window) can provide useful information about the semantic properties of those lexemes. For example, Landauer and Dumais (1997) report that the LSA model can pass a multiple-choice TOEFL synonym test. Lund, Burgess and Atchley (1995) present evidence that co-occurrence data can act as a good predictor of priming effects. Burgess and Lund (1997) demonstrate that the HAL model can produce clustering in its high-dimensional space according to the grammatical category of different lexemes.

We therefore decided to investigate the possibility of using the *Latent Semantic Analysis* (LSA) model (Landauer and Dumais, 1997; Landauer, Foltz and Laham, 1998) to see if it could provide SME with the sort of lexico-semantic information required for it to map successfully on the Faithful Dgroups (Note that although we use the LSA model, this does not indicate a particular commitment to that model alone, but rather we use it as an exemplar of the more general approach).

### Relaxing the Lexical-Identicality Constraint

Since the only commitment SME makes to the CR theory is during the formation of match hypotheses, where

it requires that relations have lexically-identical functors and the same number of arguments if they are to support a match hypothesis, SME’s code was altered so that it enforced different constraints on the formation of match hypotheses. In the modified version of SME (M-SME) two relations still have to have the same number of arguments to warrant a match hypothesis, but the *lexical-identity constraint* is relaxed. Instead of the two relations also having to have identical functors, the functors are compared with one another using the LSA model<sup>2</sup>. Only if they are assigned a score greater than a threshold value (called the *reconciliation-threshold*) is a match hypothesis formed. In this way, the relations with functors REVOLVES\_AROUND and ORBITS might be combined in a match hypothesis because the LSA model assigns them a score of 0.48.

The possibility of assigning different values (between 0 and 1) to the reconciliation-threshold generalises the original constraints that SME places on match hypothesis formation. When the threshold is set to 1 the reimplemented model performs just like the original SME because LSA only assigns lexically-identical functors a score of 1. When the threshold is set to 0 any two functors will be assigned an LSA score greater than or equal to the threshold, and so the only constraint on match hypothesis formation is that the relations in question have the same number of arguments<sup>3</sup>. It is clear that the reconciliation-threshold needs to be assigned a value that maximises the performance of M-SME.

### Setting the Reconciliation-Threshold

In order to determine a value for the reconciliation-threshold it is necessary to establish some criterion by which the *quality* of mappings can be assessed. The following experiments investigate whether such a measure can be derived from the number of match hypotheses and the SES scores of M-SME on a variety of mapping tasks.

### Experiment 2A

This experiment investigates the effect that varying the reconciliation-threshold has on the number of match hypotheses formed for each category of mapping (LS, SS, etc.). We predict that the number of match hypotheses formed for each match will decrease as the reconciliation-threshold increases because the semantic constraints on match hypothesis formation become stricter. This result will indicate that M-SME is functioning as expected. Furthermore, if the reconciliation-threshold can be used to reduce the number of match hypotheses formed then this could be used to limit the computational complexity of the mapping process.

**Method.** M-SME was used to map between the base domain and its four variants on the nine sets of Faith-

<sup>2</sup>The LSA model assigns two functors a score between 0 and 1, depending on their location in the highdimensional space defined by taking each lexeme sampled as a dimension.

<sup>3</sup>Note that the introduction of a reconciliation-threshold only affects the *formation* of mappings; the *evaluation* of mappings remains unaffected: M-SME calculates SES scores in exactly the same way as SME.

ful Dgroups, as the reconciliation-threshold was adjusted between 0 and 1.

**Results.** The results of Experiment 2A can be seen in Figure 1. The reconciliation-threshold is plotted against the number of match hypotheses formed for each category of the mapping task. This shows that the number of match hypotheses formed for each category of the mapping task decreases in a regular nonlinear fashion as the reconciliation-threshold is increased from 0 to 1.

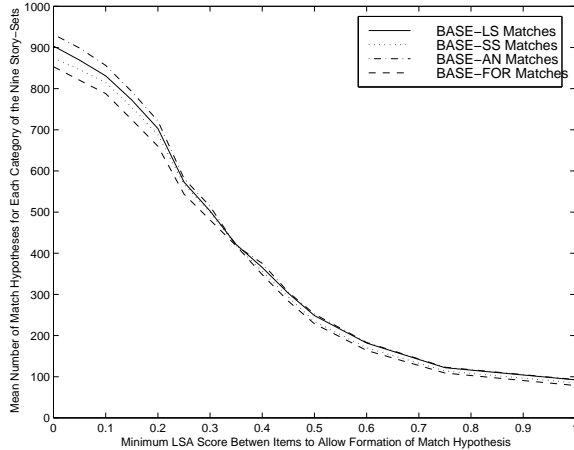


Figure 1: A plot of the number of match hypotheses that M-SME produces in matching the base stories with their four variants as the reconciliation-threshold is adjusted.

**Discussion.** The regular decrease in the number of match hypotheses formed offers preliminary evidence that M-SME is performing as expected, and that the computational complexity of the mapping process can be limited by increasing the reconciliation-threshold. However, it is possible that in doing this the semantic constraints on mappings become too strict to allow the appropriate analogical mappings to be constructed. This clearly requires further investigation.

## Experiment 2B

This experiment investigates the effect of the reconciliation threshold on the SES scores produced for each category in the standard mapping task on the Faithful Dgroups.

**Method.** M-SME was used to perform the same mapping task as in Experiment 2A, but this time the SES scores for each category were recorded as the reconciliation-threshold was adjusted from 0 to 1. We predicted that there would be a consistent separation in SES scores between those materials exhibiting  $+ST$  and  $-ST$  as the reconciliation-threshold was varied, indicating that M-SME is sensitive to the structural aspects of the Faithful Dgroups.

**Results.** The results of this experiment are shown in Figures 2-3. Figure 2 shows the SES category scores against the reconciliation-threshold. Figure 3 shows the

same data, but this time with the mapping categories split in to those which share structural commonalities with the base stories, and those which do not.

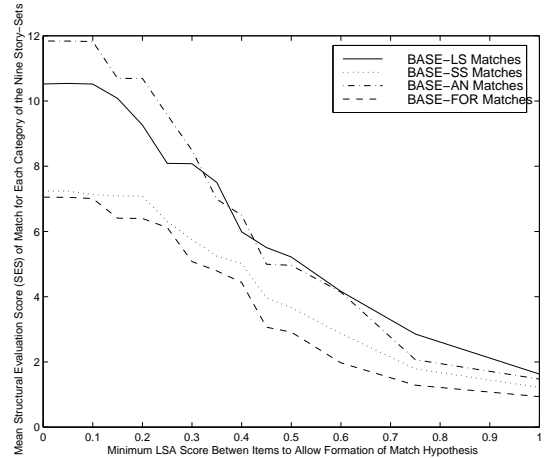


Figure 2: A plot of M-SME’s SES scores on the standard mapping task with Faithful Dgroups against the reconciliation-threshold.

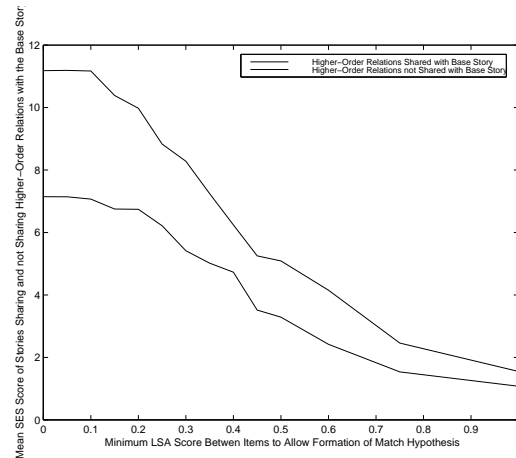


Figure 3: A plot of M-SME’s SES scores on the standard mapping task with Faithful Dgroups against the reconciliation-threshold (Dgroups are split into those exhibiting  $+ST$  and those exhibiting  $-ST$ ).

**Discussion.** Figure 3 offers preliminary evidence that M-SME is sensitive to the  $\pm ST$  factor on the Faithful Dgroups, as predicted. This represents a large improvement over SME’s performance on these materials. However, SES scores are a measure only of the *quantity* of structure that is mapped between two domains. Basing our evaluation of M-SME on SES scores alone is insufficient evidence of its success, because we need to ensure that it is sensitive to genuine analogies between domains and is not mapping inappropriate structure. So, a measure sensitive to the *quality* instead of just the *quantity*

of mapped structure is required.

### Experiment 3

To gain a useful measure pertaining to the quality of mappings made by M-SME, each of the individual alignments made in the successful global mappings were examined and rated for correctness.

**Method.** Each of the individual alignments produced by M-SME on the standard mapping task on the Faithful Dgroups were inspected and assessed for correctness (i.e. whether or not they represented genuine *analogical* alignments). The LSA score that sanctioned each alignment was also recorded, to see if the reconciliation-threshold could be set so as to prevent incorrect alignments from being made whilst still permitting correct alignments to be made.

Alignments made between the base and the SS and FOR categories were rejected, because it was unclear what would constitute a correct or incorrect alignment in these cases, as the materials were designed to share little or no structure with the corresponding base representation. The matches were performed with the reconciliation-threshold set to 0 to make the alignments generated as inclusive as possible. This was in order to collect the largest possible set of match hypotheses to see what the LSA scores were for each alignment.

Note that not *all* of the match hypotheses formed for each match were inspected, but only the ones that were included in the highest scoring global mapping for each attempted match. Although it would have been informative to consider all these hypotheses, there would have been approximately 16,200 of them<sup>4</sup>, which is too many to inspect by hand! This evaluation procedure imposes limitations on the information available. No conclusion can be drawn using this method about (i) the number of correct alignments that should have been, but are not, included within the best global mapping, and (ii) the number of incorrect alignments that are not included in the best global mapping.

**Results.** 85.99% of the alignments inspected were designated ‘correct’, whilst the remaining 14.01% were designated as ‘incorrect’. The mean LSA score between the two functors featuring in correct alignments was 0.731; the same score for incorrect alignments was 0.294. Statistical analysis showed this difference to be significant ( $t = 8.35$ ,  $df = 255$ ,  $p < 0.01$ ).

**Discussion.** The large proportion of alignments that are correct indicates that M-SME is mapping with great success on the Faithful Dgroups. The evidence of a significant separation between the LSA scores warranting the correct and incorrect alignments supports a naïve hypothesis that all match hypotheses a fixed number of standard deviations from the mean LSA score of the correct alignments could be rejected on the grounds that they are unlikely to be correct alignments. We feel that

---

<sup>4</sup>Given that there are a mean of approximately 900 match hypotheses formed (see Figure 1) for each of the 18 matches inspected (18 remain once the SS and FOR categories are discarded).

this is a bad hypothesis for the following reason:

Many functors that appear in the Faithful Dgroups are lexically-identical because they represent higher-order or structural relations that are not explicitly mentioned in the original natural language stories. For example, causal sequences and relations of temporal succession are rarely flagged explicitly in narratives, but instead have to be *inferred*. However, such relations are essential to producing the structured representations that SME and M-SME operate on. Therefore, because their lexical form is not given explicitly in the original materials they have to be assigned a canonical form (in the case of the Original Dgroups CAUSE and FOLLOWS were used chiefly). The great frequency of such functors in the Dgroups, which were generally aligned correctly, increases the mean of the LSA scores supporting correct alignments because identical functors receive an LSA score of 1. This makes the actual separation between the scores of the correct and the incorrect alignments smaller than the mean statistic indicates.

A consequence of this is that there is no one optimal value for the reconciliation-threshold that will effectively separate the correct from the incorrect alignments (because of the lack of a sufficiently distinct boundary between the two populations). Furthermore, a closer inspection of the LSA scores sanctioning correct alignments revealed that they were subject to a fairly wide distribution. If LSA is taken as a reasonable model of lexico-semantic information then this offers evidence that the relations that should be analogically aligned need not be semantically similar in a fixed way.

In this light, the nature of the structure-mapping algorithm urges caution in enforcing a prohibitively high value to the reconciliation-threshold. The structure-mapping algorithm makes match hypotheses, and combines them in an appropriate fashion to form global mappings. However, if the reconciliation-threshold is set at too high a value certain match hypotheses will not be formed. This can, in turn, inhibit further structural alignments (because match hypotheses can sanction other alignments under the *parallel-connectivity constraint*), resulting in the poor mapping performance that SME exhibits in Experiment 1B. It is sensible, therefore, to take the line of caution when it comes to setting the value of the reconciliation-threshold, and aim for a lower value that is more permissive.

The results here suggest that a suitable value for the reconciliation-threshold would be in the range 0.0-0.3. This should reduce the number of match hypotheses formed considerably (there are around 900 on average when the threshold is 0, and about 450 on average when it is 0.3; c.f. Figure 1), and thus decrease the computation required to combine the match hypotheses into global mappings, whilst preserving SES scores at a reasonable level and ensuring that a minimal number of correct alignments are prevented from being formed.

### Experiment 4

This final experiment is designed to conclusively test the mapping performance of M-SME on the Faithful Dgroups, with a fixed reconciliation-threshold.

**Method.** M-SME was used to perform the standard inter-set mapping task of Experiments 1A-B, with its reconciliation-threshold fixed to 0. The SES scores and number of match hypotheses formed were recorded for each category of match.

**Results.** The results of Experiment 4 are shown in Table 4. The results of the two-factor repeated-measure ANOVA analysis are as below.

*SES scores:* The only factor that produced a significant effect was  $\pm ST$  ( $F(1, 8) = 19.00, p = 0.02$ ). Both  $\pm SF$  ( $F(1, 8) < 1$ ) and interaction ( $F(1, 8) < 1$ ) effects were nonsignificant.

*Match hypothesis formation:* All three factors produced nonsignificant effects:  $\pm ST$  ( $F(1, 8) = 2.09, p > 0.05$ );  $\pm SF$  ( $F(1, 8) < 1$ ); and interaction effects ( $F(1, 8) = 1.40, p > 0.05$ ).

	LS	SS	AN	FOR
SES Category Mean	21.67	15.67	21.83	14.44
MH Category Mean	903.3	874.6	931.8	853.2

Table 4: The SES scores and number of match hypotheses formed with M-SME mapping on the Faithful Dgroups. The reconciliation-threshold is set to 0.

**Discussion.** The SES scores demonstrate the appropriate sensitivity to the  $\pm ST$  factor on the Faithful Dgroups, thus indicating that M-SME successfully generates analogical mappings on Dgroups that faithfully encode the lexical properties of the materials they are derived from. The number of match hypotheses is insensitive to  $\pm SF$  indicating that surface features are irrelevant to the formation of match hypotheses; this is a marked difference from the performance of SME in Experiment 1A.

## Conclusion

We have shown that SME’s commitment to the CR theory prevents it from generating analogical mappings on representations that faithfully encode lexical information (Experiments 1A-B). We then used the information provided by a co-occurrence model of semantics to produce an alternative model of analogical mapping, M-SME. Experiments 2A-B showed that M-SME functions as expected, but that there is no convenient measure of the *quality* of analogical mappings. In Experiment 3 the quality of alignments made by M-SME were inspected and rated for correctness. A detailed analysis of this data supported the idea that to maximise the quality of analogical mappings it is necessary to minimise the role that semantic constraints play during mapping. This result supports Gentner’s (1983) original insight that it is primarily *structural* constraints that determine analogical *mappings* (indeed, in Experiment 4 semantic constraints are effectively redundant in the mapping process). In the final experiment evidence was presented that M-SME is sensitive only to the structural properties of representations that faithfully encode lexical properties. Because a commitment to semantic links or the CR theory allows

coding decisions to reduce the search space that analogical mappers face, it is significant that M-SME can still produce mappings when presented with problems of this greater complexity. Whilst M-SME is a more expensive mapper overall, we think that a similarly improved model of the *retrieval* of analogies may enable the use of *contextual* information to reduce the search space in the mapping phase.

## Acknowledgements

We would like to thank Andrew Wishart for helpful comments on an initial draft of this paper.

## References

- Burgess, C., and Lund, K. (1997). Modelling Parsing Constraints with High-Dimensional Context Space. *Language and Cognitive Processes*, **12**:177-210.
- Falkenhainer, B., Forbus, K.D., and Gentner, D. (1989). The Structure-Mapping Engine: Algorithm and Examples. *Artificial Intelligence*, **41**:1-63.
- Forbus, K., Gentner, D., and Law, K. (1994). MAC/FAC: A Model of Similarity-based Retrieval. *Cognitive Science*, **19**:141-205.
- Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, **7**:155-70.
- Gentner, D., Ratterman, M., and Forbus, K. (1993). The Roles of Similarity in Transfer: Separating Retrievability from Inferential Soundness. *Cognitive Psychology*, **25**:524-575.
- Holyoak, K.J., and Thagard, P. (1989). Analogical Mapping by Constraint Satisfaction. *Cognitive Science*, **13**:295-355.
- Holyoak, K.J., and Thagard, P. (1995). *Mental Leaps: Analogy in Mental Thought*. Cambridge, MA: MIT Press.
- Hummel, J.E., and Holyoak, K.J. (1997). Distributed Representations of Structure: A Theory of Analogical Access and Mapping. *Psychological Review*, **104**(3):427-66.
- Landauer, T.K., and Dumais, S.T. (1997). A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, **104**: 211-40.
- Landauer, T.K., Foltz, P.W., and Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes*, **25**:259-84.
- Lund, K., Burgess, C., and Atchley, R.A. (1995). Semantic and Associative Priming in High-Dimensional Semantic Space. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 660-65). Pittsburgh, PA: Erlbaum.

# Selective advantages of syntactic language — a model study

Willem H. Zuidema\*

jelle@csl.sony.fr

Theoretical Biology, Utrecht University  
Padualaan 8, 3584 CH Utrecht  
The Netherlands

Paulien Hogeweg

p.hogeweg@bio.uu.nl

Theoretical Biology, Utrecht University  
Padualaan 8, 3584 CH Utrecht  
The Netherlands

## Abstract

We study a computational model of the evolution of language in groups of agents to evaluate under which circumstances syntax emerges. The fitness in the model depends on the composition of the population. We find that this fact significantly alters the evolutionary dynamics. If scores are attributed to both speaker and hearer, expressive syntax is hard to obtain. If scores are attributed only to the hearer, syntax develops, but agents lose the willingness to speak. Implications and a possible solution of this paradox are discussed.

## Introduction

Among the many differences between human language and other animal communication systems, syntax is widely acknowledged to be particularly important. Syntax allows us to combine a finite set of meaningful units into an unbounded set of combinations. It allows us to speak about events happening at other times and places. It allows us to communicate about causal relations, to phrase questions or imperatives, and to share in detail previous experiences. The emergence of syntactic language is therefore considered to be one of the major transitions in evolution (Szathmáry & Maynard-Smith, 1995).

In the traditional view, syntax reconciles the need for high expressiveness with some of the natural boundary conditions on communication such as memory limitations, errors in distinguishing sounds, or bottlenecks in the transmission of language knowledge. However, present-day language fulfills many more functions than exchanging information, including facilitating social relations, individual expression, increase of status, esthetic experience and perhaps internalizing our knowledge of the world. It is unclear in what way such functions are recent side-effects, or play an important role in explaining the origins of language.

Discussions of such issues tend to be very unsatisfactory, because they seem hardly restricted by empirical or theoretical bounds. *Computational modeling* offers a

novel approach to these issues, because such models are at least restricted by whether or not the *combination* of assumptions implemented in the model yield the hypothesized outcome: syntactic language. This paper discusses a simple computational model of an evolving group of communicating individuals and studies under which selection pressures expressive, syntactic language arises. Before describing the model architecture and results, we will first briefly discuss the theoretical background and some related work.

## Evolution of language

Probably the most well-known speculation on the origins of human language is the paper of Pinker & Bloom (1990). Pinker & Bloom argue that syntax must originate in a process of evolutionary optimization, because “natural selection” is the only explanation for the origins of complex design in nature. The paper brings together a valuable collection of findings, but from a theoretical perspective it is problematic, because it lacks precision and formalization. In its weakest interpretation the central claim is trivial (there is no doubt that only members of the human species can acquire fluency in a human language) and in its strongest interpretation (“evolution has led to genes that explicitly specify a universal rule system for language”) the claim is untenable. However, the lack of a more precise aspect to Pinker & Bloom’s work, makes it hard to position their ideas between these extremes.

Moreover, Pinker & Bloom’s paper is symptomatic for the popular fallacy in linguistics that one can only choose between two explanations: (i) language originates in a genetic evolution, or (ii) language arises as the spontaneous result of general cognitive skills and social structure. We believe that putting these two explanations in opposition, excludes the most interesting part of the story. Spontaneous pattern formation (“self-structuring”) needs a mechanism to set the right parameters, and evolution needs a plausible substrate to operate on. Viewing self-structuring as a substrate for evolution (Boerlijst & Hogeweg, 1991a) offers a fresh perspective that allows one to study how evolution, genetic information, learning, development, embodiment and social structures all interact to shape

---

\*Present address: Sony CSL, 6, Rue Amyot, 75005, Paris, France; webpage: [www-binf.bio.uu.nl/~jelle](http://www-binf.bio.uu.nl/~jelle)



human language. Note that such an interactionist account differs fundamentally from a naive “some parts of language are innate, some are learned” view.

## Computational modeling

Recent work that studied such interactions in computational models has produced a wealth of new hypotheses and insights (Hurford, 1989; Hashimoto & Ikegami, 1996; Batali, 1997; Steels, 1997; De Boer, 1999; Kirby, 2000; Nowak & Krakauer, 1999; Hurford, 2000). Such models are *relatively precise* implementations of the underlying set of assumptions, and allow one to evaluate the internal coherence of such a set. Moreover, they are *productive*, in the sense that they often show unexpected behaviors that help to generate new hypotheses and concepts. And although they are necessarily simplified representations, the fact that their behavior can be experimentally evaluated makes it possible to study more complex phenomena than with analytical methods alone. Computational models therefore pre-eminently can make tractable systems with many variables and interactions.

On the issue of the origins of syntax, a number of intriguing mechanisms have been identified using computational modeling techniques. Although very diverse, they all emphasize the fact that syntax greatly increases the number of possible forms in a language. For instance, Batali (1997), Kirby (2000) and Hurford (2000) studied how *cultural evolution* can account for the emergence of syntax. Although they use several different formalisms, the common idea in this work is that the internal knowledge of language (the infinite “I-language”) is transmitted culturally (via a finite “E-language”) from one agent to another. This “transmission bottleneck” works as a filter, in which syntactic elements of language typically out-compete non-syntactic elements, because the former are inherently used more often.

Nowak & Krakauer (1999) studied a game-theoretic model of language evolution and identify a different mechanism that can account for the emergence of syntax. Using the matrix representations of Hurford (1989), they infer a “linguistic error limit”. Given that an individual makes mistakes in distinguishing sounds with a probability that depends on the similarity between those sounds, Nowak & Krakauer calculate a limit on the number of messages an individual can convey. They show mathematically that *word formation* and *syntax* can help overcome such a limit. Moreover, they show that both non-syntactic and syntactic strategies are *evolutionary stable strategies* (i.e. cannot be invaded by other strategies). However, every mixed strategy can be invaded by every mixed strategy that uses *more* syntactic sentences. Thus, the evolutionary process should lead towards grammar.

Hashimoto & Ikegami (1996) showed that syntax can emerge in an *evolving group* of communicating agents.

The agents in their model have an internal rewriting grammar, that generates a formal language using lexical or syntactic strategies. Because there is no limit on the number of rules, both strategies could in principle generate all possible strings in the finite domain that was used. However, at the start of the simulations agents are initialized with just one rule in their grammar. Because mutations add rules one at a time, and expressiveness grows much faster with grammar size using a syntactic strategy, syntactic agents out-compete non-syntactic ones.

An important aspect of Hashimoto & Ikegami’s model is that fitness is not a fixed measure, but depends on the kind of grammars that are present in the population. This leads to some counterintuitive results. For instance, they find that the most expressive agents are not necessarily the most successful and that a score for *not being recognized* accelerates the evolution of syntax. These observations are the starting point for the model study reported in this paper.

## The model

The model reported in this paper is a variant of the model of Hashimoto & Ikegami (1996). Of the many aspects that might be relevant, we study only one particular type of interaction: between evolutionary dynamics and group structure. We therefore ignore all aspects of grammar, except for the fundamental properties of compositionality and recursion. We ignore semantics, by just attributing scores for successful parsing. And we ignore learning, by assuming that agents end up with the same internal grammar, except from some changes that result from mutations in the innate component of language.

In this simplified model we will show that evolution shapes the linguistic environment of agents, but, conversely, that the group structure also shapes the evolutionary process. This interaction guides evolution in unexpected directions, and, depending on the implemented function of language, can both facilitate and hinder the development of syntax.

The model consists of a population of agents with an internal rewriting grammar, which they inherit with some mutations from their parent. The grammars are context free grammars, with nonterminal and terminal symbols from the small alphabets  $V_{nt} = \{S, A, B\}$  and  $V_{te} = \{0, 1\}$  respectively. As an extra restriction, we don’t allow the “S” at the right-hand sides of rules. At the start of most simulations, agents are initialized with a grammar with just one rule: randomly  $S \mapsto 0$  or  $S \mapsto 1$ . Agents have the ability to derive (“speak”) and parse (“understand”) strings of 0’s and 1’s of maximum length 6, using the rules from the grammar. Within these constraints the maximum expressiveness is 126. We define compositionality as using the non-terminals A and B, and recursion as using rules that were used before in the same branch of the rewriting tree.

Agents interact in a set-up of “language games”. In every game all agents can speak one string and try to recognize the strings produced by other agents. Every generation a number of games is played and scores are attributed for successful communication. In most simulations, we use an explicit “innovation pressure”. This pressure is implemented by discounting scores with the number of times a string is already heard before, and corresponds to a semantic need for a rich repertoire of forms. We designed several scoring schemes that reflect hypotheses on the function of language. The most important schemes are labeled “communication” and “perception”:

**communication** corresponds to a selection pressure to optimize the total of exchanged information, such that both the speaker and the hearer benefit from successful communication. This pressure is implemented by a score for recognition and for being recognized.;

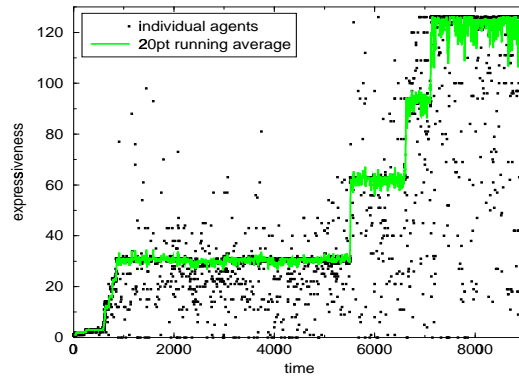
**perception** corresponds to a selection pressure to optimize the total of information received, in order to make use of the knowledge of others (as if one indirectly shares someone else’s *perception*). This pressure is implemented as a score for recognition;

We replace all agents every generation with offspring of the present population. The number of offspring of an agent depends on the total score it has received relative to other agents. Random mutations are applied to the offspring with fixed probabilities for modification of existing rules (“replace”), duplication of a random rule (“add”) or deletion of a rule (“delete”). We also implemented a mutation “shift”, that swaps a rule with the previous rule in the grammar and occurs with a probability per rule. These mutations correspond to conventions in evolutionary programming and allow for optimizing some of the relevant features of grammars, but otherwise they are more or less arbitrary.

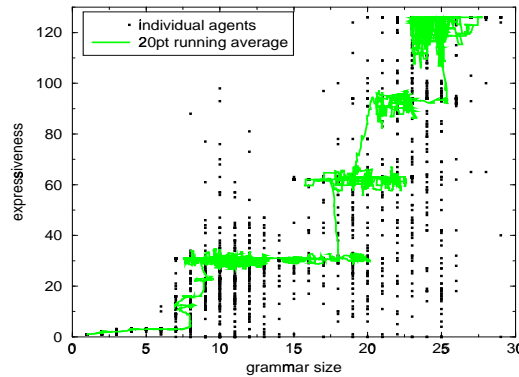
### The group effect

Fitness in this model is not a static function of an agent’s grammar (“genotype”), but it depends on the grammars of other agents too. The general observation in experiments with the model with many different parameter settings is that this fact strongly influences the evolutionary dynamics (Hashimoto & Ikegami, 1996). The success of an agent’s individual language is determined by how well it matches the language of the whole group, rather than by how much information it can encode (“expressiveness”). We call this phenomenon the “group effect”.

Figure 1(a) shows an example simulation, with a “communication” scoring scheme and “innovation”, that shows clearly some of the mechanisms that play a role. From the initial level of expressiveness of 1, the



(a) Expressiveness over 9000 generations



(b) The same run in a “phase space”

Figure 1: An example run with very clear epochal evolution. Shown are the running averages and individual agents at every tenth generation. Note that most individual points are hidden under the grey line. (a) During an epoch, expressiveness stays at a fixed level. In fact, in the first stage ( $E=31$ ) the dominant language stays exactly the same for thousands of generations. Individual agents with higher expressiveness occur, but are not able to survive in the group. (b) Grammars do vary, however, which is possible because of the neutrality in the grammar–language mapping (see text). In the phase space, one can clearly see that grammar size fluctuates during an epoch. All jumps to higher levels take place when grammars are relatively large. Such grammars are clearly larger than necessary and have a neutral tails. Parameters: default “communication” run with innovation pressure (see section “selective advantages”)

population evolves within several hundreds of generation to a level of 31. At this point, evolution has developed via selection and random mutations grammars that are redundant and not very structured, and combine several strategies in the rewriting process from the start symbols “S” to a distinct sequence of terminal characters.

For a very long time, from around generation 860 until 5510, the population remains fixed at a level of expressiveness of 31. Analysis of the language reveals that the set of strings of the majority of agents remains unchanged for this whole period. However, frequently agents appear that have a much higher level of expressiveness. This illustrates that (i) the mapping from grammar to language is very non-linear, because a single mutation can make a dramatic change in the size of the language, and (ii) there is a very strong group effect, because agents that have a much higher expressiveness (and thus are “objectively” much better), can nevertheless not persist in the population. The reason is that the languages of these agents differ too much from the language of the group. The agents therefore obtain fewer scores for being recognized and possibly even for recognizing.

Another striking observation in this simulation is that, although the languages remain unchanged for several thousands of generations, the grammars undergo a constant reorganization. This illustrates that the mapping from language to grammar is not only non-linear, but also very redundant.

Figure 1(b) shows a graph of the same simulation in a “phase space” that shows the average grammar size versus the average expressiveness at each generation. As one can clearly see, once a certain level of expressiveness is reached, the evolutionary process “wanders around” for a long time, without significant changes in the expressiveness (“neutrality”). Only when the grammars are relatively large, and thus have many unused, redundant rules, a chance event causes the population to jump to a new level of expressiveness. This chance event is that two agents mutate to the same richer language, and thus can obtain in their mutual communication enough scores to compensate for differing from the group. This mechanism relates to the idea of “neutral networks” — networks of connected points in genotype space that correspond to the same phenotype — that forms a good explanation for the occurrence of “epochs” or “punctuated equilibria” in evolving systems with a fixed fitness function (Van Nimwegen *et al.*, 1999).

### Selective advantages

While the “group effect” occurs under all parameter settings of the model, its role can be quite different for each of the scoring schemes and the initial grammars we considered. We observe compositional and recursive grammars only in about half of the parameter combinations we considered. Even if scores are explicitly discounted with the number of times a string is already used before (“innovation pressure”), expressive syntax does not necessarily emerge.

This fact is surprising, because the intuitive expectation is that expressiveness is selectively advantageous. Indeed, with (i) an *explicit* innovation pressure, the

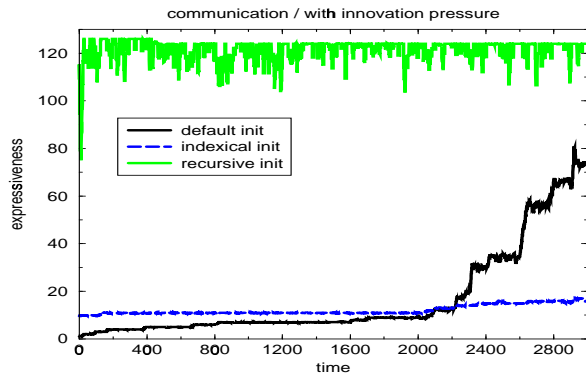


Figure 2: *Communication with innovation pressure for three different types of initial grammars. With a sufficiently large initial lexical grammar, expressive syntax can not develop.*

average score per agent has its optimum at maximal expressiveness. However, *implicitly* expressiveness influences the scores in other ways as well: (ii) expressive speakers are more likely not to be understood, and (iii) expressive listeners are more likely to understand.

This leads to an interesting interplay between each of these roles of expressiveness and the group effect. Under communication settings (ii) not being recognized is *disadvantageous*, while (iii) recognition is *advantageous* and in both scoring dimensions similarity to the group’s language is important. Under perception settings (ii) not being recognized and (iii) recognition are beneficial, while similarity to the group’s language is important for recognition, but *dissimilarity* is better for not being recognized (and thus hindering one’s competitors). Moreover, the strength of the group effect depends on the size of the group’s language and the variation within the group. In various experiments we obtained the following results:

**communication** does not lead to highly expressive grammars with the default initial grammar and without the innovation pressure. If the initial grammar is an expressive, recursive grammar, the high level of expressiveness can be maintained. In contrast, with a medium size lexical grammar, grammars remain lexical and expressiveness remains limited.

With an innovation pressure and the default initialization expressive syntax eventually does develop. In this type of runs we observe a stepwise development, with typically long intervals at the same level of expressiveness. Expressive syntactic grammars are reached only after very many generations. With an expressive, recursive initial grammar, the high level of expressiveness can be maintained. With a medium size lexical grammar expressiveness remains limited and no syntax develops (see figure 2).

With “communication” as the function of language,

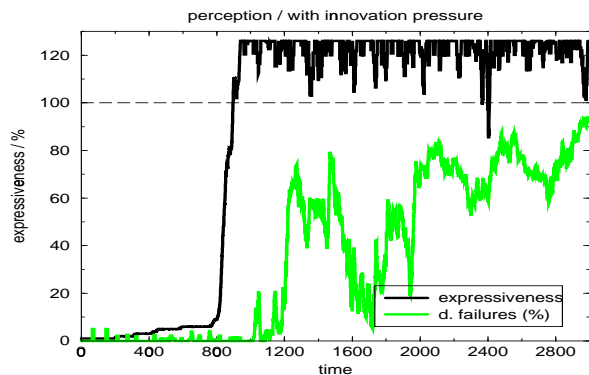


Figure 3: A typical example of a simulation with “perception” settings, the default initial grammar and an innovation pressure. Shown are the average expressiveness over time, and the percentage of failures in derivation. After around 3000 generations this percentage approaches 100, indicating that very little communication is maintained.

syntax can thus be maintained if present, but is hard to obtain. If the initial grammar is of sufficient size and of a lexical type, syntax never develops. These results are particularly interesting, as they resemble the situation that is traditionally thought to precede the emergence of grammar: large, lexical protolanguages, with communication benefits for both speaker and hearer.

**perception** shows rapid growth in expressiveness in most cases considered. With the default initialization and no innovation pressure, expressive syntax develops within a few hundred generations. With the lexical initialization it takes much longer, but the development of syntax was usually observed.

With an innovation pressure and default initial grammars the growth is generally slower than without such an innovation pressure. Infrequently, we even observe runs that remain lexical throughout the simulation. When initialized with an lexical grammar, the runs with innovation pressure show such behavior.

“Perception” thus yields expressive syntax in most cases considered (see figure 3). The benefits of *not being understood* seem to be a strong incentive to develop more expressive language. Interestingly, an innovation pressure makes the development of syntax *less* likely. Apparently, the fact that the hearer benefits from richer input hinders this development.

## Paradox

Another striking feature of perception runs is the high number of failures that occur in derivation (see figure 3). Apparently, agents develop grammars that are able to parse a high number of strings, but nevertheless frequently fail in derivation. This is possible because of the asymmetry in parsing (complete bottom-up search of the derivation tree) and derivation (random top-

down walk). This possibility was not implemented intentionally. Nevertheless, the evolutionary process discovered it and “actively” exploits it.

This observation points at a important assumption in the model: agents are forced to participate in the language game. A classic altruism problem thus arises: if speaking behavior is beneficial only for an individual’s competitors, why would it be retained in evolution? We extended the model with a parameter for probability to speak. Under perception settings this parameter indeed quickly evolves to zero.

Interestingly, these results constitute a paradox: under those circumstances that syntactic expressiveness develops, willingness to speak disappears. Under the circumstances where willingness to speak is retained, syntactic language does not develop. We studied a possible solution for this paradox in a model where agents are localized on a 2D grid and interact only with their immediate neighbors. Such spatial models are known to naturally yield altruism, because spatial patterns make multilevel evolution possible and kin selection more likely (Boerlijst & Hogeweg, 1991b).

The willingness to speak can be retained in the spatial model with perception settings. The parameter that determines the probability of an agent to speak at its turn in the language game, is initialized at 0.1. As one can see in the example of figure 4, the average value rapidly evolves to a high value close to the maximum. Spatial patterns are responsible for this selection pressure towards altruistic behavior. If one destroys the spatial patterns, also the willingness to speak disappears.

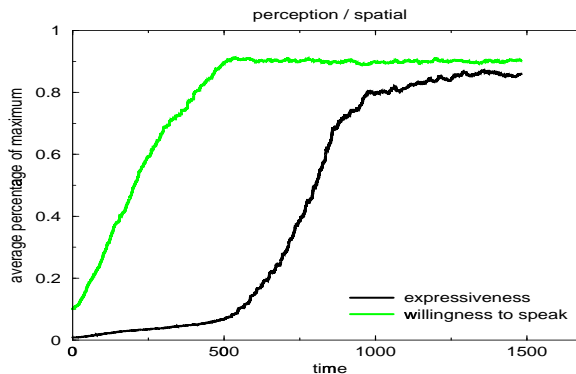


Figure 4: Perception in space. Shown is the average fraction of the maximum of expressiveness (maximum is 126) and willingness to speak (maximum is 1). Parameters are: initial population size = 2000, number of games per generation = 1, maximum string length = 6, minimum number of understanders = 0,  $m_{add} = 0.1$ ,  $m_{rep} = 0.01$ ,  $m_{del} = 0.01$ , maximum number of parsing steps = 500, maximum number of derivation steps = 60, self-interaction not allowed, discount factor 1.0, scores proportional to string length

## Discussion

Some of the striking differences in the results of different scoring schemes can be better understood by looking at a very simple game-theoretic model, where there are just two agents and two levels of expressiveness. If we work out the language games that take place in such a set-up, we find that both the low/low and the high/high situations are equilibria in the communication case, but in the case of perception only the high/high situation is an equilibrium. These results qualitatively corresponds to the results we obtained in the simulations.

The essential observation here is that, although homogeneous high expressiveness is the “best” solution, *unilateral* high expressiveness under communication setting is in fact disadvantageous. It seems a promising approach to extend this game-theoretic analysis to a more general case, with more levels of expressiveness and more interacting agents. However, many aspects of the model behavior depend on the non-linear mapping between grammar and language and can not easily be captured in such an analysis.

## Conclusions

Traditionally the origins of language are thought to be explained as either the spontaneous result of human cognitive abilities and social interactions, or the result of an evolution of our innate language capacity. This model study shows an example system where both social interaction and evolutionary updating play a role. Not because one part of language can be explained by “nurture” and another part by “nature”, but because they fundamentally interact: social interactions shape the evolutionary process and vice versa.

Also, traditionally language and the evolution of language are studied in terms of how much information about the outside world can be transmitted. Our results suggest that this might not always be the most interesting way of looking at language, because language can have its own dynamics within a group that is quite independent from how well it represents the outside world.

Moreover, this model study shows results that deviate from the traditional picture that lexical protolanguages became larger and larger until syntax became necessary. If communication is beneficial for both speaker and hearer and the population uses an extensive lexical language, syntax does not develop. If the traditional picture holds, the question arises which mechanisms are responsible for the differences.

Finally, spatial patterns have not played much of a role in speculations about the origins of language. Results from this study suggest that such spatial patterns can be relevant. The fact that present-day language shows obvious spatial patterns indicates that a global approximation perhaps excludes important

mechanisms.

Many open questions remain. For instance, under perception settings there is an *indirect benefit* of speaking that leads to high values of the willingness to speak. Why then, does this indirect benefit not result in the same disadvantage of unilateral high expressiveness that we observe under communication settings? Such intriguing issues are left for future work.

## Acknowledgements

Many thanks to Ludo Pagie for help with programming C++, and Michael Moortgat, Onno Zoeter, and all members of the Theoretical Biology group for technical and moral support and stimulating discussions. WHZ thanks Sony CSL-Paris for allowing him the time to write and present this paper.

## References

- BATALI, J. (1997). Computational simulations of the emergence of grammar. In: *Approaches to the evolution of language* (Hurford, J. et al., eds.). Cambridge University Press.
- BOERLIJST, M. C. & HOGEWEG, P. (1991a). Self-structuring and selection. In: *Artificial Life II* (Langton, C. et al. eds.), 255–276.
- BOERLIJST, M. C. & HOGEWEG, P. (1991b). Spiral wave structure in pre-biotic evolution. *Physica D* **48**, 17–28.
- DE BOER, B. (1999). *Self-Organisation in Vowel Systems*. Ph.D. thesis, Vrije Universiteit Brussel AI-lab.
- HASHIMOTO, T. & IKEGAMI, T. (1996). The emergence of a net-grammar in communicating agents. *BioSystems* **38**, 1–14.
- HURFORD, J. (1989). Biological evolution of the sausage sign as a component of the language acquisition device. *Lingua* **77**, 187–222.
- HURFORD, J. R. (2000). Social transmission favours linguistic generalization. In: *The evolutionary emergence of language* (Knight, C. et al., eds.). C.U.P.
- KIRBY, S. (2000). Syntax without natural selection. In: *The evolutionary emergence of language* (Knight, C. et al., eds.). C.U.P.
- NOWAK, M. A. & KRAKAUER, D. C. (1999). The evolution of language. *Proc. Nat. Acad. Sci. USA* **96**, 8028–8033.
- PINKER, S. & BLOOM, P. (1990). Natural language and natural selection. *Behavioral and brain sciences*
- STEELES, L. (1997). Synthesising the origins of language and meaning. In: *Approaches to the evolution of language* (Hurford, J. et al., eds.). C.U.P.
- SZATHMÁRY, E. & MAYNARD-SMITH, J. (1995). The major evolutionary transitions. *Nature* **374**, 227–232.
- VAN NIMWEGEN, E., CRUTCHFIELD, J. & HUYNEN, M. (1999). Neutral evolution of mutational robustness. *Proc. Nat. Acad. Sci. USA* **96**, 9716–9720.

**This page left blank intentionally.**

**This page left blank intentionally.**

# A Natural Bias For the Basic Level?

Annie Archambault (ANNIE@PSY.GLA.AC.UK)  
Frédéric Gosselin (GOSSELIF@PSY.GLA.AC.UK)  
Philippe G. Schyns (PHILIPPE@PSY.GLA.AC.UK)  
Department of Psychology, University of Glasgow  
58 Hillhead St., Glasgow G12 8QB UK

## Abstract

It is well established that people can categorize the same objects at different levels of abstraction (i.e., superordinate, basic, and subordinate). Of these, the basic level is known to have a privileged status that is often attributed to the organization of categories in memory. Here, we argue that the bias could in part arise from the image formation process itself—i.e., the object properties for categorization that arise from the 2D retinal projections of distal 3D objects. In the real world, people do categorize objects from a variety of viewing distances and these modify the availability of object information on the retina. In two experiments, we tested the hypothesis that the information for basic categorizations is more resistant to changes in viewing distance than that of subordinate categorizations.

Casual observers would experience little difficulty to categorize the animals in Figure 1 as exemplars of *dog* and those of Figure 2 as exemplars of *whale*. If they were “experts”, they could categorize these animals as *Saint-Bernard dog*, *Doberman dog*, *Sperm whale*, and *Humpback whale*. People can similarly apply different levels of category abstraction to the 3D distal objects that impinge on their retina.

Rosch et al.’s (1976) seminal research isolated three “natural” levels of object categorization: the superordinate (*animal*, *vehicle*, *furniture*), the basic (*dog*, *car*, *chair*), and the subordinate (*Saint-Bernard dog*, *Porsche*, *Chippendale chair*). Of these, the basic and subordinate are thought to be closer to perception and we will focus on their main differences. The former level is superior to the later in a number of ways:

(1) Categories at the basic-level are verified fastest (see also Hoffmann & Ziessler, 1983; Jolicoeur, Gluck & Kosslyn, 1984; Murphy, 1991; Murphy & Smith, 1982; Murphy & Brownell, 1985; Tanaka & Taylor, 1991).

(2) Objects are named faster at the basic than at the subordinate level (Hoffmann & Ziessler, 1983; Jolicoeur, Gluck & Kosslyn, 1984; Murphy, 1991; Murphy & Smith, 1982; Murphy & Brownell, 1985; Rosch et al., 1976; Tanaka & Taylor, 1991; Johnson & Mervis, 1997).

(3) Objects are preferentially designated with their basic-level names (Berlin, 1992; Brown, 1958; Rosch et al., 1976; Tanaka & Taylor, 1991; Wisniewski & Murphy, 1989).

(4) Throughout development, basic names are learned before subordinate names (Anglin, 1977; Brown, 1958; Rosch et al., 1976; Horton & Markman, 1980; Markman, 1989; Markman and Hutchinson, 1984; Mervis and Crisafi, 1982).

(5) Basic names tend to be shorter (Brown, 1956; Rosch et al., 1976).

The origin of the bias to the basic level is still a matter of debate. In categorization, researchers have proposed that categories at the basic level are more *differentiated* that is, “... have the most attributes common to members of the category and the least attributes shared with members of other [contrasting] categories.” (Rosch et al., 1976, p. 435) The first component of this differentiation definition has been called the *specificity* (Murphy & Brownell, 1985), or the *informativeness* (Murphy, 1991) of a category, and the second component the *distinctiveness* of a category (Murphy & Brownell, 1985; Murphy, 1991). The difference between basic and subordinate categorizations would thus stem from distinct differentiations at these two levels. But the origin of these remain unspecified.

In recognition, researchers have sought to ground the basic level advantage on object properties (i.e., feature content). Rosch et al. (1976) found that basic-level categories are the most inclusive categories at which objects look alike. This suggests that shape is an important factor in the advantage of the basic over the subordinate level. One determinant of shape is part structure. Tversky and Hemenway (1984) found—for a broad range of natural categories including objects and organisms—a little increase in the number of listed parts from the basic to the subordinate level. Parts could therefore be a main determinant of basic-levelness. Jolicoeur, Gluck, & Kosslyn (1984) proposed that objects are initially recognized at the basic level on the basis of their parts, but also that these parts index the entry point to recognition. Entry point categories are usually at the basic-level but not always. To access categories below the entry point, such as Rosch’s subordinates, *additional perceptual information* is required (see also Biederman, 1987). This additional information was, however, left unspecified. Reflecting on the state of the art in object recognition, it is fair to say that the relationships between the basic level preference and its perceptual determinants are at a standstill.

From this brief review of the literature, two main stances emerge regarding the advantage of basic level over the subordinate categorizations:

(1) Categorization researchers have argued that the organization of categories in memory produces the faster access to the basic level (e.g. Murphy, 1991).

(2) Recognition researchers have proposed that categorization is faster at the basic level because the visual system is geared to extract parts from the input, and parts represented categories at the basic level (e.g. Biederman, 1987).

We will here present and test a third, and possibly simpler alternative: The bias for the basic level could arise



from natural constraints on the image formation process that modifies the perceptual availability of object cues with changes of viewing distance.

People who recognize common objects tend to do so over a wide range of viewing distances. For example, you need to recognize your car at a distance in a parking lot, but you also need to recognize it from a closer range, when you are about to unlock its door.

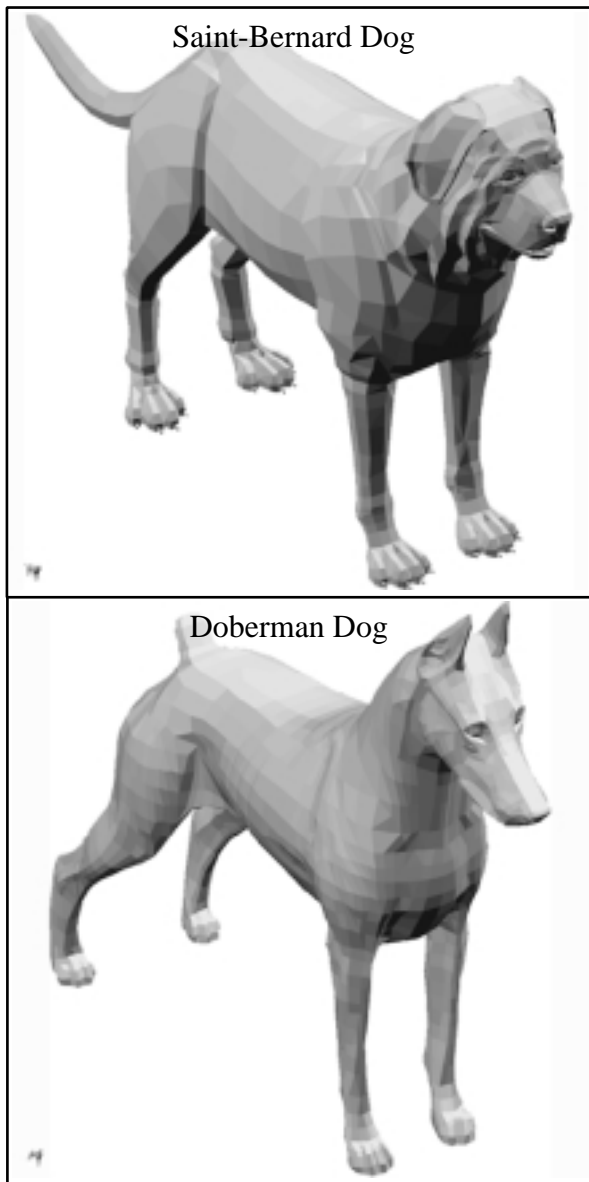


Figure 1. Three-quarter right views of the Saint-Bernard and the Doberman Dogs used in experiments 1 and 2. The figure respects the proportions of the stimuli, not their absolute sizes: the large animals occupied 12 deg of visual angle; the small ones (see dark spots at the bottom-left of each large animal) .38 deg.

However, a simple computational argument can be made that changing the size of the retinal projection also changes the information available in the image for identification. Simply put, reducing the retinal projection of an object by a factor of two reduces its sampling frequency by the same factor. Simplifying a little, if one starts with a 512x512 original image, the reduction samples one pixel every other pixel to produce a 256x256 image.

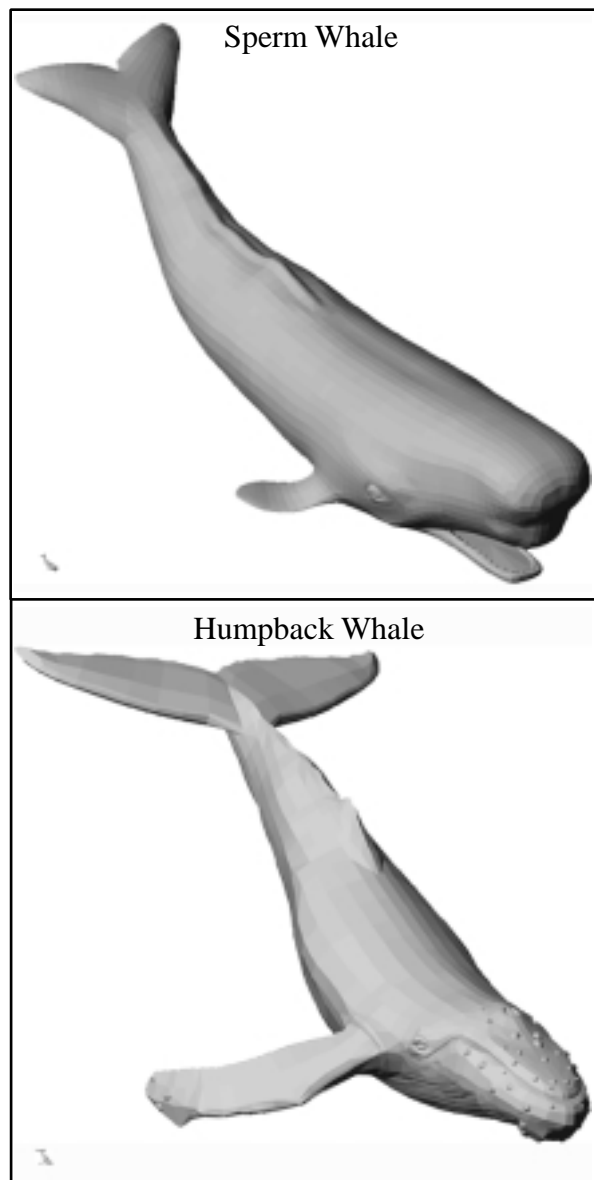


Figure 2. Three-quarter right views of the Sperm and Humpback Whales used in experiments 1 and 2. The figure respects the proportions of the stimuli, not their absolute sizes: the large animals occupied 12 deg of visual angle; the small ones (see dark spots at the bottom-left of each large animal) .38 deg.

Any variation that was expressed between any two adjacent pixels of the original image (e.g. a black and white contrast) is lost in the reduced image. (Technically, shrinking an image eliminates its high spatial frequency information). This produces a marked loss in information for visual categorization.

If we hypothesize that different basic and subordinate categorizations require visual information that resides at different scales of the stimulus, changing the scale of one stimulus could produce markedly different patterns of categorization performance. For example, removing large-scale information by reducing the size of objects could selectively impair the categorization level requiring the most specific details—i.e. the subordinate level. If this were the case, a bias to the basic level could arise from the statistics of categorization attributes over a wide range of viewing distances. Specifically, the attributes

that access the basic level could have a greater resilience over scale changes than those accessing the subordinate level. This natural bias on the availability of perceptual cues would shed a new light on the structure of the basic level. We would still not know exactly what is this “additional information” required for the subordinate, but we would know that it is resistant to scale changes.

## Experiment 1

Experiment 1 was designed to investigate the interaction between the scale of objects and their levels of categorization. Stimuli were three-dimensional (3D) gray-level computer synthesized animal categories (*bird, cow, dog, horse, frog, turtle, spider* and *whale*) (e.g., see figures 1 and 2). A similarity judgment task required participants to establish whether two simultaneously presented animals were the same either at the BASIC level (e.g., are both pictures *cow* exemplars?) or at the SUBORDINATE level (e.g., are the two pictures the same cow?). The animal pairs represented either two identical individuals (e.g., the same cow), two members of the same animal category (e.g., two different cows) or two members of a different animal categories (e.g. a cow and a bird). Animal pairs could appear in one of six possible different sizes. Each pair stayed on the screen as long as participants deemed it necessary (self-paced judgments). If the perceptual cues needed for BASIC and SUBORDINATE judgments are available regardless of the scale of the objects, participants should not differ in performing basic and subordinate similarity judgments. However, if information differs for BASIC and subordinate categorizations, a reduction in stimulus size might differently affect performance.

It is important to stress that this task involves absolute levels of information. That is, participants can use all the information present in a stimulus pair, as the two animals remained on the screen until a similarity judgment was made. Failure to notice a difference in these conditions would imply that the required information had vanished.

### Participants

Twenty Glasgow University students with normal or corrected vision were paid to participate in the experiment.

### Stimuli

Stimuli were computer-synthesized 3D animals. The set of animals was composed of 8 different animal categories (*bird, cow, dog, horse, frog, turtle, spider* and *whale*), each comprising 2 different exemplars. All animals were presented at one of six different sizes. The largest size corresponded to 512 square pixels and the smallest one to 16 square pixels. Successive divisions (by 2) of the largest pictures produced all intermediate sizes. These sizes were 256, 128, 64, and 32 square pixels. They corresponded to about 12, 6, 3, 1.5, .75 and .38 degrees of visual angle, respectively. In total, 96 stimuli were created (8 animal categories \* 2 individuals \* 6 sizes). In addition, each object could be presented from two different viewpoints (separated by 95 degrees of rotation in depth), so that when two objects appeared in a pair they would never be strictly identical pictures and people would need to recognize the represented animals to judge their similarity.

### Procedure

Before starting the experiment, each participant was instructed that they needed to make two different types of similarity judgments. To the question “Same animal category?” participants had to judge whether the two presented animals belonged to the same animal category (e.g., are both animals dogs?). To the question “Same individual?” the task was to decide whether both animals were the same exemplar (e.g., are both dogs the same individual?). Participants were told to take as long as they wished and to look very carefully at each animal pair before making a decision.

A trial started with the apparition of one animal pair on the computer monitor. The two animals appeared simultaneously and were always pictured from a different viewpoint. Participants could observe the animal pair for as long as they wished. A keypress would substitute that animal pair with a question of the screen. The question was either “Same animal category?”, “Different animal categories?”, “Same individual?” or “Different individuals?”. They then entered their judgment by pressing “yes” and “no” keys on the computer keyboard.

Experiment 1 comprised 4 main classes of trials depending on whether there was a match (vs. non-match) at the BASIC (vs. SUBORDINATE) level. Match trials at the BASIC level represented different animals from the same animal categories (e.g., two different dogs), whereas non-match trials represented animals from different animal categories (e.g., a dog and a cow). Match trials at the SUBORDINATE level represented 2 pictures of the same individual from a different viewpoint whereas non-match trials presented pictures of different individuals. With these specifications, BASIC-match and SUBORDINATE-non-match trials comprised the same animal-pairs. The experiment included 768 trials and lasted for about forty minutes. The order of trials was randomized across participants.

## Results and Discussion

Remember that Experiment 1 sought to assess the interaction between the scale of objects and the level of categorization (basic vs. subordinate). Specifically, we tested that participants were equally good at assessing similarity judgments when they were required to do it at the basic level (e.g. same animal category?) and at the subordinate level (e.g., same individual?) when the scale of objects was large. A  $d'$  measure, which includes both Hit (H) rate (saying that two animals are different when they are different) and False Alarm (FA) rate (saying that two animals are different when they are identical), was used as our dependent variable. The top of Figure 3 shows the average  $d'$ s across all subjects at the different scales and categorization levels.

A two-way, within-subjects ANOVA revealed significant main effects of size (512, 256, 128, 64, 32, 16 square pixels)  $F(5, 95) = 38.59, p < .01$ , level of categorization (Basic vs. Subordinate),  $F(1, 19) = 66.29, p < .01$ , and a significant interaction between these factors,  $F(5, 95) = 5.80, p < .01$ . Further analysis revealed that differences between levels of categorization were true for 16 square pixels,  $F(1, 19) = 119.73, p < .01$ , 32 square pixels,  $F(1, 19) = 87.38, p < .01$ , 64 square pixels,  $F(1, 19) = 6.52, p < .01$ , 128 square pixels,  $F(1, 19) = 41.44, p < .01$  and 256 square pixels,  $F(1, 19) = 13.11, p < .01$ , but

not for 512 square pixels,  $F(1, 19) = 7.14$ , *ns*.

Furthermore, the slope of the best subordinate linear predictor in Figure 2, top, is about twice as much as that of the basic (see continuous lines in Figure 3, top;  $R^2 = .63$  and  $.81$  for the best basic and subordinate fits).

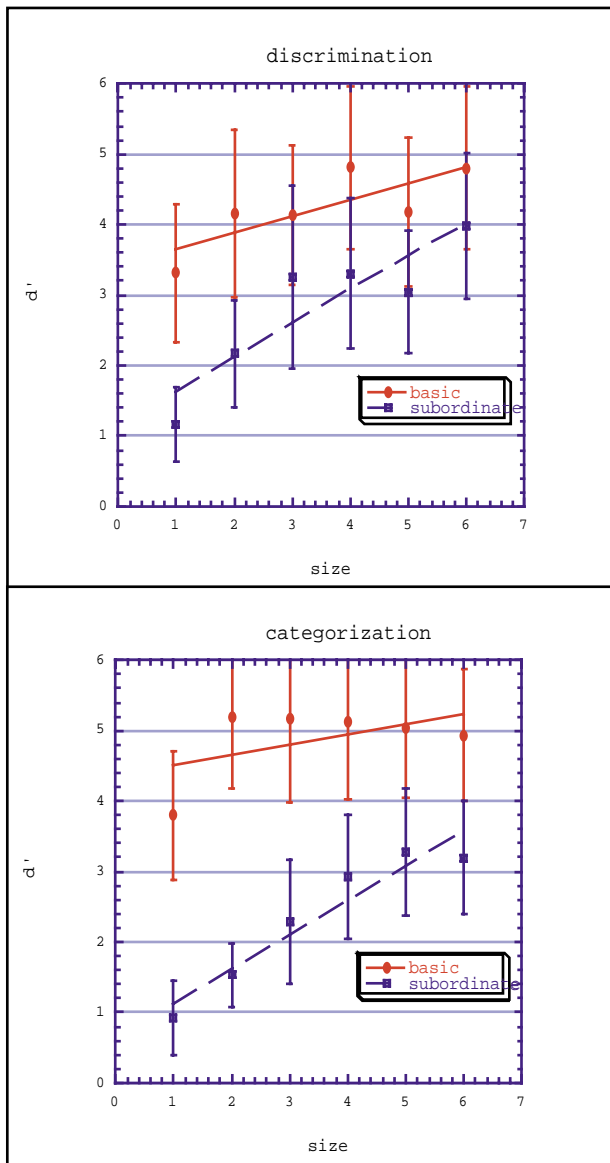


Figure 3. Average  $d'$ s with standard deviations for Experiments 1 and 2 at the basic and subordinate categorization levels. The continuous lines are the best linear predictors. (Note that size 1 = .38 deg and that size 6 = 12 deg.)

Results reveal that at smaller scales (256 to 16 square pixels) identical animals pairs were easier to distinguish at the basic level than at the subordinate level (see Figure 3, top).

## Experiment 2

Experiment 1 revealed that the information of SUBORDINATE-level judgments was less resilient to changes of scale than BASIC-level judgments. This is interesting because participants could use all the information available in the stimuli to resolve the task. It therefore suggests that the information for subordinate categorization vanished before that of basic

categorizations.

However, one could oppose that the matching task of Experiment 1 might solicit representations and processes that are atypical of everyday categorizations. For instance, it is conceivable that participants relied simply on local one feature-difference to decide that two stimuli differed—e.g., if one of the two stimuli had a tail. Of course, this would not explain how they did when the stimuli did not differ, but it still triggers the more basic problem of generalizing from the results of Experiment 1 to realistic categorizations.

Experiment 2 was designed to directly probe everyday categorization processes. One criticism that is often leveled at experiments studying the nature of visual information in categorization tasks is that they use tachoscopic conditions of stimulus presentation. Here, we made sure that the stimuli stayed on the screen for as long as the subjects felt necessary. This approach allows a measure of categorization performance in conditions of absolute information—i.e. information availability is not relative to speed of presentation.

Experiment 2 was a free categorization task. In a learning phase, participants learned to identify each of the sixteen animals at the basic and subordinate levels. They were then transferred to a categorization task where an animal (e.g., a whale) would appear on the screen (at one of 6 possible scales). After a self-paced scrutiny of the picture, participants were asked a question about the membership of the input to either a basic-level (is this a whale?), or a subordinate category (is this a sperm whale?). If the results of Experiment 1 tapped into the absolute levels categorization information then we expect the categorization results of Experiment 2 to follow a similar trend—i.e., a faster decrease of subordinate categorization accuracy with decrease in stimulus scale.

### Participants

Twenty Glasgow University students with normal or corrected vision were paid to participate in the experiment.

### Stimuli

The training set comprised gray-scale pictures of the 16 different individual animals. For each animal two pictures (showing the animals from two different view points—95 degrees apart in depth) were printed onto a white sheet of paper side by side. Pictures measured in total 10 x 10 cm. Each individual was identified by a sentence printed underneath the pictures. For example the two whales were identified as *sperm whale* and *humpback whale* (see Figure 2).

The stimuli used for the categorization task were the ones of Experiment 1 (e.g., see figures 1 and 2). The set consisted again of the 8 animal categories (2 individuals per animal category), the six different sizes and the two different viewpoints.

### Procedure

During the training phase, participants learned to identify each of the sixteen animals at the basic and subordinate levels. We tested their knowledge by presenting them with the pictures alone, one at the time, and asking them to name the animal at the basic and subordinate levels. *Perfect naming* performance was

required before going on to the categorization task. Corrective feedback was provided.

In the categorization task, participants were shown an animal on a computer monitor. Animals were presented from one of the two possible viewpoints and were displayed at one of the six different sizes. Participants were told that they could look at the animals for as long as they wanted. Once they were ready, a key press would initiate the disappearance of the animal and would display a question on the computer monitor. The question could either be basic ("Is it a cow?") or subordinate ("Is it a Friesian cow?"). Participants responded by pressing the appropriate key on the keyboard. The experiment included 768 randomized trials and lasted for about 50 minutes.

## Results and Discussion

Remember that Experiment 2 was designed to replicate results of Experiment 1 with a categorization task. We were thus interested mainly in the proportion of correct responses. For each subject, we computed  $d$ 's for all sizes and categorization levels. The bottom portion of Figure 3 shows the mean  $d$ 's across subjects for the different sizes and categorization levels.

A two-way, within-subjects ANOVA revealed significant main effects of size (512, 256, 128, 64, 32, 16 square pixels)  $F(5, 95) = 36.79, p < .01$ , level of categorization (Basic vs. Subordinate),  $F(1, 19) = 418.62, p < .01$ , and a significant interaction between these factors,  $F(5, 95) = 5.75, p < .01$ . Further analyses revealed that differences between levels of categorization were true for 16 square pixels,  $F(1, 19) = 174.53, p < .01$ , 32 square pixels,  $F(1, 19) = 379.61, p < .01$ , 64 square pixels,  $F(1, 19) = 119.23, p < .01$ , 128 square pixels,  $F(1, 19) = 68.95, p < .01$  and 256 square pixels,  $F(1, 19) = 39.41, p < .01$ , and for 512 square pixels,  $F(1, 19) = 47.91, p < .01$ .

The continuous lines on Figure 3, bottom, are the best linear predictors for the basic and subordinate  $d$ 's ( $R^2 = .26$  and  $R^2 = .92$ , respectively). The slope of the subordinate line is more than three times that of the basic one.

Results thus reveal that, animals were easier to categorize at the basic level than at the subordinate level (see Figure 3, bottom).

## General Discussion

This article tested the prediction that the preference for basic level categorizations could arise from a natural source of biases. When the retinal projection of one object shrinks in size (as happens when the object is moved further away from the observer), scale-specific visual information is lost. We tested the hypothesis that the basic and subordinate categorizations of identical objects require information that resides at different scales of the same distal stimulus. Two experiments tested these predictions.

In Experiment 1, participants in a similarity task were asked to judge whether two simultaneously presented objects had the same basic level, or the same subordinate category. We found that even though subjects could take *as long as they wanted* to inspect the object pairs, subordinate judgments were significantly more affected by a reduction in stimulus size than basic judgments. The

unconstrained inspection licenses the conclusion that we are tapping into the absolute level of information required for basic and subordinate categorizations.

Experiment 2 addressed the objection that a similarity task might trigger processes and representations that are atypical of everyday categorizations. In a categorization task, subjects had to confirm that the input belonged to a basic, or to a subordinate category. Even though subjects could again scrutinize the stimuli without any time constraint, we found that subordinate categorizations were much less resilient to changes of stimulus size.

In sum, the two experiments reported here converge on the idea that the perceptual shape cues required to resolve subordinate categorizations are more sensitive to scale changes than those required of basic categorizations. This has a number of implications for theories of basic and subordinate level categorization and recognition that we consider in turn.

Remember that we hypothesized a natural bias for the shape cues that access the basic level because these might be more resilient to variations in viewing distances. The results confirmed the hypothesis. Our results predict that the more robust default categorization strategy is to categorize objects at the basic, not the subordinate level. It is important to stress that we do not know precisely what the important basic and subordinate cues were in our experiments. However, to the extent that basic categorizations were not much affected by changes in size, we can propose that the cues present at all sizes (i.e. coarse scale cues) supported basic categorizations. For example, silhouettes were clearly present at all sizes and they could very well subtend basic categorizations—at least in the tasks considered here. An alternative could be that different cues residing at different scales can independently index the basic level, a hypothesis that has never been explored. If part extraction relies on the fine-grain edge description outlined in Biederman (1987), it seems unlikely that a part description of the objects subtended basic categorizations in our experiments, because Biederman's (1987) part description process is sensitive to scale.

## Concluding Remarks

We have shown here that size matters for subordinate categorization. One possibility for the basic level bias results from the greater resilience of basic level cues over a range of viewing distances. Future research will need to be conducted to isolate what the scale-independent and the scale-dependent cues are that support basic and subordinate categorizations.

## References

- Anglin, J. M. (1977). *Word, object and conceptual development*. New York: Norton.
- Berlin, B. (1992). *Ethnobiological classification: principles of categorisation of plants and animals in traditional societies*. Princeton, New Jersey: Princeton University Press.
- Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94, 115-145.
- Brown, R. (1958). How shall a thing be called? *Psychological Review*, 65, 14-21.
- Hamm, J. P. and McMullen, P. A. (1998). Effects of

- orientation on the identification of rotated objects depend on the level of identity. *Journal of Experimental Psychology: Human Perception and Performances*, 24, 413-426.
- Hoffmann, J. and Ziessler, C. (1983). Objectidentifikation in kunstlichen begriffshierarchien [Object identification in artificial concept hierarchies]. *Zeitschrift für psychologie*, 194, 135-167.
- Horton, M. S. and Markman, E. M. (1980). Developmental differences in the acquisition of basic and superordinate categories. *Child Development*, 51, 708-719.
- Johnson, K. E. and Mervis, C. B. (1997). Effects of varying levels of expertise on the basic level of categorisation. *Journal of Experimental Psychology: General*, 126, 248-277.
- Jolicoeur, P., Gluck, M., & Kosslyn, S. M. (1984). Pictures and names: Making the connexion. *Cognitive Psychology*, 19, 31-53.
- Markman, E. M. (1989). *Categorisation and naming in children: problems of induction*. Cambridge, Massachusetts: MIT Press.
- Markman, E. M. and Hutchinson (1984). Children's sensitivity to constraints on word meanings: Taxonomic vs. thematic relations. *Cognitive Psychology*, 16, 1-27.
- McMullen, P. A. and Jolicoeur, P. (1992). Reference frame and effects of orientation on finding the tops of rotated objects. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 807-820.
- Mervis, C. B. and Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development*, 53, 258-266.
- Murphy, G. L. (1991). Parts in objects concepts: Experiments with artificial categories. *Memory & Cognition*, 19, 423-438.
- Murphy, G. L. and Smith, E. E. (1982). Basic level superiority in picture categorisation. *Journal of Verbal Learning and Verbal Behaviour*, 21, 1-20.
- Murphy, G. L. and Brownell, H. H. (1985). Category differentiation in object recognition: Typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 11, 70-84.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382-452.
- Tanaka, J. W. and Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23, 457-482.
- Tversky, B. and Hemenway, K. (1984). Objects, parts and categories. *Journal of Experimental Psychology: General*, 113, 169-191.

# Subgoal Learning and the Effect of Conceptual vs. Computational Equations on Transfer

**Robert K. Atkinson** (atkinson@ra.msstate.edu)

Department Counselor Education and Educational Psychology; Box 9727;  
Mississippi State, MS 39762, USA

**Richard Catrambone** (rc7@prism.gatech.edu)

School of Psychology  
Georgia Institute of Technology  
Atlanta, GA 30332-0170, USA

## Abstract

Subgoal learning is examined through the use of equations that are designed to encourage a conceptual rather than computational approach to solving problems (conducting statistical tests). Learners who studied conceptually-oriented examples transferred more successfully to novel problems compared to learners who studied computationally-oriented examples. These results extend prior work on subgoal learning by demonstrating another technique for aiding subgoal learning.

## Introduction

Research suggests that learners typically struggle when they are obligated to solve problems that have different procedural requirements than those demonstrated by training problems or worked-out examples, even if those differences are relatively slight (e.g., Catrambone, 1995, 1996, 1998; Novick & Holyoak, 1991; Reed, Dempster, & Ettinger, 1985). This difficulty may stem in part from the fact that learners often represent the problem solving procedures of training problems or worked-out examples as a set of linear steps rather than forming a hierarchical representation that could permit them to successfully solve novel problems (Dufresne, Gerace, Hardiman, & Mestre, 1992; Singley & Anderson, 1989)

Educators and researchers alike are concerned with this problem. In fact, the Committee on Developments in the Science of Learning (1999) recently suggested that “a major goal of schooling is to prepare students for flexible adaptation to new problems and settings [and that] students’ abilities to transfer what they have learned to new situations provides an important index of adaptive, flexible, learning” (pp. 223). Research indicates, however, that this goal is rarely achieved (Chi, Feltovich, & Glaser, 1981; Larkin, McDermott, Simon, & Simon, 1980).

Presumably, emphasizing the structure of an example through instruction will increase flexible transfer by helping the learner look beyond the surface features of the example and test problem to find the goal-related features that can be used to solve the problem. Thus, instead of committing to memory the details of

equations as the basis for one’s problem solving knowledge, a more productive approach would be to organize this knowledge in such a way that it could support generalizations across problems in a domain. One type of knowledge structure that appears to offer the promise of enhancing this type of procedural generalization is one organized around subgoals.

## Subgoal-Oriented Instruction

As used in the present paper, a subgoal denotes a meaningful conceptual piece of an overall solution procedure. Subgoals are particularly useful to learners because they can assist them in solving novel problems since problems within a domain often share a common set of subgoals, albeit the steps for achieving the subgoals vary from problem to problem within a domain. Once learners become familiar with the typical subgoals in a domain, this knowledge can assist them in identifying which part of a previously-learned solution procedure needs to be modified in order to solve a novel problem (Catrambone, 1996, 1998).

Recently, a line of research has emerged examining the efficacy of subgoal-oriented instruction (Catrambone, 1995, 1996, 1998). In particular, this line of research has explored several techniques for designing examples that help learners to form subgoals to represent the purpose of steps in an example’s solution. Across a series of studies, Catrambone investigated the impact of making the goal structure of an example’s solution explicit by using manipulations such as the use of solution step labels or visually isolating parts of example solutions. These studies indicated that if examples are designed in such a way as to encourage subgoal learning, then learners are more likely to correctly solve new problems that involve the same subgoals but require new steps for achieving them.

These studies also suggest that example solutions that are segregated or labeled encourage learners to self-explain how the steps go together. One result of his self-explanation process is the formation of subgoals (Catrambone, 1998). This work parallels research in the text-comprehension literature on the effects of signals

or cues on text-processing strategies (e.g., Lorch & Lorch, 1995; Meyer & Rice, 1989). Just as organizational signals in text induce learners to change their text-processing strategy by cueing the important text content and its organizational structure, worked-example labels are intended to increase the likelihood that learners will discern the hierarchical conceptual structure of the problem contained in the example.

### Factors that May Influence Subgoal Formation

As previously mentioned, several structural manipulations have been found to successfully make the goal structure of a problem's solution explicit, such as by the use of labels or visual isolation. But, there might be other factors that influence subgoal information. For instance, two potential factors are the nature of equations used in examples and the presence of conceptual elaborations.

**Conceptual vs. Computational Equations.** The process of calculating sum of squared deviation scores or sums of squares (SS) for the variance terms in t-tests and analyses of variance (ANOVAs) can involve two noticeably distinct types of formulas: conceptual and computational. According to Gravetter and Wallanu (2000), the conceptual formula is useful "because the terms in the formula literally define the process of adding up the squared deviations" (p. 121). For instance, the conceptual formula for SS in a t-test,  $\sum (X - \bar{X})^2$ , translates directly into the sum of ( $\sum$ ) squared deviations  $(X - \bar{X})^2$ . This clearly captures how the variance term measures the amount of spread about the mean.

In contrast, the computational formula for SS,  $\sum X^2 - \frac{(\sum X)^2}{N}$ , permits the learner to calculate SS directly from raw scores which can lead to more efficient calculations. However, there is a notable drawback to this convenience: the computational formula conceals the true meaning behind SS. Unlike the conceptual formula, a learner cannot directly translate the terms in the computational formula into a sum of squared deviation scores. As a result, the learner may not grasp that this formula is designed to measure the amount of spread about the mean.

On the one hand, the computational approach might aid performance on problems that are just like the examples that illustrated the approach, but might make far transfer difficult. That is, in the computational approach, the equation is streamlined for doing the calculations, but "hides" what is really going on. On the other hand, the conceptual approach, although typically more cumbersome computationally, clearly shows how the variance is related to the difference of each mean from

the grand mean. Therefore the conceptual approach might aid far transfer by making it easier for the learner to determine how to adapt relevant parts of the procedure.

Thus, we hypothesize that conceptually-oriented equations will be more effective than computationally-oriented equations at helping learners acquire knowledge structured around the goal-related features of the problems they study and this translates to superior far transfer performance.

**Conceptual Elaborations.** Another factor that appears to have the potential to influence subgoal formation is the use of elaborations in example-based instruction and, in particular, conceptual elaborations. The literature contains examples of several types of elaborations that vary in the degree to which they elaborate the problem at hand. They range from elaborations involving problem solutions (Lovett, 1992) to those that focus on rules and procedures (Catrambone, 1996; Reed & Bolstad, 1991; Reed et al., 1985).

To date, the success of these various elaborations has been mixed. Although Lovett (1992) found that far transfer was facilitated by elaborated solutions, Reed and his colleagues (Reed & Bolstad, 1991; Reed et al., 1985) have found virtually no evidence to suggest that rule-based instructional elaborations—those that elaborate on the purpose and appropriateness of applying a rule or procedure in a given problem-solving context—are beneficial to learners.

In one study, Catrambone (1996) examined the relative benefits of rule-based instructional elaborations versus subgoal labels. In this study, Catrambone manipulated two factors: subgoal labels (present or absent) and rule-based elaborations (present or absent), where the elaborations consisted of supplemental material describing an alternate representation or equation that could be used to solve the problems the participants were studying. He found that the labeling manipulation enhanced transfer while the rule-based elaboration manipulation did not.

The rule-based elaborations used in the Catrambone (1996) study, however, offered "what to do" knowledge not "what it means" knowledge. This distinction is important in light of research suggesting that rules conveying "what to do" knowledge might provide little help to learners for developing a deep understanding of the rule-based system they are studying whereas knowledge about "what it means" may facilitate this depth of understanding (Riesbeck & Schank, 1989). For instance, an elaboration that describes what is meant by the term "variance" (see Appendix for an example) might be more effective than one dedicated to elaborating the procedural aspect of the variance formula.

In sum, the impact of conceptual elaborations containing "what it means" knowledge in the context of subgoal-oriented instruction remains an open question.

## Overview of Study

The aim of the study was to compare the effectiveness of conceptual and computational equations, and the use of elaboration, on performance. Performance was assessed in two ways: the time spent studying the training examples and correctness of solutions on near and far transfer problems.

## Experiment

### Method

**Participants and Design.** Participants were 215 students drawn from several educational psychology courses at a small, northeastern college who participated in the experiment for course credit. The participants were randomly assigned to one cell of a 2 x 2 x 2 factorial design. The first factor was the characteristics of the variance formulas (conceptual or computational) in the t-test example, the second was the characteristics of the variance formulas (conceptual or computational) in the ANOVA example, and the third was conceptual elaboration (elaboration or no elaboration) in the examples, described below.

**Training Phase.** Participants received an instructional booklet containing a general overview of statistical hypothesis tests and two training examples, one representing a t-test and another representing the use of an ANOVA for the same 2-group comparison. The introduction to statistical hypothesis tests described the utility of these procedures and provided an overview of the four-step hypothesis testing process common to both tests. Each training example was preceded by an overview of the test that it exemplified. This explanation described the purpose of the test without going into detail regarding how to perform the test's calculations.

Half of the participants were exposed to examples that contained conceptual elaborations designed to provide "what it means" knowledge. That is, they were designed to describe the conceptual meaning behind the various formulas used in the two hypothesis tests. The other half of the participants studied examples in which the elaborations were not present.

With respect to the t-test example, the variance formulas were either conceptual or computational in nature. Similarly, with regard to the ANOVA example, the variance formulas were either conceptual or computational in character.

Regardless of the instructional manipulations, the examples contained a number of invariant structural features. First, all of the equations used across both tests were converted to their verbal equivalents so that they were devoid of any statistical notations. Second, each of the six calculational subgoals in the two examples was either labeled or visually isolated.

The t-test subgoals were to find: sample mean for group 1, variance for group 1, sample mean for group 2, variance for group 2, pooled variance, and t-statistic. The ANOVA subgoals were to find/do: preliminary calculations, sum of squares between, sum of square within, mean squares between, mean squares within, and f-value.

The Appendix shows samples of the materials from the examples.

**Test Phase.** The test booklet contained three test problems for the participants to solve. The first test problem required the participant to apply a t-test. The second problem required them to apply an ANOVA to a 2-group situation and the third problem asked them to apply an ANOVA to novel situation involving three groups. Thus, the first two problems were near transfer while the third problem involved more far transfer in that it required the learner to adapt the equations for variance. The extension is a more straightforward, modular extension of the conceptual equations. However, the extension is less straightforward in the computational equations since it involves changes to the "interior" of the equations.

The test booklet also included two sheets that participants could refer to, one containing the condition-specific formulas for the t-test (conceptual or computational) and the other containing the condition-specific formulas for the ANOVA. Although these sheets represented the formulas in the sequence in which they were applied in the training examples they did not contain any of the values from those examples.

A binary scoring system was developed to score the problem-solving protocols. This system was designed to award participants with points for the accuracy with which they achieved each subgoal. The three test problems each contained six calculational subgoals. The correct numerical answer to the subgoal was awarded one point. For example, the correct answer to the second subgoal in the t-test problem, correct group 1 variance, was 30.2. If a participants' problem-solving protocol contained this answer, he/she was given a point.

Since most subgoals contained subcomponents, the binary system allowed us to award partial credit. This permitted us to capture the proportion of the subgoal's solution—for those participants who did not have the correct numerical answer for the subgoal—that was correct. For instance, the equation associated with the second subgoal (i.e., correct group 1 variance) in the conceptual condition was coded for the presence or absence of seven components, ranging from whether each value was present in the formula to whether the equation had the correct denominator. In this example, if a participant's problem-solving protocol had six of the seven components, he/she was awarded a .86 for the



subgoal. If the subgoal was correct except for a trivial math error, the participant received full credit (one point) for that particular subgoal.

**Procedure.** Participants were asked to study carefully the instructional booklet containing the training examples since after studying it they would be asked to solve several problems. They recorded the amount of time they spent studying each example. The participants were informed that they would not be able to refer to any of the examples while solving the problems but that they would have a copy of the formulas. This constraint was designed to increase the likelihood that participants would focus their attention on studying the examples and how they were solved.

Participants were run in groups ranging in size from 5 to 30 participants. Participants worked for approximately 75 minutes and were asked to show all their work.

## Results

To validate the scoring system that was developed, two raters independently scored a random sample of 10% of the problem-solving protocols and agreed on scoring 98% of the time. Disagreements were resolved by discussion. One experimenter independently scored the remaining problem-solving protocols.

A 2 x 2 x 2 analysis of variance was initially conducted on the study times for the two examples (i.e., t-test and 2-group ANOVA) and the correctness measures for the three test problems, using elaboration, type of t-test formulas, and type of ANOVA formulas as grouping factors. There was no systematic effect of elaboration on correctness and so, in the interest of clarity and brevity, this factor will not be discussed below in the context of correctness. Table 1 presents the mean scores for each condition on the correctness measures for the three test problems.

*Training Times for T-Test Example:* There was a significant main effect of elaboration,  $F(1, 207) = 10.11, MSE = 10.8, p < .01$ , which indicated that the participants presented with the elaborated material ( $M = 8.11$  min.) spent more time studying the examples compared to participants who studied unelaborated materials ( $M = 6.73$  min.). There were no other significant main effects or interactions.

*Training Times for ANOVA Example:* There were no significant main effects or interactions for training times on the ANOVA example.

*Performance on T-Test Problem (Near Transfer):* There was a significant main effect of t-test formula,  $F(1, 211) = 9.18, MSE = 1.32, p = .009$ , which indicated that the participants exposed to the conceptual t-test example outperformed those who studied the computational version. There was no effect on performance as a function of the version of the ANOVA example studied and there was no interaction between the factors.

*Performance on 2-Group ANOVA Problem (Near Transfer):* There were no significant main effects for this dependent measure; t-test:  $F(1, 211) = 1.06, MSE = 2.09, p = .31$ ; ANOVA:  $F(1, 211) = 0.21, p = .65$ . However, the two-way interaction between t-test equations and ANOVA equations was significant,  $F(1, 211) = 5.52, p < .02$ . Examination of the mean scores suggest a disordinal interaction, that is, the effects of the t-test factor reverse themselves as the levels of the ANOVA factor change. Specifically, for the participants provided with conceptual t-test formulas, the conceptual ANOVA group obtained a higher score than the computational group. For participants provided with the computational t-test formulas, the computational ANOVA group obtained a higher score than the conceptual group.

*Performance on 3-Group ANOVA Problem (Far Transfer):* There were no significant main effects for this dependent measure; t-test:  $F(1, 211) = 2.55, MSE = 2.65, p = .11$ ; ANOVA:  $F(1, 211) = 0.10, p = .75$ . The interaction was significant,  $F(1, 211) = 6.01, p < .02$ . Examination of the mean scores revealed the same disordinal interaction found in the 2-group problem. That is, for the participants provided with conceptual t-test formulas, the conceptual ANOVA group obtained a higher score than the computational t-test formulas. For participants provided with the computational t-test formulas, the difference was reversed.

## Discussion

The overall performance differences among the groups can be summarized as follows: the combined t-test

Table 1: Scores on Test Problems as a Function of T-Test and ANOVA Examples

	T-Test Conceptual		T-Test Computational	
	ANOVA Conceptual	ANOVA Computational	ANOVA Conceptual	ANOVA Computational
T-Test Problem (max = 6)	5.43	5.46	4.99	5.07
2-Group ANOVA Problem (max = 6)	4.16	3.61	3.50	3.87
3-Group ANOVA Problem (max = 6)	4.02	3.55	3.12	3.74

conceptual and ANOVA conceptual condition tended to outperform the other conditions consisting of the other possible combinations of t-test formulas and ANOVA formulas on near and far transfer problems. There was little evidence of improved generalization by any group as a function of having been provided with elaborations. The results suggest that the first example sets the tone for the interpretation of the second example and performance on the far transfer problem. If the first t-test example used conceptual equations, then performance on the far transfer (3-group) ANOVA problem was particularly aided if the ANOVA example was also conceptual. If the first t-test example was computationally-oriented, the performance on the far transfer ANOVA problem was better if the ANOVA example was also computationally-oriented. Thus, it appears that in order for a learner to acquire a more subgoal-oriented approach to these problems, the best pedagogical approach would be to make both examples use conceptual equations. Even if the ANOVA example was conceptually-oriented, its benefits on the far transfer ANOVA problem were reduced if the initial t-test example was not also conceptual. Consistency in the examples appears to be important for subgoal learning.

The results advance prior work on subgoal learning by demonstrating that generalization can be enhanced through the nature of the equations used in examples. Thoughtfully-designed examples that include conceptually-oriented equations seem to be an effective way to help learners solve novel problems.

Two caveats remain, however. First, under certain circumstances, the conceptual formula represents the most direct way of calculating sum of squares. In particular, when a data set consists of a small number of whole numbers and its mean is a whole number (which characterizes the data used in the present study), the resulting deviation score will be a whole number, which allows the learner to avoid the computational burden of decimals or fractions. Thus, one could argue that the advantage of the conceptual group in the present study therefore appears to be computational, rather than in increasing understanding. This suggests that a follow up study should explore the impact of presenting computational and conceptual equations to learners in situations in which the latter is clearly more cumbersome computationally (e.g., resulting means are not whole numbers and/or data set contains decimals).

Second, while the present results are consistent with the claims about benefits to transfer for learners who acquire useful subgoals (e.g., Catrambone, 1996, 1998), subgoal-learning was demonstrated only indirectly here. Thus, an important extension of the present work is to add converging measures, such as talk-aloud protocols, to determine if the transfer advantage can be clearly tied back to subgoal learning.

## References

- Bransford, J.D., Brown, A.L., & Cocking, R.R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.
- Catrambone, R. (1995). Aiding subgoal learning: Effects on transfer. *Journal of Educational Psychology*, 87(1), 5-17.
- Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(4), 1020-1031.
- Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General*, 12(4), 355-376.
- Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152.
- Dufresne, R.J., Gerace, W.J., Hardiman, P.T., & Mestre, J.P. (1992). Constraining novices to perform expertlike problem analyses: Effects on schema acquisition. *The Journal of the Learning Sciences*, 2, 307-331.
- Gravetter, F.J., & Wallnau, L.B. (2000). *Statistics for the behavioral sciences*. Belmont, CA: Wadsworth.
- Larkin, J., McDermott, J., Simon, D.P., & Simon, H.A. (1980). Expert and novice performance in solving physics problems. *Science*, 208, 1335-1342.
- Lorch, R., & Lorch, E.P. (1995). Effects of organizational signals on text-processing strategies. *Journal of Educational Psychology*, 87, 537-544.
- Lovett, M.S. (1992). Learning by problem solving versus by examples: The benefits of generating and receiving information. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 956-961). Hillsdale, NJ: Erlbaum.
- Meyer, B.J.F., & Rice, E. (1989). Prose processing in adulthood: The text, the reader and the task. In L. W. Poon, D. C. Rubin, & B. A. Wilson (Eds.), *Everyday cognition in adult and later life* (pp. 157-194). New York: Cambridge University Press.
- Novick, L.R., & Holyoak, K.J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(3), 398-415.
- Reed, S.K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(1), 106-125.
- Reed, S.K., & Bolstad, C.A. (1991). Use of examples and procedures in problem solving. *Journal of Educational Psychology*, 17(4), 753-766.
- Riesbeck C.K. & Schank, R.C. (1989). *Inside case-based reasoning*. Hillsdale, NJ: Erlbaum.
- Singley, M.K., & Anderson, J.R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard Univ. Press.

## Appendix

### Sample Materials from T-Test Example and ANOVA Example

#### *SAMPLE OF PROBLEM STATEMENT:*

A car manufacturer that makes a car called the Jupiter just came out with a new model, the Jupiter XL. Some of the modifications made to the car are expected to improve the mpg (miles per gallon) rating of the car while other modifications are not. The manufacturer has hired your firm, an independent consumer research firm, to test the new model. To determine if there is any difference between the mpg rating of the old and new models, you collect a random sample of 5 cars of the old model and 6 cars of the new model. You drive the cars along the same city route and record the average mpg rating of each car. Here are the data:

Old Model		New Model	
<u>Car</u>	<u>MPG</u>	<u>Car</u>	<u>MPG</u>
1	30	1	37
2	34	2	36
3	34	3	40
4	29	4	36
5	33	5	34
		6	33

#### *SAMPLE OF CONCEPTUAL ELABORATION FOR THE COMPUTATIONAL T-TEST VARIANCE CALCULATION*

A variance is a measure of how much the scores that make up a group deviate from the mean of a group. Even though it is not obvious in the calculation below, part of the calculation of the variance involves computing the difference between each score in a group and the mean for the group. Thus, a variance measures the variability of scores around a mean.

#### *SAMPLE OF COMPUTATIONAL T-TEST VARIANCE CALCULATION*

$$s_1^2 = \text{sample variance for group 1} = \frac{\text{sum of squared scores in group 1} - \frac{(\text{sum of the scores in group 1})^2}{\text{number of scores in group 1}}}{(\text{number of scores in group 1}) - 1}$$

$$= \frac{(30^2 + 34^2 + 34^2 + 29^2 + 33^2) - \frac{(30 + 34 + 34 + 29 + 33)^2}{5}}{5 - 1} = \frac{5,142 - \frac{(160)^2}{5}}{4} = \frac{22}{4} = 5.5$$

#### *SAMPLE OF CONCEPTUAL T-TEST VARIANCE CALCULATION*

$$s_1^2 = \text{sample variance for group 1} = \frac{(1\text{st score} - \text{mean})^2 + (2\text{nd score} - \text{mean})^2 + \dots + (\text{last score} - \text{mean})^2}{(\text{number of scores in group 1}) - 1}$$

$$= \frac{(30 - 32)^2 + (34 - 32)^2 + (34 - 32)^2 + (29 - 32)^2 + (33 - 32)^2}{5 - 1} = \frac{22}{4} = 5.5$$

#### *SAMPLE OF COMPUTATIONAL ANOVA SUM OF SQUARES (BETWEEN) CALCULATION*

*SSB* = sum of squares between groups

$$= \left[ \frac{(\text{sum of scores in group 1})^2}{\text{number of scores in group 1}} + \frac{(\text{sum of scores in group 2})^2}{\text{number of scores in group 2}} + \dots + \frac{(\text{sum of scores in last group})^2}{\text{number of scores in last group}} \right] - \frac{(\text{sum of scores in all groups})^2}{\text{number of scores in all groups}}$$

$$= \left[ \frac{(160)^2}{5} + \frac{(216)^2}{6} \right] - \frac{(376)^2}{11} = 12,896 - 12,852.36 = 43.64$$

#### *SAMPLE OF CONCEPTUAL ANOVA SUM OF SQUARES (BETWEEN) CALCULATION*

*SSB* = sum of squares between groups

$$= \text{number of scores in group 1} (\text{the mean for group 1} - \text{grand mean})^2 +$$

$$\text{number of scores in group 2} (\text{the mean for group 2} - \text{grand mean})^2 + \dots +$$

$$\text{number of scores in last group} (\text{the mean for last group} - \text{grand mean})^2 = 5(32 - 34.18)^2 + 6(36 - 34.18)^2 = 43.64$$

# A Comparative Study of Unsupervised Grapheme-Phoneme Alignment Methods

Timothy Baldwin and Hozumi Tanaka  
Dept of Computer Science  
Tokyo Institute of Technology  
2-12-1 Ookayama,  
Meguro-ku, Tokyo 152-8552 JAPAN  
{tim,tanaka}@c1.cs.titech.ac.jp

## Abstract

This paper describes and compares two unsupervised algorithms to automatically align Japanese grapheme and phoneme strings, identifying segment-level correspondences between them. The first algorithm is inspired by the TF-IDF model, including enhancements to handle phonological variation and determine frequency through analysis of “alignment potential”. The second algorithm relies on the C4.5 classification system, and makes multiple passes over the alignment data until consistency of output is achieved. In evaluation, the first algorithm proves to be greatly superior to the second, producing a word accuracy of 96.94%.

## Introduction

The task of grapheme-phoneme alignment is intrinsically related to text-to-speech conversion, and provides the basic toolset of grapheme-phoneme correspondences for use in predicting the pronunciation of a given word. While it is certainly possible to handcraft grapheme-to-phoneme mappings (see, e.g., (Allen et al., 1987; Sejnowski and Rosenberg, 1987; Huang et al., 1994; Divay and Vitale, 1997)), we suggest that it should be possible to automatically extract such data from a database of grapheme-phoneme string pairs *without any form of supervision*. Thus, given a pronunciation-annotated machine-readable dictionary, it should be possible to generate a set of aligned grapheme-phoneme (word-pronunciation) pairs reliably and fully automatically. Theoretically, the grapheme-phoneme alignment output could then be plugged into a reading machine, producing an instant text-to-speech system for any language (as per (Ling and Zhang, 1998; Black et al., 1998)).

The objective of this paper is to analyse the applicability of unsupervised learning methods to automated grapheme-phoneme alignment in Japanese. In particular, we propose an incremental learning algorithm founded upon the TF-IDF metric, and compare this to a multi-pass alignment method drawing on the C4.5 classification system (inspired by the method of (Ling and Wang, 1997)). Alignment data is first constructed by exhaustively generating all alignment mappings for a given grapheme-phoneme pair. We filter off lexically and phonologically implausible alignment candidates from this data, and feed the final set of alignment candidates into the different alignment algorithms. These algorithms then incrementally disambiguate the data to produce a unique alignment candidate for each grapheme-phoneme tuple, through analysis of frequency distribution in the data.

## Definitions

Japanese is made up of the three native orthographies of kanji, katakana and hiragana. **Kanji** characters (e.g.

“消”) derive from the Chinese writing system and are largely ideographic in nature; a single kanji character tends to have multiple pronunciations (a sample of readings for “消” include *syō*, *ki(eru)* and *ke(su)*). **Katakana** and **hiragana** (collectively described as **kana**) are isomorphic syllabaries, with each character describing a unique, mutually exclusive phoneme content; examples of hiragana and katakana are “し” (*si*) and “ゴ” (*go*), respectively. The three orthographies intermingle in modern-day Japanese texts, with hiragana generally used for inflectional affixes, case particles and stop words, katakana for loan words, and kanji for content word stems. This effect is seen in the word 消しゴム [*kesigomu*] “eraser”, which incorporates all three script types.

In targeting “graphemic Japanese”, therefore, we must consider all three writing systems. Phonemic Japanese, on the other hand, can be described through kana characters, as all kanji characters are transcribable into kana, and kana describe the full phonemic inventory of Japanese in the form of phoneme chunks. That is not to say that every kana character maps to a single phoneme, but there is a unique broad phonetic transcription associated with almost all kana characters.<sup>1</sup> It is thus trivial to complete the full grapheme-phoneme conversion process if necessary, and at the same time, our choice of kana characters as phoneme medium frees us from consideration of low-level connective restrictions between phoneme units, as this information is implicitly encoded within the orthography.

Grapheme-phoneme (“G-P”) alignment is defined as the task of *maximally* segmenting a grapheme compound (a single dictionary entry, usually constituting a single word) into morpho-phonetic units, and aligning each such unit to its corresponding phoneme unit in the phonetic transcription for that compound. Segmentation of the grapheme compound is maximal in the sense that no segment can be further segmented into aligning sub-segments. To take the example of the grapheme string 感謝-*su-ru* [*ka-n-sya-su-ru*] “to thank/be thankful”,<sup>2</sup> 感 aligns with *ka-n* in the phoneme string, and 謝 with *syā*, as indicated in *align*<sub>1</sub> of Fig. 1.

<sup>1</sup>The only exception to affect us is the kana *u*, which when not used as an inflecting suffix, is pronounced as /o/ when immediately preceding an /o/ sound within a phoneme segment, and /u/ otherwise. Here, disambiguation is possible given phoneme segment context and part-of-speech information.

<sup>2</sup>So as to make this paper as accessible as possible to readers not familiar with Japanese, kana characters are written italicised in Latin script for the remainder of this paper, with character boundaries indicated by “.” and segment boundaries (which double as character boundaries) indicated by “⊙”.

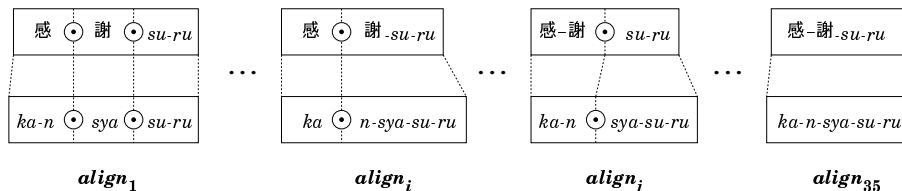


Figure 1: Candidate alignments for 感謝-su-ru [ka-n-sya-su-ru] “to thank/be thankful”

## Cognitive aspects of G-P alignment

One vital issue in grapheme-phoneme alignment is the determination of ‘atomic’ grapheme segments, that is segments which are not further divisible. Clearly, the lower bound on atom size for Japanese is a single kana or kanji character, but there is no inherent upper bound on the number of characters that can combine to form a segment, for either grapheme or phoneme segments. While it is correct to say that there is a cognitive preference to segment off individual kanji characters (possibly with kana suffices), there is equally potential for (indivisible) multiple-kanji grapheme segments, such as 白-詞 [se-ri-fu] “one’s lines”. Consequently, alignment does not simply consist of segmenting the grapheme string up into individual characters and aligning them with chunks of the phoneme string, and consideration must be given to the granularity of segmentation.

A number of inter-related cognitive factors seem to determine the “segmentability” of a grapheme string and resultant “alignability” with a given phoneme string, namely: (i) the relative frequency of each segment-level G-P sub-alignment; (ii) the cognitive immediacy of adjacent segments; and (iii) phonetic similarity to regular readings in the case of novel G-P sub-alignment.

High relative frequency of alignment refers to the situation of a given grapheme segment  $g$  commonly aligning with a given phoneme segment  $p$  (and phonological variants thereof), such as 感 invariably aligning with the reading  $ka-n$ . Clearly if the  $\langle \dots \circ g \circ \dots \rangle - \langle \dots \circ p \circ \dots \rangle$  alignment sub-schema is observed with sufficient frequency, a natural preference will arise to emulate that same alignment sub-schema wherever possible, for reasons of familiarity.

In the case that there is no alignment schema which produces familiar alignments for all individual grapheme segments, there is a tendency to preserve as much regularity to the overall alignment schema as possible by maximising the number of regular alignments and framing any irregular alignments between segment-level alignments of high cognitive immediacy. Thus, when presented with a G-P tuple such as  $\langle \text{白髮} \rangle - \langle si-ra-ga \rangle$ , where 白 is commonly associated with the reading  $si-ra$  but not  $si$  and there are no independent instances of 髮 taking a  $ga$  or  $ra-ga$  reading, there is a natural preference to uphold the single known sub-alignment for 白 and produce a forced alignment for 髮, as in  $\langle \text{白} \circ \text{髮} \rangle - \langle si-ra \circ ga \rangle$ .

Finally, if a novel alignment must be made such as  $\langle \dots \circ \text{髮} \rangle - \langle \dots \circ ga \rangle$  above, conservatism rules in that irregular readings tend to be chosen so as to be phonetically similar to established readings. In the case of 髮, the established reading is  $ka-mi$  (or  $ga-mi$  in its voiced realisation), from which the deletion of a single character produces the suggested  $ga$  reading.

In the case that the above processes do not apply to any substring of the G-P tuple, the tendency is to chunk unalignable kanji together into a single multi-kanji segment, such as occurred for  $se-ri-fu$  above.

The implications of the above observations to our statistical modelling of G-P alignment are to develop a model which gives preference to sub-alignments of high plausibility, allows irregular alignments given that the surrounding context displays high cognitive immediacy of alignment, and has the facility to “back-off” to multi-kanji segments when necessary. Our interpretation of TF-IDF is suggested to constitute such a model.

## Grapheme-phoneme alignment

Grapheme-phoneme alignment is performed as a three-stage process: (a) detection of lexical alternation and removal of lexical alternates from the input; (b) determination of all possible alignment candidates and subsequent pruning through alignment constraints; and (c) scoring of all final candidate alignments to determine the final solution.

**Lexical alternates** are defined as containing the same kanji characters in the same linear order, and coinciding in phonemic content (i.e. having the same reading). We enforce the constraint that all lexical alternates must be governed by the same basic alternation schema, allowing us to filter off alignment candidates for a given G-P tuple which are incompatible with one or more alternates of that tuple.

Given that both grapheme and phoneme segments can be of arbitrary length, **alignment candidate generation** must encompass all segmentation cardinalities. That is, for the example of a three-character grapheme string, we must consider the maximal segmentation of the string into three segments, and also partial segmentations into two segments or alternatively a single segment encompassing the full string.

Luckily, we are able to rely on strict linearity of alignment between the grapheme and phoneme strings, and in most cases can count on the alignment being isomorphic (the only exception being “grapheme gapping” – see (Baldwin and Tanaka, 1999b)). As a result, the total number of alignments is given by  $\sum_{x=0}^{l-1} C_x^{m-1} C_x^{n-1}$  in the general case, where  $m$  is the character length of the grapheme string,  $n$  the character length of the phoneme string, and  $l = \min(m, n)$ .

We are able to reduce the alignment space considerably, however, through the advent of five lexical and phonological constraints on alignment, as described in (Baldwin and Tanaka, 1999b). These constraints apply to script and syllable boundaries, the character length of aligned segments, and the number of voiced obstruents contained in a single phoneme segment. For the data targeted in evaluation, the average alignment paradigm size was reduced from 12.06 to 3.27 (a reduction of close to 75%), with no instances of the correct alignment candidate being pruned from the alignment paradigm.

It is important to realise that the application of the above constraints not only reduces the search space for statistical scoring, but can actually single out a unique

$$freq(\langle g, p \rangle) = \left| \left\{ \langle GS, PS \rangle : \exists p' \in phon\_var(p) \left\{ \langle \dots \underset{i}{\circ} g \underset{i+1}{\circ} \dots \rangle - \langle \dots \underset{i}{\circ} p' \underset{i+1}{\circ} \dots \rangle \in \{ \langle GS_{seg} \rangle - \langle PS_{seg} \rangle \} \right\} \right\} \right| \quad (1)$$

$$wfreq(\langle g, p \rangle) = SOLVED \cdot freq_{SOLVED}(\langle g, p \rangle) + UNSOLVED \cdot freq_{UNSOLVED}(\langle g, p \rangle) \quad (2)$$

$$tf-idf(\langle g, p, ctxt \rangle) = \underbrace{\frac{wfreq(\langle g, p \rangle) - UNSOLVED + \alpha}{wfreq(\langle g \rangle)}}_{tf(\langle g, p \rangle)} \log \left( \underbrace{\frac{wfreq(\langle g, p \rangle)}{wfreq(\langle g, p, ctxt \rangle) - UNSOLVED + \alpha}}_{idf(\langle g, p, ctxt \rangle)} \right) \quad (3)$$

legal solution, providing what turns out to be vital “free ride” alignment data to bootstrap the different systems with.

The alignment constraints are the only component of the overall formulation which is specific to Japanese, and the different algorithms would be equally applicable to unconstrained alignment data, making them directly transferrable to any other language.

We next turn to description of the two main unsupervised alignment methods.

## Incremental learning with TF-IDF

The first algorithm (originally proposed in (Baldwin and Tanaka, 1999a)) is based on incremental learning or “hill-climbing”, whereby the system disambiguates a single alignment paradigm at a time and incrementally updates the statistical model according to both discarded alignment candidates and the selected alignment solution. Selection of the alignment paradigm to be disambiguated is performed according to an adaptation of the TF-IDF scoring metric (Salton and Buckley, 1990), originally developed within the information retrieval fraternity for term weighting. In this, we score and rank each alignment candidate contained within the current alignment paradigm, and further rank the different alignment paradigms according to the weighted ratio between the top- and second-ranking alignment candidates. The highest-scoring alignment paradigm on each iteration is selected for disambiguation, according to the top-ranking alignment candidate described therein. We then update the statistical model, revise scores for alignment paradigms affected by the changed statistics, and rerank in preparation for the next iteration.

The utility of TF-IDF within the task of G-P alignment, stems from it weighting up terms (aligned G-P segments) which occur frequently within a given document (grapheme segment) context, but relatively infrequently within other documents (left/right adjoining grapheme and phoneme contexts). As described above, we wish to model the cognitive process of G-P alignment by maximally weighting high-frequency (regular) readings for a given grapheme string, but at the same time scoring down readings which occur primarily in a fixed lexical context, as this would tend to point to oversegmentation at the phoneme level (the phoneme context is in actual fact part of the reading for the current grapheme segment) and/or the grapheme level (the grapheme context clusters with the current grapheme segment to form a multiple-grapheme segment).

In addition to facilitating the detection of regular alignments, TF-IDF provides a means of variably “windowing” over the grapheme and phoneme strings, in that it does not involve a pre-conceived notion of segment size. Additionally, by way of taking the mean of the scores for the left/right and grapheme/phoneme contexts for each aligned G-P segment pair, we are able to weight up alignments with more highly regularised segment-level

readings, again mirroring the cognitive processing of G-P alignment.

While TF-IDF offers no immediate solution to the third cognitive issue of conservatism in cases of non-regular readings, it does allow us to handle abbreviations of regular readings—as was seen above for the *ga* reading of 髪—in that they will generally be found within the (undisambiguated) alignment paradigm of G-P tuples drawing on the regular reading.

## Counting frequencies

Clearly, to be able to apply the TF-IDF metric, we require some way of counting frequencies. For disambiguated alignment paradigms, we can rely on the absolute frequencies of segments contained within the alignment solution. For residue alignment paradigms awaiting disambiguation, on the other hand, we have an arbitrary number of alignment candidates to choose from, and no immediate way of producing an all-encompassing frequency value.

We resolve this issue by associating a single frequency count with every segment type occurring independently in a given alignment paradigm, irrespective of the number of alignment candidates it occurs within. In this way, we model the “alignment potential” of each segment. This process can be formalised as in equation (1), in the case of  $freq(\langle g, p \rangle)$  (singleton and triple segment combinations are defined in a similar fashion). Here,  $p$  is the phoneme segment aligning with grapheme segment  $g$ , and  $phon\_var(p)$  describes the set of “phonological alternates” of  $p$ . Phonological alternates are predictable instances of phonological alternation from a base form  $p$ , with the most widespread types of phonological alternation being “sequential voicing” (Tsujimura, 1996) and gemination. Fortunately, phonological alternation occurs only on syllables at phoneme segment boundaries, and phonological alternate “equivalence classes” are mutually exclusive in the main. This allows us to cluster together frequencies for members of each equivalence class, going some way toward combating the effects of data sparseness.

## The basic TF-IDF model

Our interpretation of the TF-IDF model is given in equation (3), where  $g$  is a grapheme segment,  $p$  a phoneme segment and  $ctxt$  a single phoneme or grapheme context for  $\langle g, p \rangle$  within the current alignment. As an additional facet of hill-climbing, we weight up segment frequencies contained within disambiguated alignment paradigms ( $freq_{SOLVED}$ ), over those for unprocessed alignment processes ( $freq_{UNSOLVED}$ ). This is achieved through the  $wfreq$  functions for the various segment combinations, which use the fixed  $SOLVED$  and  $UNSOLVED$  weights to prioritise disambiguated frequencies ( $0 < \alpha < UNSOLVED \leq SOLVED$ ). The subtractions by a factor of  $UNSOLVED$  are designed to discount the single occurrences of  $\langle g, p \rangle$  and  $\langle g, p, ctxt \rangle$  in the current align-

ment paradigm, and  $\alpha$  is an additive smoothing constant designed to counter the effects of low frequency counts.

As described above, consideration of lexical context for a given segment tuple  $\langle g, p \rangle$  is four-fold, made up of the single *character* immediately adjacent to  $g$  in the grapheme string and single *syllable* immediately adjacent to  $p$  in the phoneme string (or the null string in the case of a string-initial/final segment), for both the left and right directions. An individual *tf-idf* score is calculated for each of these contexts *ctxt*, and the resultant scores combined by taking the 4-way arithmetic mean. In the case of full-string segment alignment, the overall score is defined to be  $tf(\langle g, p \rangle)$ .

The overall score for all G-P segment tuples contained in the current alignment is computed according to the arithmetic mean of the respective combined *tf-idf* scores, with the proviso that full kana-based grapheme segments are excluded from computation.

Additional allowances for affixation and conjugation are made according to the method described in (Baldwin and Tanaka, 1999b).

## Selective sampling

As described above, a single alignment paradigm is selected for disambiguation on each iteration, and the statistical model updated by way of incrementing frequencies for segment alignments contained in the alignment solution (according to the *SOLVED* weight), and decrementing frequencies deriving from segments occurring in disallowed alignment candidates and not the alignment solution. The method of “selectively sampling” a single alignment solution on each iteration, is performed by calculating a score for each alignment paradigm, and disambiguating the highest scoring paradigm according to the top-scoring alignment candidate contained therein. The score for the alignment paradigm is calculated according to the “weighted log odds” discriminative ratio  $s_1 \log \frac{s_1}{s_2}$ , where  $s_1$  is the score for the top-ranking alignment candidate and  $s_2$  that for the second-ranking alignment candidate within the current alignment paradigm. Intuitively, this balances up maximisation of both  $s_1$  and the disparity between  $s_1$  and  $s_2$ , such that we are after not only alignment candidates which score well, but also alignment paradigms where the top-scoring alignment candidate has a clear empirical advantage over other candidates.

## Multi-pass classification

The second algorithm is inspired by the research of Ling and Wang (1997), who applied the C4.5 classification system (Quinlan, 1993) to unsupervised alignment of English G-P tuples. Specifically, C4.5 was used to predict the phonemic equivalent of English words (graphemic strings), by way of outputting a phoneme for each constituent character in a given character window and combining phonemes to give an aligned phonemic equivalent for the original word. A phonetic transcription for the original word was then used to independently generate alignment candidates, and the alignment candidate most similar to the C4.5-constructed alignment chosen as the alignment solution. Similarly to our incremental learning method, alignment solutions are then fed back into C4.5 as training data, for use in aligning subsequent words. Ling and Wang implemented a number of heuristics to improve the performance of their basic method, including ordering the system inputs in ascending order of alignment cardinality, delaying making a decision in cases of multiple alignment candidates being equally similar

to the C4.5-constructed alignment, and cross-validating held-out partitions of the alignment data against the remainder of the data. The final alignment precision over 33,121 English words exceeded 99.5%.

The alignment accuracy on English is certainly impressive, and suggests the method as promising for Japanese G-P alignment. Unfortunately, however, the case of Japanese G-P alignment is considerably more complex than that of English. Most importantly, as noted above, Japanese phoneme segments often extend over multiple characters for single character grapheme segments even, whereas in the case of English, grapheme segments almost always map onto single phonemes. It was thus possible for Ling and Wang to enumerate the 40 or so possible phoneme segments and have C4.5 choose between them in predicting the phonemic equivalent of each grapheme segment. If we attempted to do the same for Japanese, we would end up with a total of over 56,000 phoneme segments in the case of the data set used in evaluation, making the classification task unmanageable. Additionally, English uses only 26 letters (assuming uniform case), whereas our test data contains 4429 grapheme character tokens and 167 phoneme character tokens. This blow out in the search space suggests the need for a different classification approach.

On the implementation side, Ling and Wang are unable to use negative evidence from discarded alignment candidates, a possibility we look to. We also use the certainty factor values returned by C4.5 in scoring the plausibility of different alignment candidates.

## Algorithm basics

We apply the basic fixed window method suggested by Ling and Wang, but instead of inputting only grapheme context to return a phoneme, we input both grapheme and phoneme contexts to return a binary judgement on the plausibility of a coincident segment boundary existing at the centre of the two context windows. Grapheme and phoneme context is set to 3 characters on either side of the segment boundary, making for a combined window size of 12 characters. To give an example based on *align*<sub>1</sub> from Fig. 1, the classifier “\_, \_, 感, 謝, su, ru, \_, ka, n, sya, su, ru” should produce a judgement of true, corresponding to the leftmost inter-segment boundary in *align*<sub>1</sub> (the ‘\_’ token indicates an empty character beyond the boundaries of the original word, and the underlined component is the grapheme window). “\_, \_, 感, 謝, su, ru, ka, n, sya, su, ru, \_”, on the other hand, is associated with a negative judgement within the context of *align*<sub>1</sub>, as despite segment boundaries existing at the centres of the two context windows, they do not coincide under alignment.

At the same time as classifying input for segment boundary compatibility, C4.5 can be set to output a certainty factor  $cf : 0 \leq cf \leq 1$  for each class. This is particularly useful in comparing the overall plausibility of alignment candidates with little or no segment boundary overlap. In the case of the tuple 大使 [*ta-i-si*] “ambassador”, for example, we need to choose between the three alignment candidates of  $a_1 = \langle \text{大} \odot \text{使} \rangle - \langle \text{ta} \odot \text{i-si} \rangle$ ,  $a_2 = \langle \text{大} \odot \text{使} \rangle - \langle \text{ta-i} \odot \text{si} \rangle$  and  $a_3 = \langle \text{大使} \rangle - \langle \text{ta-i-si} \rangle$ , the second of which ( $a_2$ ) is correct. In sizing up  $a_3$  against  $a_1$  and  $a_2$ , we are making a judgement as to the plausibility of the single segment boundary distinguishing each alignment candidate pairing. However, if  $a_1$  were determined to be more plausible than  $a_3$ , and equivalently  $a_2$  more plausible than  $a_3$ , how could we choose between  $a_1$  and  $a_2$ ? Here, we apply the certainty factors in transferring evaluation across to a numeric comparison.

So as to limit comparison of alignment candidates to only those determined to be legal by the alignment constraints, we cluster “homogeneous” legal alignment candidates together into “packed alignment arrays” and individually score each alignment candidate described therein. **Packed alignment arrays** are of the form  $\langle \text{大?使} \rangle - \langle \text{ta?i-si} \rangle$ , for example, where ‘?’ indicates an optional segment boundary aligning with the corresponding boundary in the opposing string (note that packed alignment arrays can also contain fixed segment boundaries, the score for which contributes to the overall score of alignment). **Homogeneous** alignment candidates are defined as not having any crossing-over of alignment and having all coincident segments aligning similarly.  $a_1$  and  $a_2$  from above are not homogeneous (due to the “大” and “使” segments aligning differently), producing the two packed alignment arrays  $\langle \text{大?使} \rangle - \langle \text{ta?i-si} \rangle$  and  $\langle \text{大?使} \rangle - \langle \text{ta-i?si} \rangle$  for “大使”. A combined score for each alignment candidate is determined based on the average segment boundary certainty factor, and alignment candidates realised in multiple packed alignment arrays (such as  $a_3$  in the two presented arrays) are given the minimum score out of those realisations. The optimal alignment candidate for a given alignment paradigm is then that which produces the maximum mean certainty factor.

While we do not have any pre-annotated training data (preserving the true unsupervised status of our method), we do have disambiguated positive evidence from the “free ride” data disambiguated by the alignment constraints. We can also construct negative evidence from alignment candidates disallowed by the alignment constraints,<sup>3</sup> although here there is potential for segment-level overlap with the correct alignment. We thus take only those segment boundary windows not found within the final alignment paradigm. Finally, we stretch the boundaries of unsupervised learning somewhat in providing the system with positive instances for each hiragana and katakana character, aligning it as a single character in the grapheme and phoneme strings. These instances combine to form the “bootstrap data” with which the system is initialised.

We further order the G-P tuple data set in ascending order of alignment cardinality, in the manner of Ling and Wang, so as to give the system the chance to make easier decisions early on and use the resulting evidence in making decisions of greater complexity later on.

### First pass

In the first pass over the data, C4.5 is initialised with the bootstrap data from above, and run over each alignment paradigm in turn. The classifying decision tree is then updated each time an alignment paradigm is disambiguated, based on the positive evidence described by the alignment solution. Due to the potentially dubious nature of evidence arising from this first pass, we commit ourselves to an alignment solution only in the case that there is no tie in combined certainty factor; in the case of a tie, we reserve our decision for subsequent passes. Additionally, we feed only positive evidence back into C4.5.

### Second and subsequent passes

In the second and subsequent passes, we classify each alignment paradigm according to the combination of the

<sup>3</sup>We take only negative evidence produced through alignment incompatibility between otherwise legal alignment candidates for lexical alternates, and also that produced by Lyman’s Law (Vance, 1987).

original bootstrap data and any positive or negative data generated for other alignment paradigms in the preceding pass, holding out data pertaining to the current alignment paradigm. Negative data is progressively added into the training data throughout the second pass, and maintained through subsequent passes.

In the case of a tie in alignment score, we take the first alignment producing that score. Note that packed alignment arrays are listed in descending order of the number of individual alignment candidates they describe, such that by taking the first alignment candidate to produce the maximum alignment score, we are giving it credit for having won out over a larger number of alignment candidates.

We continue iterating over the data until the combined alignment output converges, that is we attain consistent output over successive passes.

## Evaluation

The proposed systems were tested on a set of 5000 G-P tuples containing at least one kanji, randomly extracted from the combined EDICT Japanese-English<sup>4</sup> and Shinmeikai (Nagasawa, 1981) dictionaries. Any lexical alternates of the 5000 G-P tuples were further added to the test set to give the alignment constraints full scope for application (expanding the test data out to 6503 instances), and the original 5000 G-P tuples manually aligned for use in system performance evaluation. The extra lexical alternate data is used only in applying the alignment constraints and has no bearing on subsequent evaluation. The annotated alignment data was, of course, not available to any of the system configurations at execution time.

The test data was additionally pre-processed into alignment paradigms and sorted into ascending order of alignment cardinality, so as to ensure that the input to the two systems was identical.

The baseline word accuracy for this test suite, based on random selection of an alignment schema from the final alignment paradigm for each G-P tuple, is 44.75%.

Looking first to the incremental TF-IDF learning method, we tested the algorithm with different settings for the parameters *SOLVED*, *UNSOLVED* and  $\alpha$ , and found that the respective values of 1.0, 0.5 and 0.05 produced the best word accuracy of 96.94%. Higher values of  $\alpha$  tended to produce greater levels of under-alignment (chunking together of grapheme and phoneme segments into single super-segments), whereas lower levels of  $\alpha$  produced greater levels of over-alignment (intra-segment segmentation). Larger values of *SOLVED*, on the other hand, tended to inflate the overall rate of both over- and under-alignment. Interestingly, most errors were homogeneous with the correct alignment.

The multi-pass classification method produced a word accuracy of 47.34% on the first pass and 58.18% on the second and third passes, halting on completion of the third pass due to coincidence of output with the second pass. Ties in alignment score were produced for only 78 of the 5000 annotated G-P tuples on the first pass, such that we were able to produce alignments for 4922 G-P tuples. On the first pass, most errors took the form of underalignment, whereas errors on the second and third passes were generally instances of overalignment or non-homogeneous with the correct alignment.

Based on these figures, the incremental TF-IDF learning method is clearly superior to the multi-pass classification

<sup>4</sup><ftp://ftp.cc.monash.edu.au/pub/nihongo>



method, both in terms of raw accuracy and the degree of error in the case of incorrect output. The 58.18% word accuracy for the multi-pass classification approach also contrasts starkly with the 99.5% word accuracy claimed by Ling and Wang for English G-P alignment, although it does well outperform the baseline word accuracy.

To further compare the accuracies of the different methods, we calculated the progressive alignment accuracy over corridors of 250 alignment solution outputs, based on the order of output. In the case of the multi-pass classification method, the order of output corresponds to the order of the original data, in increasing order of alignment cardinality, whereas for the incremental TF-IDF learning method, the order of output is determined by the discriminative ratio values. For both methods, however, the first 895 outputs are the “free ride” alignment paradigms of cardinality one, at word accuracy 100%. The progressive word accuracies for pass 1 and passes 2 and 3 of the multi-pass classification method are presented separately as “MP-P1” and “MP-P2/3”, respectively. We include evaluation of a number of variations on the basic TF-IDF method (“BASIC”) to verify the efficacy of the discriminative ratio, firstly by way of a random sampling method, where a random alignment paradigm is disambiguated at each iteration (“RAND”), and secondly by way of a non-incremental method where all alignment paradigms are disambiguated according to the initial top-scoring alignment candidate and output in descending order of the discriminative ratio value (“DUMP”). The various progressive word accuracies are given in Fig. 2.

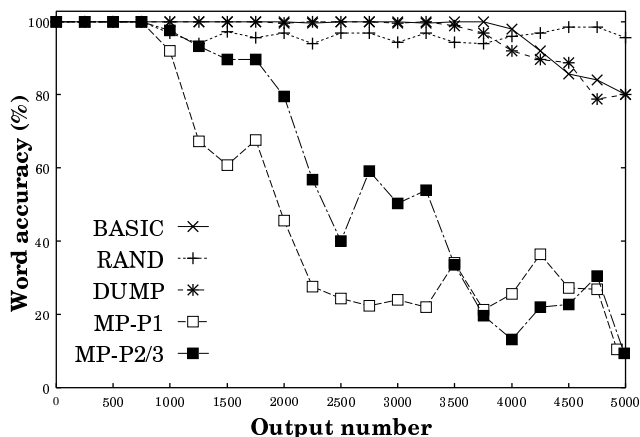


Figure 2: Progressive alignment accuracy

We can see a clear correlation between alignment cardinality and accuracy for the multi-pass classification method, and also a clear performance gain for the second pass over the first pass. The performance gain for the basic TF-IDF method over random sampling (word accuracy 96.64%) and the non-incremental method (word accuracy 96.20%) is more subtle, although the basic method does return higher word alignment accuracy and the output is more consistently accurate over the first 4000 or so annotated outputs, pointing to the success of the selective sampling method.

## Conclusion

In conclusion, we have proposed two fundamentally different methods of unsupervised G-P alignment, and tested them on a set of 5000 Japanese G-P tuples. The first method centres around an adaptation of the TF-IDF met-

ric, and iterates over the data, hill-climbing as it goes. The second method, inspired by Ling and Wang (97), uses C4.5 to determine segment boundary compatibility for combined G-P context windows, and selects the most plausible overall alignment candidate according to the confident factor values returned by C4.5. It makes multiple passes over the data, incrementally enhancing alignment accuracy as it goes. The TF-IDF-based learning method returned a word accuracy of 96.94% in evaluation, surpassing the 58.18% 3-pass word accuracy for the C4.5 multi-pass classification method by a large margin. In the future, we are interested in running the different methods over data for other languages, and expanding evaluation.

## References

- Allen, J., Hunnicutt, M., and Klatt, D. (1987). *From Text to Speech: The MITTalk System*. CUP.
- Baldwin, T. and Tanaka, H. (1999a). The applications of unsupervised learning to Japanese grapheme-phoneme alignment. In *Proc. of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, pages 9–16.
- Baldwin, T. and Tanaka, H. (1999b). Automated Japanese grapheme-phoneme alignment. In *Proc. of the International Conference on Cognitive Science*, pages 349–54.
- Black, A., Lenzo, K., and Pagel, V. (1998). Issues in building general letter to sound rules. In *Proc. of the 3rd ESCA Workshop on Speech Synthesis*, pages 77–80.
- Divay, M. and Vitale, A. (1997). Algorithms for grapheme-phoneme translation for English and French: Applications for database searches and speech synthesis. *Computational Linguistics*, 23(4):495–523.
- Huang, C., Son-Bell, M., and Baggett, D. (1994). Generation of pronunciations from orthographies using transformation-based error-driven learning. In *Proc. of the International Conference on Speech and Language Processing*, pages 411–4.
- Ling, C. and Wang, H. (1997). Alignment algorithms for learning to read aloud. In *Proc. of the 15th International Joint Conference on Artificial Intelligence*, pages 874–9.
- Ling, C. and Zhang, B. (1998). Grapheme generation in learning to read English words. In Mercer and Neufeld, editors, *Proc. of the 12th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI'98*, pages 184–95. Springer.
- Nagasawa, K., editor (1981). *Shinmeikai Dictionary*. Sansendo Publishers.
- Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
- Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–97.
- Sejnowski, T. and Rosenberg, C. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145–168.
- Tsujimura, N. (1996). *An Introduction to Japanese Linguistics*. Blackwell.
- Vance, T. (1987). *An Introduction to Japanese Phonology*. New York: SUNY Press.

# Discovering and Describing Category Differences: What makes a discovered difference insightful?

Stephen D. Bay (sbay@ics.uci.edu)

Michael J. Pazzani (pazzani@ics.uci.edu)

Department of Information and Computer Science

University of California, Irvine

Irvine, CA 92697, USA

## Abstract

Many organizations have turned to computer analysis of their data to deal with the explosion of available electronic data. The goal of this analysis is to gain insight and new knowledge about their core activities. A common query is comparing several different categories (e.g., customers who default on loans versus those that don't) to discover previously unknown differences between them. Current mining algorithms can produce rules which differentiate the groups with high accuracy, but often human domain experts find these results neither insightful nor useful. In this paper, we take a step toward understanding how humans interpret discovered rules by presenting a case study: we compare the responses of admissions officers (domain experts) on the output of two data mining algorithms which attempt to find out why admitted students choose to enroll or not enroll at UC Irvine. We analyze the responses and identify several factors that affect what makes the discovered rules insightful.

## Introduction

Data collection is a daily activity of many organizations in business, science, education, and medicine. Large databases are routinely collected and with the advent of computers to process the information, these organizations want to analyze the data to gain insight and knowledge about the underlying process behind the data. The data usually represents information on their core business, and an important task is understanding the differences between various client groups. For example, bank loan officers may be interested in analyzing historical loan data to understand the differences between people who are good and poor credit risks. Admissions officers at UC Irvine (UCI) are interested in analyzing admissions data to understand the factors which influence an admitted student's choice to enroll at UCI. It is important that the discovered differences both be true and accurate descriptions of the data as well as being acceptable and understandable by the end users.

A common technique for discovering group differences from data is to apply a data mining algorithm to automatically find rules from the data. For example, after analyzing loan data we might find that people with graduate degrees are good loan risks (i.e. grad-degree  $\rightarrow$  low-risk). There have been many studies which investigate the accuracy of rules that describe category differences, but very few which investigate how humans interpret the results.

In this paper, we focus on two issues relating to the interpretation of discovered rules by human domain experts: First, algorithms for automatically finding group differences can be categorized broadly into discriminative and characteristic (or

informative) approaches (Rubinstein & Hastie, 1997). In discriminative approaches, the algorithms attempt to find differences that can be directly used to classify the instances of the groups. In characteristic approaches, the algorithms attempt to find differences in the class descriptions, some of which may also be highly predictive but are not necessarily so. We investigate if human domain experts have a preference for either strategy. Second, there are many objective measures of rule quality and typically mining algorithms seek rules that optimize these measures. For example, with if-then rules of the form  $A \rightarrow C$  (antecedent implies consequent), many algorithms attempt to maximize the confidence which is the conditional probability of the consequent being true given the antecedent ( $P(C|A)$ ). The assumption is that rules that score highly on the objective measure are useful to domain experts. The problem is that while there are many objective measures of pattern quality, such as support (Agrawal, Imielinski, & Swami, 1993), confidence (Agrawal et al., 1993), lift (also known as interest) (Brin, Motwani, Ullman, & Tsur, 1997), conviction (Brin et al., 1997) and many others, none of the measures truly correlate with what human domain experts find interesting, useful, or acceptable. The reality is that most mined results are not useful at all. For example, Major and Mangano (1995) analyzed rules from a hurricane database and reduced 161 rules to 10 "genuinely interesting" rules. In a more extreme, but common case, Brin et al. found over 20000 rules on a census database from which they learned that "five year olds don't work, unemployed residents don't earn income from work, men don't give birth" and other uninteresting facts. Thus we investigate the relationship between human subjective measures of rule usefulness to objective measures of rule quality.

We answer our research questions, "Is a discriminative or characteristic approach more useful for describing group differences?" and "How do subjective and objective measures of rule interest relate to each other?" by reporting on an analysis of discovered rules by human domain experts. We analyzed UCI admissions data to understand the groups of students that decide to enroll or not enroll at UCI given an offer of admission. After discovering rules with two different algorithms, we then showed the rules to human domain experts and asked them to rate the rules according their *insightfulness*, i.e. did the rule expand their knowledge about the admission process? After obtaining experts results, we then analyzed the responses to compare and contrast discriminative and characteristic approaches as well as objective and subjective measures of rule quality.

In the remainder of this paper, we first highlight the differences between discriminative and characteristic approaches and describe the mining algorithms used. We then describe the knowledge discovery task: analyzing admissions data to improve the recruitment process at UCI. We examine domain experts responses and compare them to quantitative measures of rule quality. We conclude by discussing related work and examining possible directions for future work.

## Background: Discovering Category Differences

Mining algorithms for finding category or group differences can be classified as discriminative or characteristic. Discriminative miners attempt to find differences that are useful for predictive classification with a high degree of accuracy. Characteristic miners attempt to find significant differences in the class descriptions. This can result in rules that are highly predictive as with discriminative mining, but predictiveness is not a requirement of the mined rules. Discriminative miners look for one key set of features that distinguish the categories while characteristic miners look for all important differences between the categories.

For example, a discriminative difference would be that students who do not enroll at UCI are much more likely to have a GPA greater than 4 and live more than 100 miles from UCI than students who do enroll. Ninety eight percent of these students (GPA greater than 4 and distance from UCI greater than 100 miles) reject UCI's admission offer and do not enroll. Knowing that a student has these characteristics allows us to classify them with high accuracy.

A characteristic difference would be that 39.8% of students that enroll at UCI are English native speakers compared with 47.9% of students who do not enroll. Clearly this difference affects many students, but knowing that a student is an English native speaker does not give us much information about whether the student will enroll or not. It contains information that is not useful for prediction, but nevertheless may be important to an analyst attempting to understand the two groups.

Formally, we can describe the two approaches as follows: Let  $X$  be the set of attributes and values with which we describe the differences.  $X$  can be a single attribute value pair such as  $X = \{\text{native language} = \text{English}\}$  or it can be a conjunction, e.g.  $X = \{\text{GPA} > 4 \wedge \text{UCI distance} > 100 \text{ miles}\}$ . Let  $y$  be the class or category. Then discriminative approaches attempt to find  $X$  such that the following equation is maximized.<sup>1</sup>

$$|P(y = c_1|X) - P(y = c_2|X)| \quad (1)$$

Characteristic approaches attempt to find  $X$  such that

$$|P(X|y = c_1) - P(X|y = c_2)| \quad (2)$$

is maximized. Note that we can relate these two equations with Bayes Rule:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \quad (3)$$

<sup>1</sup>The exact form of the equation that is maximized can vary somewhat from this definition, but all discriminative approaches concentrate on finding large differences in  $P(y|X)$ .

Thus we can always convert from one to the other. Although the forms can be made equivalent, the difference is that the  $X$  that optimizes/maximizes Equation 1 is not necessarily the same as the  $X$  that is best for Equation 2.

We now describe two algorithms representative of the approaches. C5 (Quinlan, 1993) which is a discriminative approach and STUCCO (Bay & Pazzani, 1999) which is a characteristic approach.

## A Discriminative Approach: C5

A discriminative approach to distinguishing two or more groups from each other is to use a rule learner or decision tree to learn a classification strategy. In this paper, we use the program C5 which is an updated version of C4.5 (Quinlan, 1993). It is a workhorse of the Machine Learning community and is a gold standard to which many new algorithms are compared.

Given two categories  $c_1$  and  $c_2$ , C5 attempts to find sets of variables such that Equation 1 is maximized and so that as many examples in the database are covered by rules as possible. C5 performs greedy heuristic search to develop a decision tree. Starting at the root of the tree, C5 selects an attribute-value test to partition the feature space. Each partition is represented by a child node and is then recursively divided with more tests. The tests are chosen to create child nodes which tend to be mainly of one class.

After finding the tree, C5 can then convert it to rules and remove unnecessary terms imposed by the top down tree structure. It does this by following the path from the root to every leaf and constructing one rule for each path. The rules contain every term that appears in nodes along the path. C5 then tests each term that appears in a rule and removes terms that offer no predictive benefit.

## A Characteristic Approach: STUCCO

Here we briefly review the STUCCO algorithm for mining contrast sets. The reader is directed to (Bay & Pazzani, 1999, 1999b) for a more detailed description.

STUCCO is a complete mining algorithm that searches for contrast sets, conjunctions of attribute-value pairs, that have substantially different probabilities across several distributions or groups. The goal is to find contrast sets where the value of Equation 2 is greater than a threshold  $\delta$ .

STUCCO takes a two stage approach to mining. In the first stage, STUCCO searches for all possible contrast sets that meet the criteria. In the second stage, STUCCO summarizes the mined results to present only a small set of rules.

STUCCO organizes the search for contrast sets using set-enumeration trees (Rymon, 1992) to ensure that every node is visited only once or not at all if it can be pruned. Figure 1 shows an example set-enumeration tree for four attribute-value pairs. STUCCO searches this tree using breadth-first search; it starts with the most general terms first, i.e. those contrast sets with a single attribute-value pair such as  $\text{sex} = \text{female}$  or  $\text{UCISchool} = \text{Engineering}$ . These sets are the easiest to understand and will have the largest support. It then progresses to more complicated sets that involve conjunctions of terms, for example,  $\text{sex} = \text{female} \wedge \text{UCISchool} = \text{Engineering}$ .

During search, STUCCO scans the database to count the support of all nodes for each group. It examines the counts

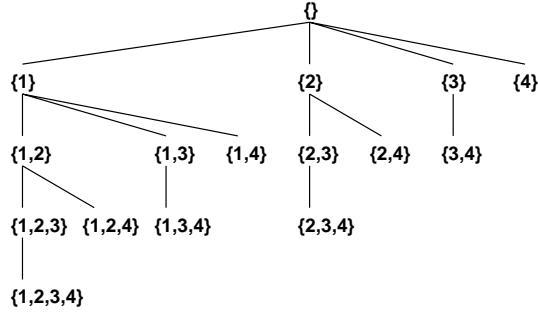


Figure 1: Example search tree for four attribute-values pairs  $\{1,2,3,4\}$ .

to determine which nodes meet the criteria and which nodes should be pruned. STUCCO also explicitly controls the search error to limit false discoveries by keeping careful track of the number of statistical tests made to verify Equation 2 and adjusting the  $\alpha$  level for individual tests to control the overall Type I error rate.

In the second stage, STUCCO summarizes the mined results by showing the user the most general contrast sets first, those involving a single term, and then only showing more complicated conjunctions if they are surprising based on the previously shown sets. For example, we might start by showing the contrast sets  $\text{sex} = \text{female}$ ,  $\text{UCISchool} = \text{Engineering}$ , and  $\text{GPA} > 4$ . STUCCO would then move on to showing more complicated sets such as  $\text{sex} = \text{female} \wedge \text{UCISchool} = \text{Engineering}$  or  $\text{UCISchool} = \text{Engineering} \wedge \text{GPA} > 4$ , and finally  $\text{sex} = \text{female} \wedge \text{UCISchool} = \text{Engineering} \wedge \text{GPA} > 4$ . The conjunctions are only shown if their frequencies could not be predicted from the subsets using a log-linear model (Everitt, 1992). This hierarchical approach eliminates many uninteresting results and can reduce the number of mined results by more than an order of magnitude leaving a small set of rules for a user to view.

## Analysis of UCI Admissions Data

At UCI, the admissions office collects data on all undergraduate applicants. The second author serves on a campuswide committee whose goal is to analyze this data to identify changes that could be made to admissions policies that would improve the quality, quantity, and diversity of students that enroll at UCI. Currently the admissions officers typically analyze the data by manipulating spreadsheets and thus they can only form simple summaries and do not perform detailed multivariate analyses that would be provided by a data mining algorithm.

Here, we report on an analysis of the 1999-2000 enrollment data to identify differences between students who chose to enroll and those who did not for all students accepted at UCI. There were a total of 13344 students given admission offers, of which 3871 accepted and enrolled at UCI and 9473 who did not. For each student, the data contains information on variables such as ethnicity, UCI School (e.g. Arts, Engineering, etc.), sex, home location, first language, GPA, SAT scores, Selection Index Number (SIN) which is a composite

score formed from GPA and SAT scores, statement of intent to enroll, etc. We joined the data with a zipcode database and added fields for the distance to UCI and to other UC schools. Numeric variables, such as SAT scores and distances were manually converted into nominal variables at thresholds that are meaningful for the admissions office.

We ran STUCCO and C5 on the data to obtain contrast sets. For STUCCO we used the following parameter settings:  $\delta = 1\%$  and global  $\alpha = 1$ . For C5 we used the default parameter settings except we set the misclassification costs to balance the different group sizes (typically only 30% of admitted students will enroll). This was necessary as without cost balancing C5 would fail to find any rules distinguishing the two groups and would resort to a default strategy of always predicting that the students would not enroll (the more common class).

Both C5 and STUCCO produce results in their own particular format. To make interpretation easier and to eliminate any bias from the presentation format, we converted the results into an equivalent set of English sentences describing the differences using an identical sentence structure for both C5 and STUCCO. We translated the numeric results associated with the outputs of STUCCO and C5 into *yield* and *gain* which are meaningful quantities for the admissions officers. Yield is the percentage of students that enroll; gain is the difference in the number of students that would enroll if the yield was identical to the average yield. The results can be ordered by gain to highlight the differences that have the largest effect. Rule 1 shows a sample result converted automatically to English text. The full set of results are too big to be shown in this paper, but Appendix A presents a small subset and all examples used in this paper are actual findings.

*Rule 1.* Students who are Korean and have a Selection Index Number between 6000 and 6500 are more likely to enroll with a 30% higher yield than average. This represents a gain of 66 students.

We expressed yield relative to the average. In this example, the yield was 30% higher than average yield (25.6%) which is the percentage of all students who accepted UCI's offer. Thus the yield for this category is  $55.6\% = 30\% + 25.6\%$ . The gain of 66 students is a measure of the *effect size* of the rule, i.e. how many students does it affect? In general, we believe the effect size is a domain independent factor that contributes to how insightful a rule is. For example, in a loan default problem the effect size would indicate how many additional loans that default can be attributed to customers that meet the conditions of the discovered rule.

The discriminative and characteristic approaches resulted in two very different rules sets describing the differences between students who enroll and do not enroll. Table 1 shows the size (as measured by the number of terms in the conjunction) and number of rules mined by C5 and STUCCO. It shows the results for all of the mined rules and the best 30 as measured by gain (we used only the best 30 rules for our experiment in the next section). Examining the table we see that C5 returned far more results than STUCCO and that the individual sets tended to be larger and more complicated. While more complex results are undesirable, by itself, it is not an indication that one method is better than another. Table 2

summarizes the average gain (magnitude only) of the results and in Figure 2 we plot the discovered rules according to their yield difference (with respect to the average yield) and gain. We summarize the differences as follows:

- C5 tends to produce longer rules than STUCCO. The average number of terms in a C5 rule was 3 whereas for STUCCO the average is 1.7.
- Compared with STUCCO, C5 produced rules with higher yield differences but smaller effect sizes.
- C5 produces many rules that are clearly uninteresting because their effect size is very small. For example, on the full set of returned rules the median effect size was only 5 for C5. In contrast, STUCCO’s median was 132.

Table 1: Summary of Results for C5 and STUCCO (S).

Size	All		Best 30	
	C5	S	C5	S
1	25	42	6	17
2	46	24	16	11
3	62	10	6	2
4	35	1	2	
5	21	1		
6	5			
total	194	78	30	30

Table 2: Summary of Effect Size Results for C5 and STUCCO (S). Values were calculated based on magnitude only.

	All		Best 30	
	C5	S	C5	S
median	5	132	125	273
mean	32	175	173	306
min	0	28	34	165
max	495	683	495	683

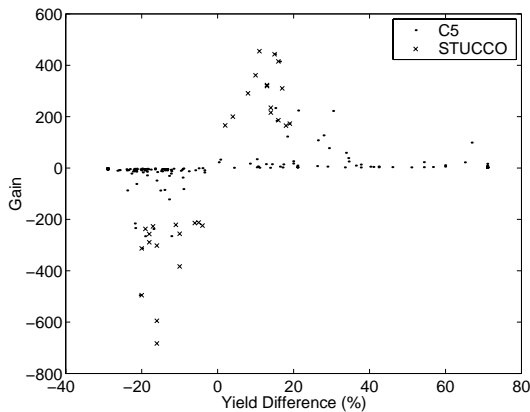


Figure 2: Comparing discriminative and characteristic rule sets.

## Experimental Evaluation

In this experiment, we showed the results from data mining to admissions officers at UCI and asked them to rate the rules according to the insight they provide.

To give an example of an insightful rule, consider Rule 2. Admissions officers at UCI have been uncertain of the effect that the proximity to UCI and other UC campuses plays in student’s college choice. Students who live at home with their parents substantially reduce the cost of higher education. It was well known that students who live close to UCI are more likely to accept offers, but little was understood about how this interacts with other variables. Rule 2 provides insight into this: it suggests that UCI competes fairly well for students with UCLA, UCSD and UC Riverside.

*Rule 2.* Students who live within 30 miles of UCI and live within 30 miles of another UC school are more likely to enroll with a 10% higher yield than average. This represents a gain of 329 students.

As another example of what an admissions officer would find insightful, consider Rule 3. It suggests that UCI does an extremely poor job of recruiting bright students who have not yet declared a major. This is probably because recruiters treated non-declared majors as confused students who needed help rather than as bright students who wanted to explore their options. Due to this discovery, UCI is changing the way it approaches recruiting undeclared students, particularly those with high GPAs.

*Rule 3.* Students who have a GPA greater than 4, and are undeclared majors are less likely to enroll with a 15% lower yield than average. This represents a loss of 123 students.

**Subjects.** The subjects were 4 faculty and staff at the University of California, Irvine who are actively involved in the admissions process and expressed an interest in viewing the results of a computer analysis of admissions data to find factors relating to student recruitment. The subjects did not receive any compensation.

**Stimuli.** The stimuli consisted of two sets of statements corresponding to the outputs of C5 and STUCCO. Each set consisted of 30 rules, 15 describing students with increased yield and 15 describing students with decreased yield. Appendix A shows the 15 increasing yield statements for STUCCO. Within the group of increasing or decreasing yield statements the rules were sorted by gain (largest first). The yield and gain values were rounded to the nearest integer. The subjects were not aware of the algorithm that generated each set of rules.

**Procedures.** Each subject was shown the two sets of rules and were asked to “consider the statements in the context of being an admissions officer whose goal is to improve the quality, quantity, and diversity of students that enroll.” We then asked the subjects to rate each statement on its insightfulness using a scale from -3 to +3, with -3 corresponding to not insightful and +3 corresponding to insightful. After viewing both sets of statements, the subjects were asked to indicate which set they preferred overall.

**Results & Discussion.** Table 3 shows the mean ratings of the experts for both STUCCO and C5. It is clear from the values that the experts are using different scales and that the rules were not equally insightful to all. Experts 1, 2, and 4 rated STUCCO higher than C5. Using a group  $t$  test, the differences were significant for E2 and E4 at the 0.001 level. The difference in ratings were not significant for Experts 1 and 3.

Expert 1 indicated no preference for STUCCO or C5, but the remainder all stated that they preferred the rules which were learned by STUCCO.

Table 3: Mean Ratings for C5 and STUCCO

	Expert			
	E1	E2	E3	E4
STUCCO	0.7	-0.1	1.57	1.23
C5	0.43	-0.87	1.87	0.23
t(58)	0.69	3.48	-1.38	4.87

We pooled the STUCCO and C5 ratings for each expert and then calculated the correlation of the experts ratings with the objective measures of rule quality: yield difference, gain (effect size), and rule size (number of conjuncts). For our calculations we used the magnitude of the yield difference and gain. The results are shown in Table 4. For yield difference and rule size there were no significant correlations with insightfulness. For gain, we found a significant relation for Experts 1 and 2. With a  $t$  test of a correlation coefficient ( $H_0: \rho = 0$ ), the results are significant at the 0.01 level or better. This suggests that the effect size is a factor in determining how insightful domain experts find the discovered differences. Since STUCCO finds many rules with large effect sizes, experts seem to prefer the STUCCO rules.

Table 4: Correlation of Ratings to Objective Rule Measures

	Expert			
	E1	E2	E3	E4
Yield Diff.	0.1406	-0.0827	0.1664	-0.1139
Gain	0.5598	0.4316	0.2469	0.0848
Rule Size	-0.2449	-0.2224	0.0787	-0.2620

We believe there are other factors that influence the ratings of insight given to a rule. In particular, some rules are already well known to the admissions officers. In addition, some admission officers have a particular focus (e.g., minority students) and would be more interested in rules of that type. We tabulated the inter-correlation of the experts in Table 5 using the pooled C5 and STUCCO responses. The results are surprisingly in that the correlation between experts is very low. Figure 3 plots the ratings of E2 and E4, the experts with the highest correlation. This suggests that insight is very subjective and there are important individual differences, possibly relating to the prior knowledge of the task.

Table 5: Correlation of Ratings Between Experts

	Expert		
	E2	E3	E4
E1	0.2748	0.1028	-0.1329
E2		0.1894	0.2778
E3			-0.1613

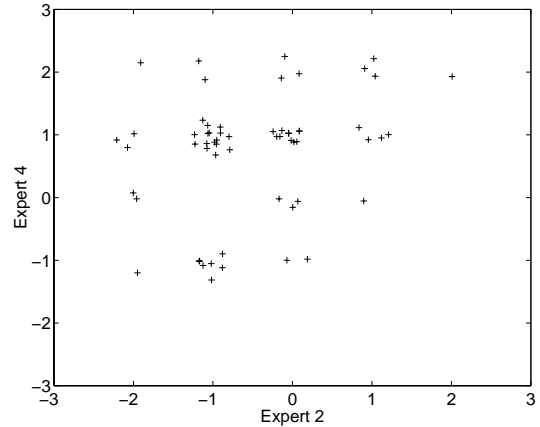


Figure 3: Expert 2 versus Expert 4. <sup>2</sup>

## Related Work

Insightfulness is an extremely difficult notion to capture, and this work has only begun to investigate this concept. Blake and Pazzani (2000) have taken an orthogonal approach to understanding when a rule is insightful. They examined how background knowledge encoded in an electronic knowledge base could be used to remove uninteresting rules from a set. In contrast this work examines how the discovery strategy (discriminative or characteristic) affects the insight of rules found and how insightfulness correlates with objective measures of rule quality.

Clearly, before a rule can be insightful to a person, it must be understood and considered valid. In the past, researchers have considered understandable as synonymous with “short” and thus designed mining algorithms with a strong bias towards rules that are short and accurate under this assumption (Karalic, 1996; Craven, 1996). While this makes intuitive sense, there have been no studies which quantitatively confirm this. There have been two studies which indicate that perceived validity of the mined rules affects the credibility and willingness to use mined results: Pazzani, Mani, and Shankle (1997) examined the effect of monotonicity relationships on rule acceptance in diagnosing potential Alzheimer patients. They found that regardless of rule accuracy neurologists were unwilling to use rules which violated the intent of the diagnostic test. Pazzani and Bay (1999) looked at the effect of incorrect signs on the credibility of regression equations and likewise found that equations where the sign of a variable differed from subjects expectations were rated poorly. An interesting result of their study was that longer regression equations were more credible than shorter equations.

Silberschatz and Tuzhilin (1996) suggested that interestness is a subjective quality that depends on the individual. However, they did not test this theory quantitatively with human subjects. Our results with inter-expert agreement support their theory.

<sup>2</sup>A small amount of random jitter has been added to the points.

## Conclusions and Future Work

We asked the following two questions in our paper: “Is a discriminative or characteristic approach more useful for describing group differences?” and “How do subjective and objective measures of rule interest relate to each other?”

We answered these questions by conducting a study of admissions officers and their responses to the outputs of the data mining algorithms C5 and STUCCO. Our main findings are that (1) characteristic differences are more useful to domain experts than purely discriminative differences. (2) Many objective measures of rule quality correlate poorly with expert opinions on what is insightful, but there is some evidence that effect size is important. (3) Rule insightfulness is highly subjective as even experts examining the rules for the same task do not correlate well with each other.

This paper presented an initial study of how experts with a particular task in mind evaluated data mining discoveries. It is important to use experts because they represent the intended users of the mining programs and will have a distinct purpose in mind when evaluating results. However, the limitation is that experts are inherently rare and thus we were only able to obtain the responses from four people. We plan on conducting a larger study on a less specialized domain so that we can involve more subjects.

## Acknowledgments

This research was funded in part by the National Science Foundation grant IRI-9713990.

## Appendix A

We show here the 15 positive yield rules with the largest gain found by STUCCO.

1. Students who live within 30 miles of UCI are more likely to enroll with a 11% higher yield than average. This represents a gain of 455 students.
2. Students who have a Selection Index Number between 6000 and 6500 are more likely to enroll with a 15% higher yield than average. This represents a gain of 443 students.
3. Students who have a GPA between 2.75 and 3.5 are more likely to enroll with a 16% higher yield than average. This represents a gain of 415 students.
4. Students who live within 30 miles of UCI and live within 30 miles of another UC school are more likely to enroll with a 10% higher yield than average. This represents a gain of 361 students.
5. Students who are from Orange County are more likely to enroll with a 13% higher yield than average. This represents a gain of 323 students.
6. Students who are from Orange County and live within 30 miles of UCI are more likely to enroll with a 13% higher yield than average. This represents a gain of 319 students.
7. Students who scored less than 500 on their SAT Verbal are more likely to enroll with a 17% higher yield than average. This represents a gain of 310 students.
8. Students who scored between 500 and 600 on their SAT Math are more likely to enroll with a 8% higher yield than average. This represents a gain of 291 students.
9. Students who have a Selection Index Number between 6000 and 6500 and scored between 500 and 600 on their SAT Verbal are more likely to enroll with a 14% higher yield than average. This represents a gain of 235 students.
10. Students who have a Selection Index Number between 6000 and 6500 and scored between 500 and 600 on their SAT Math are more likely to enroll with a 14% higher yield than average. This represents a gain of 216 students.

11. Students who scored between 500 and 600 on their SAT Verbal are more likely to enroll with a 4% higher yield than average. This represents a gain of 200 students.
12. Students who have a GPA between 2.75 and 3.5 and have a Selection Index Number between 6000 and 6500 are more likely to enroll with a 16% higher yield than average. This represents a gain of 186 students.
13. Students who have a Selection Index Number between 5000 and 6000 are more likely to enroll with a 19% higher yield than average. This represents a gain of 173 students.
14. Students who live within 30 miles of another UC school are more likely to enroll with a 2% higher yield than average. This represents a gain of 166 students.
15. Students who have a GPA between 2.75 and 3.5 and scored between 500 and 600 on their SAT Math are more likely to enroll with a 18% higher yield than average. This represents a gain of 165 students.

## References

- Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining associations between sets of items in massive databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 207–216.
- Bay, S. D., & Pazzani, M. J. (1999). Detecting change in categorical data: Mining contrast sets. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 302–306.
- Bay, S. D., & Pazzani, M. J. (1999). Detecting Group Differences: Mining contrast sets. *under review*.
- Blake, C., & Pazzani, M. J. (2000). Identifying Insightful Association Rules. *under review*.
- Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 255–264.
- Craven, M. W. (1996). *Extracting Comprehensible Models from Trained Neural Networks*. Ph.D. thesis, University of Wisconsin-Madison.
- Everitt, B. S. (1992). *The Analysis of Contingency Tables* (second edition). Chapman and Hall.
- Karalic, A. (1996). Producing more comprehensible models while retaining their performance. In *Proceedings of Information, Statistics and Induction in Science*, pp. 54–65.
- Major, J. A., & Mangano, J. J. (1995). Selecting among rules induced from a hurricane database. *Journal of Intelligent Information Systems*, 4, 39–52.
- Pazzani, M. J., & Bay, S. D. (1999). The independent sign bias: Gaining insight from multiple linear regression. In *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*, pp. 525–530.
- Pazzani, M. J., Mani, S., & Shankle, W. R. (1997). Comprehensible knowledge-discovery in databases. In *Program of the Nineteenth Annual Conference of the Cognitive Science Society*.
- Quinlan, J. R. (1993). *C4.5 programs for machine learning*. Morgan Kaufmann.
- Rubinstein, Y. D., & Hastie, T. (1997). Discriminative vs informative learning. In *Proceedings Third International Conference on Knowledge Discovery and Data Mining*, pp. 49–53.
- Rymon, R. (1992). Search through systematic set enumeration. In *Third International Conference on Principles of Knowledge Representation and Reasoning*.
- Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6).

# *Harmonia loosely praestabilita*: discovering adequate inductive strategies

Hilan Bensusan  
Christophe Giraud-Carrier  
Department of Computer Science, University of Bristol,  
Bristol BS8 1UB, UK  
hilanb,cgc@cs.bris.ac.uk

## Abstract

Landmarking is a novel approach to inductive model selection in Machine Learning. It uses simple, bare-bone inductive strategies to describe tasks and induce correlations between tasks and strategies. The paper presents the technique and reports experiments showing that landmarking performs well in a number of different scenarios. It also discusses the implications of landmarking to our understanding of inductive refinement.

## Introduction

One of the central goals of cognitive science is to uncover mechanisms that allow agents to produce and manage knowledge. Although informed by existing theories of living organisms, the chief contribution of artificial intelligence, is to investigate knowledge mechanisms in abstract, that is, independently of their psychological or neurological plausibility. Machine learning endeavours to study induction, one of the basis of knowledge production. It considers different inductive strategies, their performance in different scenarios.

Not surprisingly, different inductive strategies are adequate for different inductive tasks. Theoretical results show that there is no inductive strategy that can perform well in every conceivable task (Schaffer, 1994). Some practitioners of machine learning reacted to this predicament by insisting that not every conceivable inductive tasks is equally deserving of attention. If we concentrate on a subset of all conceivable tasks, some people claim that we should restrict ourselves to “real world problems”, we can find a small number of strategies that can handle induction (Rao, Gordon & Spears 1995). The problem arises when we try to give a precise definition for the set of “real world problems”. In any case, we face correlations between sets of tasks, or problems, and induction strategies. Strategies perform well only in a subset of the set of all tasks, this subset is often called the *area of expertise* of a strategy. Machine learning is then left to discover, by induction, correlations between inductive strategies and their area of expertise. One way of doing this is by automating this search for correlations between tasks and strategies. This process is often called *meta-learning* and a number of different approaches has been proposed (see Bensusan (1998,1999), Giraud-Carrier & Hilario (1998), Giraud-Carrier & Pfahringer (1999), Lindner & Studer (1999)). Meta-learning has a number of general consequences for the study of cognition.

This paper explores some of the general consequences of a new way of doing meta-learning, called *landmarking*. The technique has been introduced recently (Bensusan & Giraud-Carrier 2000; Pfahringer, Bensusan & Giraud-Carrier 2000) and some new results are reported here. Landmarking searches for correlations between tasks and inductive strategies by exploring the similarities between different strategies in order to locate the task in a map of areas of expertise. The discovery of similarities between strategies can prove to be a tool to refine inductive strategies and, ultimately, a way to sketch an explanation of human inductive success.

This paper is organised as follows. Next section introduces landmarking. The following section presents experiments that assess its performance. Then we consider some of its implication for the general study of induction and cognition. A last section concludes the paper.

## Meta-learning through landmarking

Meta-learning is the endeavour to automatically discover correlations between tasks and inductive strategies. To simplify without loss of generality, let’s concentrate on supervised learning tasks.<sup>1</sup> These tasks are composed by a set of examples described by attribute values and classified according to a target function. The induction of the difference in extension of the predicates “lemon” and “watermelon”, for example, may include attributes such as COLOUR, SHAPE, SIZE. Something YELLOW, EGG-SHAPED, SMALL qualifies as lemon whereas something GREEN, ROUND, BIG is a watermelon. If the attributes that describe the example are not well-chosen, learning could be very difficult. Consider, as an example, the following worse set of attributes for the “lemon-watermelon” problem above: IS IT A VEGETABLE?, IS IT A FRUIT?, DOES IT FLY?. The two examples are now described as NO, NO, YES. The importance of the example description derives from the fact that inductive strategies rely on representations to generalise. Successful inductive hypotheses are the ones that can represent accurately the similarities and the differences relevant to the task.<sup>2</sup>

<sup>1</sup>Although there are different uses of the terms “induction” and “learning”, in this paper we shall use the terms as interchangeable.

<sup>2</sup>Data representation is important because every learning strategy has what machine learning calls a *representational bias*, a preference for hypotheses with a specific representa-



Meta-learning tasks are inductive tasks. Here, the examples, instead of being lemons or watermelons, are inductive tasks classified according to the best inductive strategy to tackle them. Thus, we have: TASK1 -> NAIVE BAYES, TASK2 -> BACKPROPAGATION, TASK3 -> NEAREST NEIGHBOR etc where each of the inductive strategy mentioned after the arrow is the best strategy for the task before the arrow.<sup>3</sup> The meta-learning task is to use these examples to learn how to classify new tasks in terms of the most suitable inductive strategy. The crucial question for meta-learning is therefore how to describe tasks.

Different approaches to task description have been proposed. These include the use of statistical features of the dataset in the task (Michie et al. 1994) and the use of features of a decision tree representation of the task (Bensusan 1998; Bensusan 1999). In the latter, an inductive hypothesis, namely the one produced by a decision tree induction method, is used to describe the task. Landmarking also makes use of specific inductive methods to describe the task, but makes use of the method's performance rather than the method's induced hypothesis.

The basic idea of the landmarking approach is that the performance of an inductive strategy on a task uncovers information about the nature of the task. Tasks are described by a set of attributes corresponding to the performance of simple, efficient strategies on them. These strategies are expected to indicate which other, more refined strategy is the best to tackle the task. They act, therefore, as landmarkers, indicating where, in the space of all areas of expertise, the task belongs. It explores empirically the relationships between areas of expertise of different learners.

The kind of inference on which landmarking relies can be illustrated with the help of Figure 1. The rectangle represents a set of inductive tasks and the ellipses represent subsets of the set of tasks where a given inductive strategy performs well, that is, areas of expertise. Assume that  $i1$ ,  $i2$ , and  $i3$  are taken as landmarkers. In this case, landmarking concludes that problems on which both  $i1$  and  $i3$  perform well, but on which  $i2$  performs poorly, are likely to be in  $i4$ 's area of expertise etc. Of course, the proximity of the areas of expertise of two strategies indicates some similarity between the inductive mechanisms behind them. For landmarking purposes, however, it is sufficient to concentrate on so-to-speak cartographic considerations. Tasks are described by how some landmarkers fare on them. Exploring the meta-learning potential of landmarking amounts to investigating how well a landmark learner's performance hints at the location of the respective learning tasks in

tion (Haussler, 1989; Russell & Grossof 1990). Thus, most Decision Tree induction algorithms prefer simpler decision trees, most rule induction algorithms prefer simpler rules. There is a trade-off between the need for good input representation and the strength of the strategy's preference (Craven & Shavlik, 1995).

<sup>3</sup>For a survey of the most used inductive strategies including all those to be mentioned in this paper consult Mitchell (1997).

the expertise map.

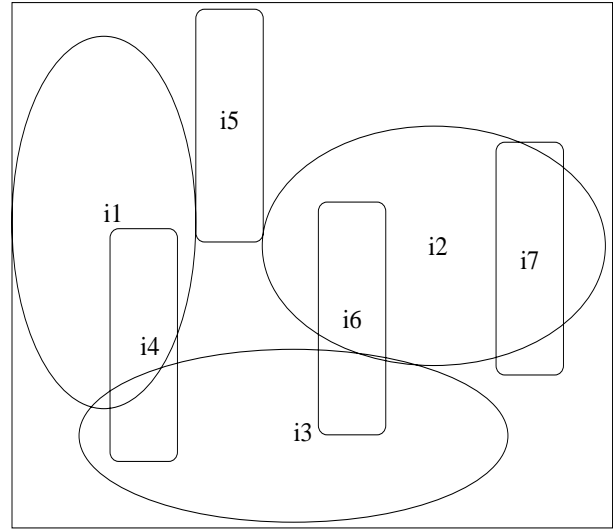


Figure 1: Example of map of areas of expertise

Landmarking relies on some simple and efficient inductive strategies to signpost the location of a task in a map of expertise areas. Landmarking discovers experimentally which inductive strategies are similar enough to have neighbouring areas of expertise. It therefore finds neighbourhoods of inductive strategies and, ultimately, draws a map of areas of expertise. While other approaches represent tasks in a way only indirectly related to their location in an expertise map, landmarking faces them as points in the map to be described in terms of their distance to some known milestones.

Landmarking is a tool to discover the areas of expertise of a learning device. In fact, this is the very goal of machine learning research: to establish the strength and the scope of different inductive strategies. In addition, it highlights which tasks fail to belong to the area of expertise of any of the existing inductive strategies. It can therefore direct the search of new strategies towards areas of the expertise map not currently covered by any learning method. If successful, it can guide the crafting of new inductive methods and work as the basis for a model of inductive refinement. Let's turn now to some experiments designed to measure its success.

## Experiments with landmarking

A number of experiments to test landmarking are reported in (Bensusan & Giraud-Carrier 2000). The results show that landmarking successfully meta-learns in a number of different scenarios. Successful results mean that selection of inductive strategies can be done through the information contained in the performance of some landmark inductive strategies. In this section, we summarise some of these results and present new ones.

Experiments on landmarking can be done only through selecting a set of landmarkers. The landmarkers change according to what we call the *learners pool*, i.e., the set of target learners from which one must be

selected. It remains to be investigated how close can we get from a universal set of landmarkers that can be used in any learners pool. In the experiments reported here, we used the following set of landmark learners. For each, we include the motivation for its inclusion in the set.

1. *Decision node*: A single decision node is chosen according to C5.0’s information gain-ratio (Quinlan 1993, Mitchell 1997). The node is then used to classify test examples. This landmark learner aims to establish closeness to linear separability.
2. *Randomly chosen node*: A randomly chosen attribute is used to split the training set and classify the test examples. This landmark learner informs about irrelevant attributes.
3. *Worst node*: Here the gain-ratio information criterion is used to pick up the least informative attribute to make the single split. This landmarker, together with the first one, is supposed to tell us something else about linear separability: if neither the best nor the worst attribute produce a single well performing separation, it is likely that linear separation is not an adequate learning strategy.
4. *Naive Bayes*: The training set is used to estimate the probabilities required to use the Bayes theorem to classify test cases (Mitchell, 1997). This landmark learner intends to measure how conditionally independent the attributes are, given the class.
5. *1-Nearest Neighbor*: The test set is classified based on the classification of the closest training example (Mitchell 1997). This landmark learner measures how close instances that belong to the same class are.
6. *Elite 1-Nearest Neighbor*: This computes 1-Nearest Neighbor on a subset of all attributes. This elite subset is composed by the most informative attributes if the gain-ratio difference between them is smaller than  $0.1^4$ . Otherwise, the elite subset is a singleton and the learner acts like a decision node learner. This landmark learner intends to establish whether the task is a relational one, that is, if it involves parity-like relationships between the attributes (Clark & Thornton, 1997). In relational tasks, no single attribute is considerably more informative than others.
7. *Majority-class guesser*: The test set is classified according to the most common class in the training set. This landmark learner intends to inform about the frequency of the majority class.
8. *Linear Discriminant*: A linear approximation of the target function is sought (Gama, 1999). This landmark learner intends to measure closeness to simple linear separation.

---

<sup>4</sup>This threshold is based on the results reported in Ben-susan (1999).

The performance of the different landmarkers are given by the average performance on all the existing examples of the induction problem, the so-called *instance space* of the induction made from 10 different subset of examples (*training sets*) of equal size. This approach, although different from the standard practice of *cross-validation* where the sets of examples are drawn without replacement, is an efficient way to estimate how the landmark learners perform in each task. Therefore, inductive tasks are described by landmarker’s performance values. The task is then labelled by the learner with greater average accuracy on the task, using a 10-fold cross-validation approach. Each task can be labelled by a learner’s name or as “tie” when the difference in performance between the best and the worst learner is less than 10%. A (meta-)dataset with 5 examples described by 4 landmarkers looks as follows:

```
0.42187,0.46875,0.46250,0.30781,Ripper
0.45312,0.42187,0.45000,0.26250,IB
0.54687,0.56250,0.45937,0.29844,C5.0tree
0.51562,0.59375,0.43750,0.28750,MLP
0.43750,0.51562,0.43125,0.27812,tie
```

Given the (meta-)dataset, the meta-learner aims at finding correlations between the performances of the learners in the pool and that of the landmarkers.

In the first experiment, we compared landmarking with an existing approach to task description for meta-learning. This approach uses a number of information-theoretical properties of the data to describe the task (Michie et al. 1994). We implemented this information-theoretical approach by considered the following 6 features defined on literature as meta-attributes: Entropy of the class, Average entropy of the attributes, Mutual information, Joint entropy, Equivalent number of attributes, Signal-to-noise ratio. The task was to select among the following 10 learning algorithms: C5.0, C5.0 with boosting, C5.0 rules, Multi-layer perceptron trained with backpropagation (MLP), Radial-based function networks (RBF), Linear discriminant, Ltree (see Gama, 1999), Naive Bayes (NB), Instance-Based inducer (IB) and Ripper. Landmarkers 1,2,3,4,5,6,8 were used. 320 Boolean tasks were considered. The 10 learning algorithms in the learner pool were also used for meta-learning in all experiments. Error rates were based on stratified 10-fold cross-validation. Results are given in Table 1. The first line reports the error rate of the default class that, in this case, was “tie”.

The table shows that landmarking outperforms the information-based task description and therefore it is a suitable competitor. Notice that landmarking outperforms the information-based approach with all of the 10 meta-learners. Moreover, the difference in error is around 10% with the three C5.0 meta-learners. The table also shows that adding the information-based features to describe the task impairs landmarking performance.

Next, we considered a number of learners pools with two inductive strategies. Learners pools were composed by pairs of the following inductive strategies: C5.0(with boosting), C5.0(rules), Naive Bayes (NB), Instance-

Table 1: Comparison between different ways to describe tasks: performances of the landmarking approach (L), the information-based approach (Info) and the combined approach (Combined) using 10 different meta-learners.

Meta-learner	Land	Info	Combined
Default Class	0.460	0.460	0.460
C5.0boost	0.248	0.360	0.295
C5.0rules	0.239	0.333	0.301
C5.0tree	0.242	0.342	0.314
MLP	0.301	0.317	0.320
RBFN	0.289	0.323	0.304
LD	0.335	0.311	0.301
Ltree	0.270	0.317	0.286
IB	0.329	0.366	0.342
NB	0.429	0.407	0.363
Ripper	0.292	0.314	0.295
Average	0.298	0.339	0.312

Based induction (IB), Ripper and Multi-layer perceptron (MLP). Landmarkers 1,2,4,5,6,7,8 were used. Tasks were classified as a *tie* between the two strategies when the average error difference between the learners in the pool was less than 0.1. We used 927 artificially generated Boolean and MONK-like datasets (Thrun et al, 1991). Boolean instance spaces had between 5 and 12 attributes. The error rates given in table 2 are the average 10-fold cross-validation error of 5 inductive strategies used for meta-learning: IB, MLP, C5.0boost, Ripper and Radial Basis Function Network Induction (RBF).

Table 2: Landmarking to choose between pairs of learners

Learner pool	Error
NB-IB	0.383
NB-MLP	0.179
NB-Ripper	0.181
C5.0boost-MLP	0.246
C5.0boost-NB	0.359
C5.0rules-Ripper	0.204

In a different experiment, we looked at the suitability of inductive strategies and groups of similar inductive strategies. We considered that a task is suitable for a learner if it performs better than the average of 10 standard learners: C5.0, C5.0rules, C5.0boost, MLP, RBF, Linear Discriminant, Ltree, NB, IB and Ripper. For this experiment we used only landmarks 1,2,3 and 6 as they are all decision node based and are arguably enough to diagnose at least whether decision tree induction is a good way to approach the task. We used 222 tasks from the set used in the previous experiment and the 10 standard learners mentioned above to perform the meta-learning induction. We looked at the suitability of IB, NB, C5.0boost, neural network inductive strategies (MLP and RBF) in general (NN), rule induction strate-

gies (Ripper and C5.0rules) and decision tree strategies (C5.0, C5.0boost, Ltree). The error rates given in table 3 are the average 10-fold cross-validation error of the 10 inductive strategies used for meta-learning.

Table 3: Suitability of inductive approaches. Error rates for the default class prediction and for meta-learning with landmarking are given.

Approach	Default class	Landmarking
IB	0.420	0.297
NB	0.380	0.298
C5.0boost	0.510	0.449
NN	0.440	0.386
Rules	0.370	0.281
Trees	0.470	0.390

These results show that most meta-learners produce error levels smaller than the default error class and often the difference is substantial. Notice that error rate figures don't reflect the overall performance, that is the accuracy of the selected learning model. In another experiment, we tried to estimate this by using the 222 Boolean problems as tasks of a meta-learning training set and 18 other tasks to test the hypotheses and compare the selected approach with the best performing one. The 18 tasks of the test set were from the standard repository of benchmark induction problems maintained by the University of California at Irvine (UCI); these are commonly considered to be "real world problems". We chose the following problems: mushrooms, abalone, crx, sat, acetylation, titanic, waveform, yeast, car, chess(king-rook-vs-king), led7, led24, tic-tac-toe, MONK1, MONK2, MONK3, satimage, quiscas.

The results reported for this experiment are the average error difference between the best choice and the selected choice in the 18 UCI problems. If the average is in fact better than the chosen model, we consider the error difference between the chosen model and the average. Similarly if the meta-learner had chosen against the model that in fact is better than the average of the 10 learners. Here we used only C4.5 (Quinlan, 1993) as meta-learner. Average error difference appear in table 4.

Table 4: Average error difference between best and chosen option in the 18 UCI datasets

Approach	Error difference
IB	0.0356
NB	0.0165
C5.0boost	0.0443
NN	0.0314
Rules	0.0360
Trees	0.0211

The small average error difference shows that the chosen strategy, even when is not the best, performs well. It shows that landmarking seldom make choices that per-

form considerably worse than the best alternative. This is confirmed further by an experiment in the same scenario. Now we used only the 14 UCI tasks listed above as training set and tested the C4.5 hypothesis in the remaining four UCI tasks (MONK2, MONK3, satimage, quis-clas). The results obtained have a greater variation than the previous one but shows that in some cases landmarking perform completely accurately. Table 5 summarises the new results.

Table 5: Average error difference between best and chosen option in 4 UCI tasks after training on 14 UCI tasks only

Approach	Error difference
IB	0.0675
NB	0.0605
C5.0boost	0.0000
NN	0.0000
Rules	0.0443
Trees	0.0172

These results, although still preliminary, show that landmarking is capable to select inductive approaches. They suggest that it pays off to run bare-bone, landmark inductive strategies on a number of tasks and learn how their performance relates to that of other, more fleshed-out strategies. This far, we have indicated how the performance of simple inducers in a task can be used for meta-learning. We move now to the significance of landmarking for a general theory of induction.

### Discovering inductive strategies

*For me [...] the problem of induction is a problem about the world: a problem of how we, as we are now [...], in a world we never made, should stand better than random or coin-tossing chances of coming out right when we predict by inductions that are based on our innate [...] similarity standard. Darwin's natural selection is a plausible partial explanation.*  
W. V. O. Quine

One of the problems of explaining human (and animal) cognitive practices in general and inductive practices in particular is to account for success. Humans are remarkably good at inducing in familiar environments and seem to make heavy use of their background knowledge accumulated through inductions made in their lifetime history or received as cultural material. Studies on human induction on tasks similar to the MONK problems have established that prior knowledge influences the rate of concept learning, and the logical form of concepts formed during learning is a function of the logical form of the concepts previously acquired (Pazzani, 1991). In general, humans rely on previous acquisition of concepts and common-sense knowledge about the area to learn new concepts (Wisniewski & Medin, 1994; Heit, 1994). Background knowledge and the ability to meta-learn enable humans, when for instance engaged with scientific

theory building, to perform successful inductions from one or few examples.

Human inductive trajectory from innate instincts to refined theories about the world is Quine's view of the problem of induction: a problem about the world. A plausible partial complement to Darwin's natural selection is to find a model of exploiting previous induction experience to boost performance. Such model, of course, has to accommodate the partial explanation role that natural selection plays. The inductive trajectory towards greater efficiency in familiar environments had its origins in evolutionary selection of relevant inductive mechanisms. Recent there have been attempts to characterise human innate inductive tendencies in terms of learning biases (Elman et al., 1996; Dessalles, 1998). Leaving aside the question of how our inductive practices are guided by our innate instincts, we can sketch a model of the human inductive trajectory according to which our similarity standards by means of which we generalise are partly product of evolution, partly a consequence of a gradual process of refinement. We claim that landmarking can be part of an account of inductive refinement.

Landmarking is a technique to select the most adequate inductive strategy for a task, but it can also be seen as an instrument for inductive refinement. It suggests ways in which better, increasingly appropriate inductive strategies, can be constructed from rudimentary ones. Landmarkers are simple inductive strategies that can characterise tasks. Thus, they can outline new inductive strategies to adequately cover areas of the expertise map; describing the area in terms of how different learning biases fare there is a step towards constructing more refined biases that can tackle it. As a way to describe tasks, landmarking has far-reaching consequences beyond strategy selection: to landmark a group of tasks could be the first step towards the development of an inductive strategy to tackle it. This is arguably what happens when a scientist applies various simple methods to a problem in order to get information about what more sophisticated method to develop. This could also be what happens when new problems had to be addressed by humans with only few, unrefined inductive tools. Landmarking is a way to discover relationships between different strategies and, as such, to establish what is needed to ease learning. In this sense, it not only bears similarities with other methods that exploit the nature of the task to decide which way to go (Clark & Thornton, 1997) but also can be seen as a general framework for those methods as it describes tasks only in terms of a portfolio of learning performances. The emerging picture is one where the records of failure and success of the current induction tools are used to inform how these tools need refinement. Successful learning, landmarking suggests, might require learning with previous mistakes and accomplishments.

### Conclusions

*Wär nicht das Auge sonnenhaft, die Sonne koennt'es nie erblicken.* Goethe, Zahme Xenien, Werke, Weimar 1887-1918, bk 3, 1805.

Landmarking is a strategy to describe tasks so that no more than a small class of efficient learning algorithms is required. Tasks are described by their position in the expertise map. It can also be used to locate and explore expertise *terra incognita*. It can be seen as part of a model of inductive refinement whereby the description of a task in terms of landmarks offers the raw material for the development of new induction tools. The picture offered by this model is one in which human inductive abilities are roughly tuned to their environment; no survival and no refinement could start from a completely alien inductive toolkit. Evolution gives part of the explanation. But the gradual refinement that sharpens the kit and assembles new instruments is what turns the original *harmonia loosely praestabilita* into an inductively adapted species.

### Acknowledgements

This work is part of the METAL project supported by an ESPRIT Framework IV LTR Grant (Nr 26.257). We wish to thank the members of the Consortium for useful comments and discussion.

- Bensusan, H. (1988). God doesn't always shave with Occam's Razor – learning when and how to prune. *Proceedings of the 10th European Conference on Machine Learning* (pp. 119-124). Berlin: Springer.
- Bensusan, H. (1999). Automatic bias learning: an inquiry into the inductive basis of induction. *PhD Thesis* School of Cognitive and Computing Sciences, University of Sussex, UK.
- Bensusan, H. & Giraud-Carrier. (2000) Discovering task neighbourhoods through landmark learning performances. *Submitted*
- Clark, A., & Thornton, C. (1997). Trading spaces: computation, representation and the limits of uninformed learning. *Behaviour and Brain Sciences*, 20, 1, (pp. 57-90).
- Craven, M. & Shavlik, J. (1995). Investigating the Value of a Good Input Representation. in: *Computational Learning Theory and Natural Learning Systems*, ed: Petsche, T. and Judd, S. and Hanson, S., Cambridge, MA, USA: MIT Press.
- Dessales, J-L. (1998). Characterising Innateness in artificial and natural learning. *Proceedings of the Workshop on Human and Machine Learning, 10th European Conference on Machine Learning*, Technische Universität Chemnitz, Chemnitz, Germany.
- Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996) *Rethinking Innateness*. Cambridge, MA, USA: MIT Press.
- Gama, J. (1999). Discriminant trees. *Proceeding of the Sixteenth International Conference on Machine Learning*, (pp. 134-142), San Mateo, CA: Morgan Kaufmann.
- Giraud-Carrier, C. & Hilario, M., ed. (1998). *ECML'98 Workshop Notes - Upgrading Learning to the Meta-Level: Model Selection and Data Transformation*, Technische Universität Chemnitz, Chemnitz, Germany.
- Giraud-Carrier, C. & Pfahringer, B., ed. (1999). *Proceedings of the ICML'99 Workshop on Recent Advances in Meta-Learning and Future Work*, J. Stefan Institute, Ljubljana, Slovenia.
- Haussler, D. (1989). Quantifying Inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial Intelligence*, 32, 2, (pp. 177-222).
- Heit, E. (1994). Models of the Effects of Prior Knowledge on Category Learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 20, 6, (pp. 1264-1282).
- Lindner, G. and Studer, R. (1999). AST: Support for Algorithm Selection with a CBR Approach. *Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases*. (pp. 418-423). Berlin: Springer.
- Michie, D., Spiegelhalter, J. & Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification*. London: Ellis Horwood.
- Mitchell, T. (1997). *Machine Learning*. New York, NY, USA: McGraw Hill.
- Pazzani, M. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 17, 3, (pp. 416-432).
- Pfahringer, B., Bensusan, H. & Giraud-Carrier, C. (2000). Meta-learning by landmarking various learning algorithms. To appear in: *Proceedings of the Seventeenth International Conference on Machine Learning, ICML'2000*.
- Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning* San Mateo, CA, USA: Morgan Kaufmann.
- Rao, R.B. and Gordon, D. and Spears, W. (1995). For every generalization action, is there really an equal and opposite reaction? *Proceedings of the Twelfth International Conference on Machine Learning, ICML'95* San Mateo, CA, USA: Morgan Kaufmann.
- Russell, S. J. & Grossof, B. (1990). Declarative bias: an overview. in: Benjamim, P. (ed.) *Change of representation and inductive bias*. (pp. 267-308). Dordrecht, Netherlands: Kluwer.
- Schaffer, C. (1994). A Conservation Law for Generalization Performance. *Proceedings of the Eleventh Conference on Machine Learning, ICML'94*. San Mateo, CA, USA: Morgan Kaufmann.
- Thrun, S. et alii (1991). The MONK's problems - a performance comparison of different learning algorithms. Technical report CMU-CS-91-197. School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, USA.
- Wisniewski, E, J. & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, 18, (pp. 221-281).

## Path and Manner Priming: Verb Production and Event Recognition

**Dorrit Billman** ([dorrit.billman@psych.gatech.edu](mailto:dorrit.billman@psych.gatech.edu))

School of Psychology; 274 Fifth St  
Atlanta, GA 30332 USA

**Angela Swilley**

School of Psychology; 274 Fifth St  
Atlanta, GA 30332 USA

**Meredyth Krych** ([krych@psych.stanford.edu](mailto:krych@psych.stanford.edu))

Psychology Department, Jordan Hall  
Stanford, CA 94025 USA

### Abstract

Path and manner are important organizing dimensions of verb lexicons. We investigated how priming with path verbs, manner verbs, or no priming might influence event processing. Before watching a videotaped target event, subjects were primed by path and manner verbs accompanying other, unrelated events. We found effects of priming verbs on the verbs subjects produced to describe an unlabeled event. We found effects of verb produced on subsequent recognition. We compare these effects from self-generated verbs with effects from experimenter-produced verbs.

### Introduction

Language and vision provide two powerful systems for learning from the world. Information acquired from what we see and what we are told is the basis for much of our knowledge of the world (Jackendoff, 1987). How does language influence processing of visually presented information? Researchers have tested for effect of language on nonlinguistic cognition within- and between- languages.

Tests for within-language effects vary the term or expression accompanying nonlinguistic information and look for effects of language on nonlinguistic cognition. Typically, this research is motivated by questions about effects of schema or expectations on memory, not questions about language per se. For example, experiments from tests of top-down effects (Carmichael, et al, 1932; Gentner & Loftus, 1979; Schooler & Engstler-Schooler, 1990) to tests of eyewitness testimony (Hall, Loftus, & Tousignant, 1984; Loftus & Palmer, 1974; McCloskey & Zaragoza, 1985) have found effects of accompanying labels or descriptions.

Tests for between-language effects have been motivated by the Whorfian (1956) hypothesis. Recent studies have found effects of languages on a variety of cognitive tasks (Gopnik & Choi, 1990, Hoffman, Lau, & Johnson, 1986, Shatz, Martinez, Diesendruck & Akar, 1995; also noneffects Malt et al, in press) including visual memory (Levinson, 1996). Most relevant to the domain we investigate is the

research on event representation, specifically path versus manner information. Researchers have found between-language effects on how path versus manner is expressed (Berman & Slobin, 1994; Naigles et al 1998) and affects similarity (Naigles, personal communication).

Our research investigates within-language effects on event memory for path versus manner information. Specifically, we look at how alternative, descriptive verbs effects visual recognition. We were particularly interested in how path and manner verbs affect memory for path and manner information because these aspects of events seem particularly prominent and important.

Manner verbs refer to the way in which a figure carries out a motion. "Hop," "skip," and "jump" are examples of English manner verbs. Path verbs refer to the trajectory over which a figure moves, typically with respect to another reference object. "Rise," "arrive," and "cross" are examples of English path verbs. Manner and path are two of only a handful of aspects of motion events which are typically conveyed by the verbs of a language. This privilege suggests both aspects are central, important information in other aspects of event cognition. Languages seem to select one of these aspects to be normally conveyed by the verbs, with other information typically carried by 'satellite' constructions outside the verb (Talmy, 1985). In English the verb lexicon is organized around manner information and path information is typically conveyed by expressions outside the verb, specifically prepositional phrases. Many other languages, including Romance languages, typically convey path information in the verb and manner in satellites. Nevertheless, within any one language there is variation in verb meaning as well: English has a handful of path verbs, most lower frequency and of Latinate origin (Levin, 1993). Thus manner and path are two important aspects of verb representation that are systematically expressed in language, their method of expression differs across languages, but there is also some variation within a language.

Our previous research (Billman & Krych, 1998) capitalized on within-language variation (verb choice) to investigate how language information and visual information might be coupled. We presented participants

with video-taped events accompanied by either manner or path verbs. Participants returned for a visual recognition test in which no verbs or labels were presented. The test required discriminating the old items from new items with changed manner of motion or else changed path of motion. We found that type of verb initially spoken by the experimenter interacted with the type of recognition errors. Specifically, hearing a path verb (“exiting”) made participants more likely to correctly reject a changed path foil relative to hearing a manner verb (“skipping”) and hearing a manner verb aided rejection of changed manner foils relative to hearing a path verb.

The current experiment also looks for this disordinal interaction of language at encoding with type of recognition error. It also uses much the same presentation of events at encoding and test. However, the language manipulation is more indirect.

In the current experiment, the participant generated a verb describing the target events and we looked for effects of this participant-generated verb on recognition. We tried to influence the participants' choice of verb by priming (encouraged by other priming effects in language, Bock, 1990). Our primes were experimenter-provided manner or path verbs for unrelated events shown before the target.

We look for 1) effects of priming condition on the type of verbs generated, 2) effects of self-generated verbs on type of recognition, and 3) also for a direct effect of priming condition on type of recognition errors. We expected that effects of labeling by self and by another would be similar, and to this extent expected to replicate and extend our previous findings. However, there might also be differences. Listening to language generated by others might be more likely to focus a listener on aspects of the event not already attended to. Production processes might be more strongly influenced by language-internal factors such as markedness, frequency, and existence of alternative similar forms.

## Method

### Participants

Ninety-nine Georgia Tech students received course credit for participation. Data from the 75 self-reported, monolingual native English-speaking students are reported here.

### Procedure

On the first day participants viewed a series of everyday events. In the Path and in the Manner Condition, some events were labeled with a verb by the experimenter and these labeled trials served as primes. In the No Language (unprimed) Condition no events were labeled by the experimenter. In all conditions there were a few target events unlabeled by the experimenter, and for these the participants were asked to generate their own descriptive verb. On the second day participants returned for the recognition test. No language was provided or generated for any recognition trial. Participants judged whether a presented

scene was identical to one they had seen on the first day or differed in any respect.

### Encoding Session

Participants were told they would see a series of short video-taped events and that they should watch these very carefully. They were told that for some events the experimenter would ask them to write down a verb describing what was happening in the event and they would be asked to do so by questions such as “what is the woman doing?” presented right before the event began. In the Manner and Path conditions the experimenter spoke a descriptive verb or the question roughly four seconds before the event began; in the No Language condition the experimenter said “next scene” to alert the participants, instead of a descriptive verb. An unrelated filler task followed encoding.

### Recognition Session

On the next day participants took a difficult recognition memory task viewing video clips with no accompanying description. All items concerned the scenarios they had seen the day before. Subjects judged whether each video was “identical” to the original clip or differed in any way. Participants responded by marking one end of six-point scale for old items (“Sure Old”) and the other end for new items (“Sure New”). Responses were scored as correct or incorrect in the analyses here. After the recognition task, subjects described events but this data is not reported here.

### Stimuli

At both encoding and recognition, participants viewed video clips of everyday events involving human agents. They lasted 3 to 20 seconds with five seconds of black between scenes. The critical events were designed in sets of three: one original, target event and two foils. The Path Foil changed the path along which the figure moved in the original, target event, while the Manner Foil changed the manner of movement of the figure. Two orders of encoding and of recognition tapes were used.

### Encoding Stimuli

The originally-presented target events were designed to be good examples of both a path and a manner verb, for example, a child **skipping** through a living room to **exit** through the front door, or a woman **crossing** a road, **jogging**. These were the items for which the participants produced descriptive verbs. There were six target events: skip/exit, jog/cross, tiptoe/ascend, float/rise, hop/enter, and fly/descend. Immediately before a target scene, 2-3 priming items were presented. Priming events illustrated unrelated motion events. In the Manner Condition, the experimenter labeled these priming events with English manner verbs while in the Path condition, the experimenter labeled primes

Table 2.  
Verbs Produced for each Event

Type	1 Ascend/ Tiptoe	2 Exit/ Skip	3 Descend/ Fly	4 Enter/ Hop	5 Cross/ Jog	6 Rise/ Float
PATH	Ascend (1) Gohome(1)	Leave(10) Exit (1)		Enter(22) Arrive (3)	Cross(5)	Rise(43)
MANNER	Walk(11) Tiptoe(5) Step(3)	Skip(56) Frolick(1) Hop(1) Prance(1) Tror(1) Walk(1)	Fly (17) Glide(1)	Walk (24) Hop(5)	Jog(43) Run(22)	Float(24) Fly(1) Soar(1)
COMBO	Climb(40) Climb...(6) Go+Climb(1)	SkipOut (1)		WalkThru(1)		
OTHER	Exercise(2) Go(1) Move(1) Progress(1) Hurt(1)	Play(1)	Land(57)	Knock(6) Visit(5) Move(3) Go(1) Pretend(1)	Exercise(5)	Move(1) Stop(1)

with English path verbs and in the No Language condition they were unlabeled.

Thirty events were presented at encoding: 6 target events, 1 description-practice event, 15 priming events, and 8 fillers (to increase the diversity of events presented). Target events, the practice event, the filler events, and the priming events which immediately preceded each target event were identical across all three conditions.

**Recognition Stimuli.**

The 30-item recognition test presented old and new versions of filler ( 8old/8new) and of the target events (6 old and 12 new). Each original target event had a foil with a changed path and a foil with a changed manner. The changes in these foils were designed to be great enough so that the verb originally generated to describe the original event (e.g., “skipping”) would not describe the foil event. For example, in the manner foil for the skip/exit scene the child galloped rather than skipping and in the path foil the child stopped in the door rather than exiting.

Table 1.  
Design of Target and Foils for Recognition Test

Target	Exit	Skip
Path foil	<i>Approach (not exit)</i>	Skip
Manner foil	Exit	<i>Gallop (not skip)</i>

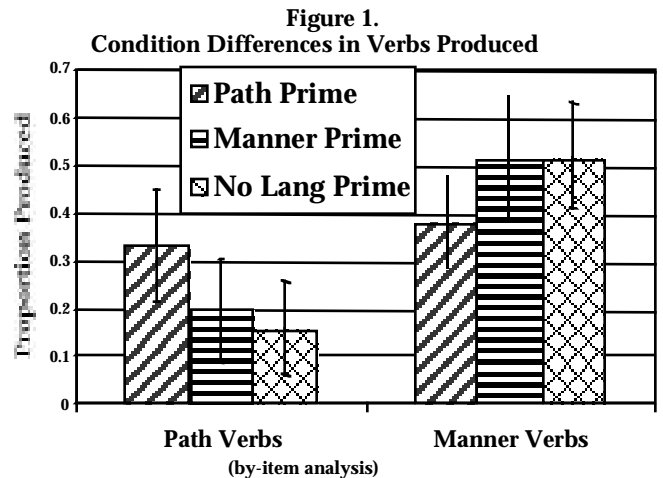
**Design**

Encoding Condition (Manner Verb Prime/ Path Verb Prime/ No Language Prime), a between-subject variable was

crossed with Recognition Item Type (Path Foil/ Manner Foil/ Old), a within-subject factor.

**Results**

We asked how priming affected verb production, how verb production affected recognition, and whether there was a direct effect of priming on recognition. Data analyses throughout are done by-item. Although this gives us small n, it allows a stable unit of analysis both for effects of condition and for conditionalized effects of verb produced.



**Verbs Produced.**

Details of the verbs produced are shown in Table 2. Events varied in the variety of verbs produced and degree of concentration in a few dominant responses. The scenes had been designed to be good illustrations of specific verbs (listed as the event identifier), but they might also be



described by other verbs. The descend/fly scene was the most homogeneous, with 98.7% of responses in the two most dominant verbs, “fly” and “land.” Interestingly, the plane did not in fact land in the original scene, but was very widely classified in terms of the normal activity in the scenario. The enter/ hop and ascend/tiptoe scenes were the most varied with 64% and 72% of responses in the two most dominant verbs. These were also the most varied in terms of numbers of different verbs used and use of phrases. These scenes also evoked verbs focused on additional or more abstract aspects than the simply the movement of the figure.

**Effects of Priming on Verb Type Produced.**

We were particularly interested in whether priming with path or manner verbs would alter the proportion of path and manner verbs produced. Figure 1 shows how the proportion of manner and path verbs produced was influenced by the priming condition. (The proportion of verbs classified either as Combination or as Other was between 28% and 31% across the three conditions). Since we have other response categories, numbers of manner and numbers of path verbs do not necessarily trade off and can be analyzed as two levels of the production variable.

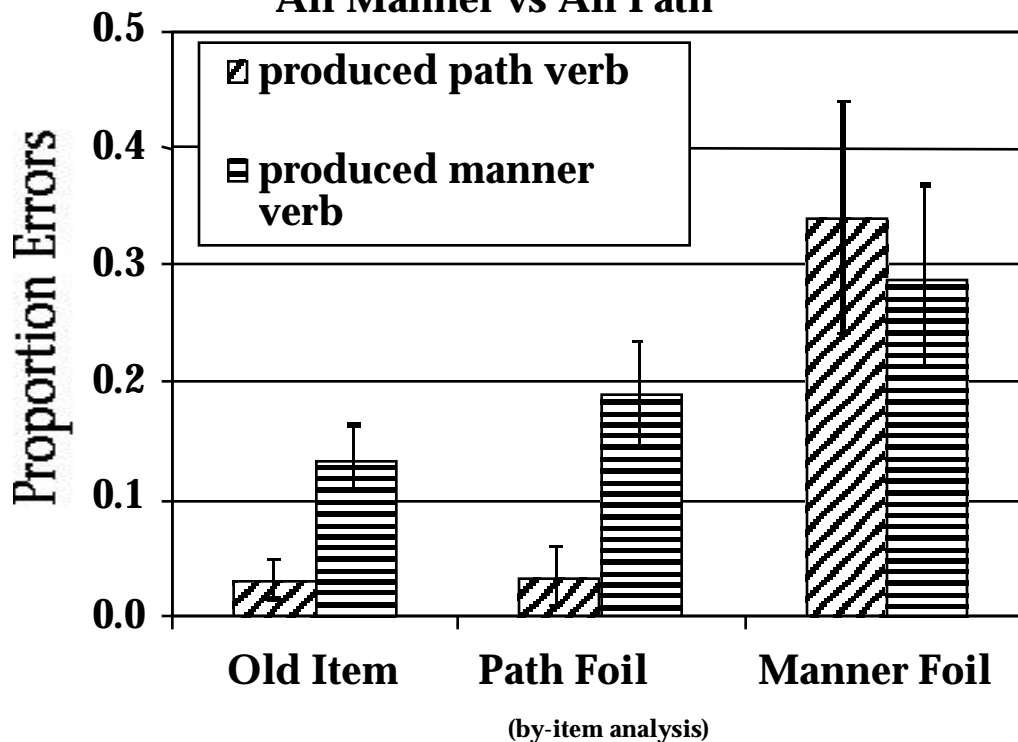
The interaction of priming condition with type of verb produced was significant ( $F[2,10]=8.33, p=.007$ ). Path-priming produced more path verbs and fewer manner verbs than either Manner-priming or no verb priming, which look

similar. Overall, there was not a main effect of priming condition on proportion of combined path or manner responses; 71% of produced verbs were manner or path in the Path-Primed condition compared to 72% in Manner-Primed and 69% in No Verb Priming ( $F<.01$ ). Overall, 49% of responses were manner verbs and 21% were path verbs. Although this preference for manner verbs seems large, items are highly variable and the difference is not significant in a by-item analysis ( $F[1,5] = 2.20, p = .20$ ).

**Effects of Producing Path vs. Manner Verbs on Recognition.**

Given that a manner or path verb was produced, is this production related to subsequent recognition judgments? Figure 2 shows that producing a path versus manner verb benefits recognition in path foils and old items, with a small harmful effect on manner foils. A 2x3 ANOVA (by-item) found that the type of verb produced interacted with type of recognition item in influencing number correct ( $F[2,8]=6.59, p=.020$ ). [Reduced df reflect loss of one event where no path verbs were produced]. The effect of item type was also significant ( $F[2,8]=6.66, p=.02$ ) with the highest error rates coming from false manner recognition, but in this test there was no overall effect of verb produced because of the tradeoff on path versus manner foils. Follow-up analyses localized the effect. A 2x2 ANOVA including manner and path foils but not old items, now showed a significant effect of item type,  $F[1,4]=7.10, p=.056$ , and a significant interaction  $F[1,4]=10.75, p=.031$ ,

**Figure 2.**  
**All Manner vs All Path**



but no main effect of verb produced. Further, paired t-tests showed an advantage of producing a Path Verb for reducing errors on Path Foils ( $p=.017$ ) and on Old Items ( $p=.035$ ), but not on Manner foils.

**Effects of Producing Dominant ‘Path’ or Manner Verb on Recognition.** A second analysis complemented the Manner/Path Produced analysis reported above. The Manner/Path analysis above mixed very different types and frequencies of verbs for each event. Further, for two of the events it excluded the most frequently used verb conveying path information. These two events had common verb responses which carried path information but

not significant, nor was condition,  $F's < 1$ , but item type was,  $F(2,71)=31.8, p<.001$ .

Table3. Proportion Errors by Condition & Item Type

ERRORS	Old Items	Path Foil	MannerFoil
Path Primed	.07	.13	.38
Manner Primed	.11	.14	.32
No Language	.12	.13	.32

## Conclusion

### Summary

We found an effect of priming condition on what verbs subjects produced to self-describe events. Manner verbs were produced more often in the manner-primed than path-primed condition; path verbs were produced more often in the path-primed than manner primed condition. The unprimed condition looked similar to the manner-primed condition.

The fact that we are able to produce this priming in verb use suggests that the linguistically analyzed dimensions of manner and path may be "psychologically real" and influence on-line performance tasks, such as verb generation.

We found that the nature of the descriptive verb produced by participants predicted their later recognition. Errors on manner foils were more likely when a path rather than manner verb had been produced and errors on path foils were more likely when a manner rather than path verb had been produced.

The pattern of results here replicates and extends our earlier studies with experimenter-provided verbs.

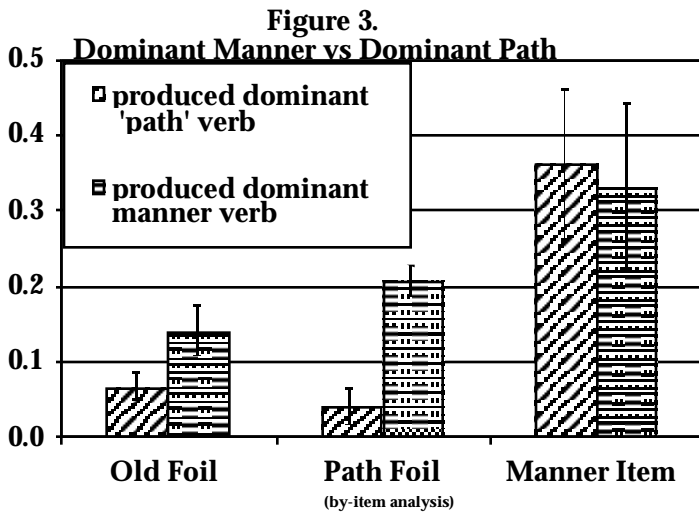
### Interpretation

These findings extend our understanding of how language is implicated in the perception and memory for events. Linguists have analyzed the verb lexicon as organized around distinctions of manner and path (Talmy, 1985). We found that use of path versus manner verbs primes different path or manner verbs used in describing unrelated scenes. This suggests that the dimensions relevant to a formal analysis of the verb lexicon also guide access and verb choice. Manner and path may act as psychological dimensions, perhaps both guiding access in the lexicon and attention in event perception.

The similarity between the recognition findings in this experiment and our prior findings suggests that whether someone hears or produces a verb, the effect is similar: distinctions in meaning carried by that verb influence recognition.

### Future Work

Additional analyses of this data will investigate verb frequency and verb discrimination. Performance with path verbs departs from performance with manner verbs or no language and we are interested in understanding the possible variety of factors which produce this asymmetry between manner and path.



which were not simple path verbs and hence were not included in the Manner-Path Verb analysis. For the “descend/fly” event, no true path verbs were produced and “land” (classified in the Other verb type) was by far the dominant response. For the “ascend/tiptoe” event, “climb” (classified as Combination) was the dominant response, which includes manner as well as path information. Since the path component of these two verbs was clearly the relevant aspect for these scenes, we designed these ‘path-verbs’ for a supplemental analysis. In this Dominant ‘Path’-Manner analysis, we looked at the effect of two verbs for each event: the one most frequently used ‘path’ or the one most frequently used manner-verb. This analysis includes more data than the first, but fewer verbs.

The results parallel the first analysis. In the 3 (Item Type) x 2 (Dominant Verb Produced) ANOVA (by-item), item type was significant,  $F(2,10)=6.70, p=.014$ , the interaction of item type and dominant verb produced was significant,  $F(2,10)=7.25, p=.011$ , but not the effect of verb produced  $F(2,10)=3.85, p=.107$ .

### Direct Effect of Priming on Recognition Type.

We also measured whether the subjects primed with path or manner (or unprimed) differed in recognition error types, not considering what sort of verb they generated, as shown in Table 3. The interaction of condition and item type was

Sometimes participants generated verbs which discriminated the target and foil event and sometimes the verbs did not discriminate. For example, if a participant said "running" this would apply to both to the original jogging scene and to the dash manner-foil, hence not discriminating target from foil. Analyzing effect of whether a path or manner verb does or does not distinguish foil from target will help identify how the verbs have their effect.

We are also interested in identifying what information about an event is made more memorable by different verbs, and what the mechanism of influence is. Verbs might exert their influence in guiding attention at encoding, in providing a more structured or integrated representation, or in serving as a separate retrieval cue during recognition.

## References

- Berman, R. A. & Slobin, D. I. (1994). Relating Events in Narrative: a Crosslinguistic Developmental Study. Hillsdale, NJ: Lawrence Erlbaum Assoc.
- Bekerian, D.A. & Bowers, J.M. (1983) Eyewitness testimony: Were we misled? Journal of Experimental Psychology: Learning, Memory, and Cognition, 1, 139-145.
- Billman, D.O., & Krych, M. (1998). Path and manner verbs in action: Effects of "Skipping" or "Exiting" on event memory. In Program of the Twentieth Annual Conference of the Cognitive Science Society. Erlbaum: Hillsdale, NJ.
- Bock, K. (1990). Structure in language: Creating form in talk. American Psychologist, 45, 1221-1236.
- Carmichael, L., Hogan, H.P., & Walter, A.A. (1932). An experimental study of the effect of language on the reproduction of visually perceived forms. Journal of Experimental Psychology, 15, 73-86.
- Choi, S. & Bowerman, M. (1992). Learning to express motion events in English and Korean: The influence of language specific lexicalization patterns. In Lexical and Conceptual Semantics, B. Levin & S. Pinker (Eds.) Elsevier/Blackwell: Cambridge, MA.
- Gentner, D. & Loftus, E. (1979). Integration of verbal and visual information as evidenced by distortions in picture memory. American Journal of Psychology, 92(2), 363-375.
- Gopnik, A. & Choi, S. (1990). Do linguistic differences lead to cognitive differences? A cross-linguistic study of semantic and cognitive development. First Language, 10, 199-215.
- Hall, D.F., Loftus, E.F., & Tousignant, J.P. (1984). Postevent information and changes in recollection for a natural event. In G.L. Wells & E.F. Loftus (Eds.) Eyewitness Testimony. Cambridge University Press: Cambridge.
- Hoffman, C., Lau, I. & Johnson, D. R. (1986). "The Linguistic Relativity of Person Cognition: An English-Chinese Comparison." Attitudes and Social Cognition.
- Jackendoff, R. (1987). On beyond zebra: The relation of linguistic and visual information. Cognition, 26, 89-114.
- Kay, P. (1996) Intra-speaker relativity. In Rethinking Linguistic Relativity, J.J. Gumperz & S.C. Levinson (Eds). Cambridge University Press: Cambridge.
- Levin, B. (1993). English Verb Classes and Alternations: a Preliminary Investigation. U. Chicago Press: Chicago.
- Levinson, S.C.. (1996) Relativity in spatial conception and description. In Rethinking Linguistic Relativity, J.J. Gumperz & S.C. Levinson (Eds.). Cambridge University Press: Cambridge.
- Malt, B.C., Sloman, S.A., Gennari, S., Shi, M., & Wang, Y. (in press). Knowing versus naming: Similarity and the linguistic categorization of artifacts. Journal of Memory and Language.
- McCloskey, M. & Zaragoza, Z. (1985) Misleading post event information and memory for events: Arguments and evidence against memory impairment. Journal of Experimental Psychology: General, 114, 3-18.
- Munnich, E., Landau, B., & Doshier, B. A. (November, 1997). Universals of spatial representation in language and memory. Poster session presented at the 38th Annual Meeting of the Psychonomic Society, Philadelphia, PA.
- Naigles, L. R., Eisenberg, A. R., Kako, E. T., Hightler, M., & McGraw, N. (1998). Speaking of motion: verb use in English and Spanish. Language and Cognitive Processes
- Schooler, J.W. & Engstler-Schooler, T.Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. Cognitive Psychology, 22, 36-71.
- Shatz, M. J., Martinez, I. M., Diesendruck, G., & Akar, D. (SRCD March 30- April 2, 1995--Indianapolis). "The Influence of Language on Children's Understanding of False Belief."
- Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), Language Typology and Syntactic Description, vol. 3: Grammatical Categories and the Lexicon. Cambridge University Press.
- Whorf, B. L. (1956). Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf. T. B. Carroll (ed.). Cambridge, MA: MIT Press.

# Reasoning from shared structure

Sergey Victor Blok (s-blok@northwestern.edu)  
Dedre Gentner (gentner@northwestern.edu)

Department of Psychology, Northwestern University,  
Evanston, Illinois

## Abstract

Two experiments contrasted the predictions of the similarity-coverage model of category-based induction with those of a structure-based account. We focused on the two theories' ability to account for the paradoxical fact that both monotonicities (increases in argument strength with the addition of premises) and non-monotonicities (*decreases* in argument strength with addition of premises) occur in human reasoning. The results are mainly in accord with the structure-based account and are inconsistent with the similarity-coverage account.

## Introduction

### Monotonicity and Induction

Humans routinely make inductive inferences, and the principles that guide these inferences have received a great deal of empirical attention (López, 1995; McDonald, Samuels & Rispoli, 1996; Osherson, Smith, Wilkie López & Shafir, 1990; Sloman, 1993). One principle that has both intuitive and empirical support is *monotonicity* – the principle that confidence in an inductive inference should increase with the number of supporting premises. For example, Osherson et al. showed that adults preferred Argument B over Argument A.

- A. All FOXES have sesamoid bones,  
All PIGS have sesamoid bones,  
Therefore, all GORILLAS have sesamoid bones
- B. All FOXES have sesamoid bones,  
All PIGS have sesamoid bones,  
All WOLVES have sesamoid bones  
Therefore, all GORILLAS have sesamoid bones.

However, robust *nonmonotonicities* have also been documented. Osherson et al.'s participants chose Argument C over D.

- C. All FLIES have sesamoid bones,  
Therefore, all BEES have sesamoid bones.
- D. All FLIES have sesamoid bones,  
All ORANGUTANGS have sesamoid bones,  
Therefore, all BEES have sesamoid bones.

Sloman (1993) and McDonald et al. (1996) have also documented nonmonotonic responding in adults. Even more strikingly, Lopez, Gelman, Gutheil & Smith (1992) showed nonmonotonicity effects very early in development; in fact, nonmonotonicity effects were reliably obtained earlier than monotonicity effects. People appear to believe that more premises make for a stronger argument, except when more premises make for a weaker argument. How can we reconcile these apparently contradictory phenomena?

### Similarity-Coverage Model

A pioneering theory of argument strength is the Similarity-coverage model (SCM) of Osherson et al. (1990). The two components of SCM are *similarity* -- the extent of feature overlap between premise and conclusion categories -- and *coverage* -- the average similarity of the premises and the instances of the lowest level taxonomic category that includes both the premises and the conclusion. The similarity-coverage model predicts monotonicity when the additional premise is a member of the same lowest level superordinate category as the initial premises and the conclusion. It predicts nonmonotonicity when the additional premise is not a member of the lowest level superordinate category. Thus nonmonotonicity can be seen as a kind of dilution effect, as illustrated by Osherson et al.'s (1990) data in (1) and (2), respectively.

- (1) a. ROBINS, SPARROWS / SEAGULLS >  
b. ROBINS / SEAGULLS<sup>1</sup>
- (2) a. ROBINS, RABBITS / SEAGULLS <  
b. ROBINS / SEAGULLS

Argument (1) is monotonic; adding the extra premise SPARROW in (1a) adds an additional piece of premise support without diluting the category coverage, because it fits within the lowest-level category (BIRDS) that applies in the single premise case (1b). In contrast, the additional premise RABBITS in (2a) raises the lowest-level common category to ANIMALS, thus diluting the category coverage. Thus the SCM can successfully predict some instances of monotonicity.

---

<sup>1</sup> Research in this area typically uses so called "blank" or opaque properties – such as 'has sesamoid bones' to ensure that belief in the conclusion is derived from the premise statements, rather than from prior beliefs about the truth of the conclusion. We will omit property names from further examples.

However, as Sloman (1993) noted, there are other instances of nonmonotonicity that are not explainable by dilution of category coverage. His participants found (3b) to be stronger than (3a).

- (3) a. CROCODILE, KINGSSNAKE / ALLIGATOR
- b. CROCODILE / ALLIGATOR

Even though the lowest level taxonomic category (REPTILE) does not change across these arguments, nonmonotonicity<sup>2</sup> is observed. Sloman acknowledges, however, that his own feature-based induction theory is also unable to explain nonmonotonicities.

### Structure-Based Induction

We propose a *structure-based induction* approach that uses structural overlap instead of overall similarity or feature overlap to predict argument strength. Our model is very different from the previous theories in that we explicitly assume that the evaluation of argument strength is accomplished by a process of aligning the representations of the premise(s) and the conclusion.

Specifically, we assume that the perceived strength of an induction from premise to conclusion depends on the goodness<sup>3</sup> of the common schema. For the one-premise case, this idea is closely related to similarity in Osherson et al.'s account and with feature overlap in Sloman's account. But when there are multiple premises, we postulate a *premise comparison process* whereby a common schema is derived from the premises. This schema is then aligned with the representation of the conclusion statement.

This variant of the *progressive alignment hypothesis* (Kotovsky and Gentner, 1996; Kuehne, Gentner & Forbus, 2000; Kuehne, Forbus, Gentner & Quinn, 2000) states that carrying out a comparison involves alignment of structured representations (e.g. Gentner & Markman, 1997).

There is evidence that structure-mapping theory captures some important aspects of inductive reasoning. Wu and Gentner (1998) told participants that a conclusion had attribute  $a_1$ . They were also told that two different premise kinds  $P_1$  and  $P_2$  also had  $a_1$ . Participants were then given the option of inferring an attribute from  $P_2$  that was causally connected to  $a_1$  or an attribute from  $P_1$  that was not causally connected to  $a_1$ . Results indicated that people strongly preferred to reason from a causal base ( $P_2$ ) over an attribute base ( $P_1$ ). See Clement & Gentner (1996) and Lassaline (1996) for related findings.

The SBI view makes several specific predictions. First, it predicts that *monotonicity* (in at least the weak sense)

<sup>2</sup> Monotonicity can be interpreted in the strong sense of *increasing monotonicity* or in the weaker sense of *non-decreasing monotonicity*. Note that even the latter, weaker sense is violated by these examples.

<sup>3</sup> We will use the term *goodness* of the common schema as a shorthand for *structural evaluation*; it depends on the size and depth of the common schema.

will result when the additional premise is alignable with the other premises and the conclusion. Second, conversely, *nonmonotonicity* should result when the additional premise is not alignable with the premises (even if it is alignable on other grounds with the conclusion).

These two assertions predict the monotonicity of argument (1) and the nonmonotonicity of (2). A further point is that the predictions of the SBI model do not rely on taxonomic category structure. Neither monotonicity nor nonmonotonicity are influenced by whether the additional premise belongs to the lowest common category that includes the premises and the conclusion. Thus SBI explains Sloman's example (3) above by noting that the goodness of alignment between the premise CROCODILE and the conclusion ALLIGATOR is diminished by first aligning CROCODILE with KINGSSNAKE.

The third prediction of SBI is that the properties inferred depend on the particular aligned schema. That is, people base their inferences (even about nominally blank properties) on the specific alignment between premises and conclusion, and not on a general sense of similarity. Because the quality of the premise-conclusion alignment determines both the specific properties people are willing to infer *and* the argument strength, we expect a strong association between these two (see Heit & Rubinstein, 1994, for a related proposal).

### Experiment 1. Two vs. three premises

In this experiment, we varied category coverage and alignability in order to contrast the predictions of the similarity-coverage model and the structure-based approach. We used five variants of each argument: a two-premise item plus four kinds of additional premises that were added to make three-premise arguments (Table 1).

Table 1. Sample base two-premise item and the additional premise in the four variant conditions in Experiment 1.

ROBINS, EAGLES, ... / BATS			
Coverage			
Alignment	C+	C-	
	A+	SEAGULLS	AIRPLANES
	A-	DOGS	TV's

The premises and the conclusion of the two-premise arguments shared a common relational schema, such as *flight* or *underwater habitat*. The three-premise arguments were constructed by adding an additional premise to the two-premise arguments. There were four types of additional premises, constructed according to a 2x2 design of *alignability* with the two-premise schema and *category coverage* – i.e., whether the additional premise belonged to the lowest level category spanning the two premises

and the conclusion (hereinafter abbreviated *spanning category*).

For example, given the two-premise argument ROBIN, EAGLE / BAT, the aligned schema presumably involves flight and the spanning category is ANIMAL. The four kinds of additional premises are as follows:

1. **A+C+** type: Alignable with the 2-premise schema (*High Alignment*) and a member of the lowest-level spanning category (*High Coverage*).  
e.g., ROBINS, EAGLES, SEAGULLS / BATS
2. **A-C+** type: Not alignable with the 2-premise schema (*Low Alignment*), but a member of the spanning category (*High Coverage*).<sup>4</sup>  
e.g., ROBINS, EAGLES, DOGS / BATS
3. **A+C-** type: Alignable with the 2-premise schema (*High Alignment*), but not a member of the spanning category (*Low Coverage*).  
e.g., ROBINS, EAGLES, AIRPLANES / BATS
4. **A-C-** type: Not alignable with the 2-premise schema (*Low Alignment*), nor a member of the spanning category (*Low Coverage*).  
e.g., ROBINS, EAGLES, TV'S / BATS

## Method

37 Northwestern University undergraduates were presented with 40 inductive arguments, one at a time on a computer, and asked to rate them according to “how well the conclusion follows from the premises.” There were eight sets, each with five argument types (8 two-premise arguments plus 4 x 8 = 32 three-premise arguments).

For example,

*Fact:*

All ROBINS have property F.

All EAGLES have property F.

*Therefore,*

All BATS have property F.

After rating all the arguments, participants were given a printed packet with the forty arguments they had just rated and were asked to write down their best guess about the property associated with each argument. They were also given the option of skipping any items for which no property had come to mind.

## Predictions

Table 2 summarizes the predictions of the two models. The structure-based induction model predicts monotonicity for alignable types (A+C+ and A+C-) and non-monotonicity for non-alignable types (A-C+ and A-C-),

<sup>4</sup> The extra premise for the A-C+ type always belonged to the same superordinate as the conclusion. This had the effect of giving the A-C+ type the highest relative coverage of any of the 3-premise arguments, as defined by the similarity-coverage model. Importantly, the A-C+ type had higher coverage than the A+C+ type, providing a very strong test of the alignment model against the coverage model.

relative to the two-premise arguments. The similarity-coverage model predicts monotonicity for high coverage types (A+C+ and A-C+), and nonmonotonicity for low coverage types (A+C- and A-C-).

Table 2. Summary of predictions of the two models

Theory	Prediction
SCM	<b>A+C-</b> , <b>A-C-</b> ≤ 2P ≤ <b>A+C+</b> , <b>A-C+</b>
SBI	<b>A-C+</b> , <b>A-C-</b> ≤ 2P ≤ <b>A+C+</b> , <b>A+C-</b>

Another line of prediction concerns the subjects' guesses about the blank properties. According to the structure-based view, the same process of structure-mapping that gives rise to the goodness of the common schema also gives rise to its specific content. Thus we predict (1) people's *confidence* in their property guesses will increase with their subjective argument strength; (2) the *uniformity* of property guesses will increase with their subjective argument strength; and (3) both the confidence and the uniformity of property guesses will be greater for alignable types than for non-alignable types.

## Results

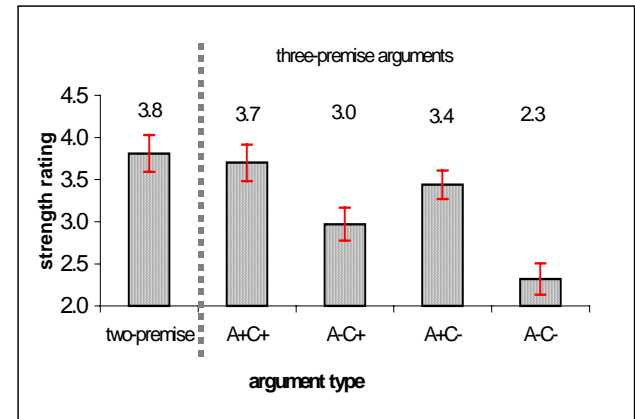


Figure 1. Argument strength ratings for five argument types in Experiment 1 (error bars are 95% confidence intervals).

**Argument Strength Ratings.**<sup>5</sup> Figure 1 shows the mean ratings across items. As predicted by the structure-based account, monotonicity (in the weak, though not the strong form) held when the additional premise was alignable. That is, there were no significant differences in judged strength between two-premise arguments ( $M = 3.81$ ;  $SD = 1.31$ ) and either the A+C+ type ( $M = 3.70$ ;  $SD = 1.28$ ) or the A+C- type ( $M = 3.44$ ;  $SD = 1.25$ ),  $t(36) = 1.46$ ,  $p > .008$ ,  $t(36) = 2.49$ ,  $p > .008$  respectively. Also as predicted, nonmonotonicity held when the additional premise was nonalignable. Arguments of the A-C+ type ( $M =$

<sup>5</sup> We performed six planned comparisons on the mean argument strength for each subject within a type, setting the two-tailed Bonferroni corrected alpha value at 0.008.

2.97;  $SD = 1.15$ ) and the A–C– type ( $M = 2.32$ ;  $SD = 1.14$ ) were rated reliably lower than two-premise arguments,  $t(36) = 4.34$ ,  $p < .008$ ,  $t(36) = 5.96$ ,  $p < .008$ , respectively.

There were no significant differences based on category coverage. In the crucial comparison of the two models, we found that A+C– arguments were rated reliably stronger than the A–C+ type,  $t(36) = 3.01$ ,  $p < .008$ , suggesting that alignability, not category coverage, best predicts the effects of adding a third premise to a two-premise argument.

**Property Guesses.** To test the relation between argument strength and likelihood of listing a property (confidence), we scored the listings on whether a participant chose to guess a property. There were 1241 guesses and 439 (26% of the total) “no guess” responses. The highest proportion of guesses was elicited by the two-premise and the A+C+ argument types (94% and 90%, respectively). The A+C– argument type also elicited a high proportion of guesses (85%). The A–C+ and A–C– types elicited substantially fewer property guesses (65% and 36%, respectively). Overall, the proportion of property guesses closely mirrored the argument strength ratings,  $r = .82$ ,  $p < 0.0001$ .

To test our predictions concerning property uniformity, we rated the content of the property guesses. When subjects were presented with alignable arguments (i.e., two-premise, A+C+ or A+C– types), subjects almost unanimously provided guesses specific to the hypothesized common schema. When presented with a non-alignable argument, subjects tended to provide general and haphazard guesses and tended to disagree about the nature of the blank property. To test this intuition, we asked two naive raters to score the property listings on the basis of *coherence*. Confirming our hypothesis, alignable arguments elicited highly focused patterns of property guesses, while non-alignable ones displayed little agreement between subjects, as observed by our independent raters. Mean coherence rating across the forty different arguments were correlated with argument strength at  $r = 0.599$ ,  $p < 0.0001$ .

## Discussion

The results of Experiment 1 largely bear out the predictions of the structure-based induction model. The effect of adding a premise to a two-premise argument depends entirely on whether the third premise is alignable with the schema that holds in the two-premise argument. If the third premise is alignable, the argument strength remains constant; if the third premise is nonalignable, the argument strength decreases. The predictions of the similarity-coverage model were not borne out for either monotonicity or nonmonotonicity. The SCM predicts monotonicity if the third premise belongs to the lowest-level spanning category of the two-premise argument; and nonmonotonicity when the third premise forces an increase in the level of the spanning category. Neither prediction held.

The most direct contrast between the models is to compare A–C+ items (low alignability but high coverage) with A+C– items (high alignability but low coverage). Participants found the A–C+ premise sets to be a far weaker inductive base than A+C–, with its specific schema. For example, argument (4a) was weaker than (4b):

(4a) ROBINS, EAGLES, DOGS / BATS

(4b) ROBINS, EAGLES / BATS

Thus increasing in the number of premises even while holding coverage constant can result in nonmonotonicity if the alignment is diminished. Indeed, (4c) is judged stronger than (4a), despite clearly having poorer coverage

(4c) ROBINS, EAGLES, AIRPLANES / BATS

In short, our nonmonotonicity findings support the claims of the structure-based framework over those of the coverage model.

The property guess findings were also consistent with the predictions of the structure-based framework. There was a strong connection between considering an argument strong and having a clear idea of what property was being inferred. This observation is consistent with our claim that the process at work here is an alignment process that results in a specific common schema.

Overall the results are encouraging. However, one point requires discussion. We found evidence of *nondecreasing* monotonicity but not of *increasing* monotonicity. There was no *increase* in argument strength for any argument type. This contrasts with Osherson et al.’s (1990) report that strength increased from two- to three-premise arguments. We suspect much of the difference stems from the fact that, whereas we used a single-argument rating task, Osherson et al. used a choice task. Comparing arguments to choose the stronger could have led to heightened contrast between the two- and three-premise arguments.

Structure-mapping does not predict a steady increase in argument strength as additional premises are added.<sup>6</sup> However, it does predict an increase when going from one-premise to two-or-more-premise arguments (always provided the added premise(s) are alignable), because alignment highlights the common structure (Gentner & Wolff, 2000). To test this prediction, we asked subjects to rate single-premise arguments matched to the multi-premise arguments used in Experiment 1. This will allow us to compare (albeit across experiments) the strength of one-premise vs. three-premise arguments.

A second motivation for Experiment 2 was to rule out a possible confound, namely, that the gain in strength for the additional premises was simply due to an increase in overall similarity (or feature overlap, on Sloman’s (1993) account) brought about by the additional premise, rather than by interactions among the premises as claimed by the structural account.

<sup>6</sup> This is because progressive alignment cannot *increase* the size of the common schema. Thus if increases in argument strength do occur when, say, 11 premises are increased to 12, the explanation must lie with other factors beyond alignment.

## Experiment 2. Single-premise arguments

Participants evaluated single-premise arguments. For each argument, the premise was the *additional premise* used in Experiment 1. For example, for the BATS item in Table 1, the four arguments tested in Experiment 2 were

- (A+C+)<sup>7</sup> SEAGULLS / BATS
- (A-C+)<sup>7</sup> DOGS / BATS
- (A+C-)<sup>7</sup> AIRPLANES / BATS
- (A-C-)<sup>7</sup> TELEVISIONS / BATS

The first question is whether, as predicted by structure-mapping, single-premise arguments will be weaker than their three-premise alignable counterparts in Experiment 1. The second question is whether the relative strengths of the three-premise arguments in Experiment 1 are mirrored by the strengths of the corresponding single premises (Thus undermining our premise-comparison account.)

### Method

16 Northwestern University undergraduates saw 32 single-premise arguments (8 items x 4 types) and rated them for strength. The procedure was identical to that in Experiment 1, except that the arguments were given in printed form, rather than on a computer.

### Results

We contrasted the mean argument strengths by argument type between Experiments 1 and 2.<sup>8</sup> Figure 2 presents the mean strength ratings across argument types.

As predicted by structural framework, among alignable types, there was a reliable advantage for three-premise over one-premise arguments (a difference of 1.28,  $t(51) = 3.90$ ,  $p < 0.001$ ). For non-alignable types, this difference was 0.53,  $t(51) = 1.84$ ,  $p > 0.05$ , n.s. Also, as predicted, planned comparisons within alignable types revealed reliable differences between the three-premise A+C+ ( $M = 3.70$ ,  $SD = 1.28$ ) and the single-premise (A+C+)<sup>7</sup> types ( $M = 2.49$ ,  $SD = 0.88$ ),  $t(51) = 3.43$ ,  $p < 0.005$ . A reliable contrast was also observed between the three-premise A+C- ( $M = 3.44$ ,  $SD = 1.25$ ) and the single-premise (A+C-)<sup>7</sup> type ( $M = 2.08$ ,  $SD = 0.85$ ),  $t(51) = 3.95$ ,  $p < 0.001$ .

Planned contrasts for the non-alignable types revealed a non-reliable difference between the three-premise A-C+ type ( $M = 2.97$ ;  $SD = 1.15$ ) and the single-premise (A-C+)<sup>7</sup> type ( $M = 2.85$ ;  $SD = 1.22$ ),  $t(51) = 0.34$ ,  $p > .70$ , n.s.

So far, the results are consistent with the structural account. However, a reliable difference was also observed between the three-premise A-C- type ( $M = 2.32$ ;  $SD = 1.14$ ) and the single-premise (A-C-)<sup>7</sup> type ( $M = 1.38$ ;  $SD = 0.54$ ),  $t(51) = 3.15$ ,  $p < 0.005$ . This result is not predicted by the structural account.

<sup>7</sup> We will refer to single-premise versions arguments by adding a prime to the three-premise symbol: e.g., (A+C+)<sup>7</sup>.

<sup>8</sup> Because the sample sizes across the two experiments were not equal, we also performed a set of more conservative non-parametric analyses, which revealed the same pattern.

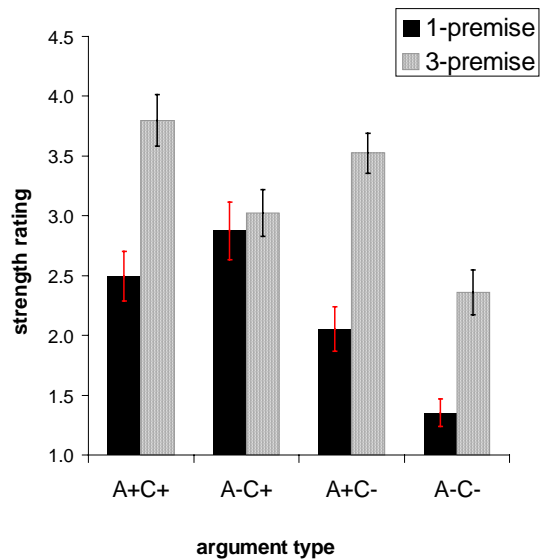


Figure 2. Argument strength ratings for four argument types in Experiment 2 and Experiment 1 (Error bars are 95% confidence intervals).

Turning to the second question, we found that the pattern of strength among single-premise arguments could not account for the three-premise results in Experiment 1. Indeed, the mean strength in the one-premise arguments was significantly *higher* for the nonalignable premises than for the corresponding alignable premises.<sup>9</sup> This is the *opposite* direction from what happened in Experiment 1, where there was an alignability advantage for the three-premise versions of these arguments. This means that the alignability advantage in Experiment 1 cannot result simply from independently accruing similarity or feature overlap across the premises.

### Discussion

Our hypothesis that alignable three-premise arguments would exhibit strong monotonicity relative to their single-premise counterparts was supported. For both of the alignable types (A+C+ and A+C-), three-premise arguments received higher ratings than their respective single-premise counterparts.

<sup>9</sup> That is, the mean strength of (A-C+)<sup>7</sup> arguments was significantly higher than for (A+C+)<sup>7</sup> arguments ( $M = 2.87$ ,  $SD = 1.21$ ;  $M = 2.49$ ;  $SD = 0.85$ , respectively,  $t(16) = 2.665$ ,  $p < 0.025$ ). (A-C+)<sup>7</sup> arguments were also rated reliably higher than the (A+C-)<sup>7</sup> arguments ( $M = 2.05$ ;  $SD = 0.83$ ),  $t(16) = 3.407$ ,  $p < 0.025$ .



## General Discussion and Conclusion

These experiments offer support for the structure-based model of induction. The alignment approach predicts both nonmonotonicities and monotonicities accurately. When the additional premises are alignable, argument strength increases between one- and multiple premises, and is weak-monotonic from two- to three- premises. Strong monotonicity holds for alignable added premises.

Osherson et al's (1990) similarity-coverage model predicts monotonicity except when the additional premise forces a taxonomically higher spanning category. But the results of Experiment 1 showed nonmonotonicity even when category coverage was constant, as well as weak monotonicity despite a decrease in coverage. Across the board, (weak) monotonicity was observed between two- and three-premise cases for just those cases where the additional premise was alignable. The pattern in Experiment 2 was similar: With one exception, monotonicity between one- and three-premise arguments was observed only for alignable arguments.

Further evidence that argument strength judgments involve thinking about the specific relational schema, as opposed to overall similarity, comes from the property listings in Experiment 1. When given alignable third premises, subjects not only rated the arguments as strong, they also had clear opinions on what "Property P" might have been, and those guesses were highly uniform. These findings are consistent with there being a specific schema that emerged from the alignment.

What is the broader significance of these findings? First, premise comparison process must be a part of argument strength models. We have documented both (A) and (B) occurring *simultaneously*:

(A)  $P1, P2, P3 / C > P1, P2, P4 / C$  [Exp. 1]

(B)  $P3 / C < P4 / C$  [Exp. 2],

Since the same premises are added to both sides in going from B to A, this reversal cannot be explained in terms of accruing overall similarity or total feature overlap. It requires an explanation in terms of premise interactivity. People are not integrating individual premise-conclusion argument strengths (e.g. "P1/C + P2/C + P3/C") but aligning premises to determine what aspects of the premises *as a set* are relevant to the argument.

The evidence for premise interactivity presented here poses a challenge to the feature-based induction theory (Sloman, 1993). As an important theoretical alternative to the coverage model, the feature-based theory assumes that instead of computing category coverage, people are assessing total feature overlap between the premises and the conclusion. Monotonicity is predicted because the addition of a premise must either increase total feature overlap or maintain it. The addition of a premise can never decrease total feature overlap, so nonmonotonicities cannot be predicted. The systematic nonmonotonicities we have observed, as well as the evidence of premise interactivity, are inconsistent with the current formulation of the feature-based induction model.

Sloman (1993) has suggested an extension to the feature-based model -- a premise comparison mechanism that weighs common features of the premises more heav-

ily than unique attributes. This might allow the feature-based model to predict some nonmonotonicities. However, it is unclear which common features of the premises will be weighted over others. An important forte of the structure-based model is that it *constrains* similarity by treating matching attributes that play similar roles in their respective concepts as more similar than matching attributes that do not (Medin, Goldstone & Gentner, 1993). Thus, inductive inferences are appropriately constrained.

## Acknowledgments

This research was supported by the National Science Foundation and the Office of Naval Research. We thank Ken Kurtz, Sven Kuehne, Jeff Loewenstein and Kathleen Braun for their assistance at all stages of this project. We are also grateful to an anonymous reviewer for helpful comments on an earlier draft of this paper.

## References

- Clement, C. A., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, 15, 89-132.
- Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52, 45-56.
- Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20(2), 1-12.
- Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, 67, 2797-2822.
- Kuehne, S.E., Forbus, K.D., Gentner, D. & Quinn, B. (2000). SEQL- Category learning as incremental abstraction using structure mapping. *Proceedings of the 22nd meeting of the Cognitive Science Society*.
- Kuehne, S.E., Gentner, D. & Forbus, K.D. (2000). Modeling infant learning via symbolic structural alignment. *Proceedings of the 22nd meeting of the Cognitive Science Society*.
- Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(3), 754-770.
- Lopez, A., Gelman, S. A., Gutheil, G., Smith, E.E., 1992. The development of category-based induction. *Child Development* 63, 1070-1090
- McDonald J., Samuels M. & Rispoli J. (1996). A hypothesis-assessment model of categorical argument strength, *Cognition*, 59, 199-217
- Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, 97, 185-200.
- Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, 25, 231-280.
- Wu, M., & Gentner, D. (1998). Structure in category-based induction. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 1154-1158.

# Problem Solving: Phenomena in Search of a Thesis

**Bruce D. Burns** (burnsbr@msu.edu)  
Department of Psychology, Michigan State University  
East Lansing, MI 48824-1117, USA

**Regina Vollmeyer** (vollmeyer@rz.uni-potsdam.de)  
Institut für Psychologie, Universität Potsdam, Postfach 601553  
14415 Potsdam, Germany

## Abstract

Problem solving has been the study of a set of phenomena rather than a set of theories. Newell & Simon's (1972) concept of search has proved very useful for describing problem solving but it is not a testable theory. We point out that without testable theories, thought about problem solving cannot progress through the interaction of thesis and antithesis. Problems solving requires theories and we propose a specific form of multispace search theory. The hierarchical three-space theory of problem solving can be derived from existing literature, and proposes that interactive search occurs in instance space (states of the problem), rule space (possible rules that govern the problem), and model space (the general understanding of the problem). This theory could be used to generate testable predictions regarding the interaction of spaces and provides a way to try to unify diverse phenomena.

## Problem Solving: What is the theory?

Sixty years ago, Maier (1940) noted that problem solving was frequently cited as a barren field of inquiry. He argued that this blandness is due to the proliferation of experimental tasks which render generalization difficult. It is bland because there is a set of phenomena, but no underlying explanation of them.

Has this blandness diminished since Maier wrote this? If one examines the way the field presents itself to its first line of consumers, undergraduates taking introductory courses in cognitive psychology, then it is arguable that the same problems identified by Maier (1940) continue to bedevil problem solving research. When one picks up a typical introductory cognitive psychology book and turns to the sections on perception, attention, or memory, then one finds a lively description of competing theories and the evidence used to support/discredit them. These are on-going debates so different books present these debates in different ways depending on the biases of the author. In contrast, there is a remarkable similarity between different books when one turns to the section on problem solving. Anderson (2000) is a typical example. He covers procedural knowledge and search, operators (including analogy), operator selection (including means-ends analysis and the Tower of Hanoi),

representation (including functional fixedness), and set effects. Other books may differ in their details, but cover the same basic ground. What is noticeable by their absence, are theories of problem solving.

There has been progress since 1940, in particular Newell and Simon's (1972) idea of problem solving as search of a problem-space. This has been very valuable both for psychological and computational approaches to problem solving, but it is not a complete theory (as others have also noted, such as VanLehn, 1989). As a language for thinking about problem solving, search has proved to be useful and enduring. However, it makes no testable predictions, so there can be no competing theory.

Sternberg (1995) emphasizes the dialectical progression of ideas in psychology. As described by Hegel (1807/1931), the dialectic begins with a viewpoint that is proposed and believed (a thesis), but in response a competing view arises (an antithesis), and eventually the best features of both are melded into a synthesis. Then the process starts again. Sternberg argues that much of the history of psychology can be seen in terms of the dialectic. This progression cannot occur in the field of problem solving because there is no thesis, therefore there can be no antithesis, and there can be no synthesis. Anderson (2000) and other introductory cognitive psychology books illustrate that we know quite a lot about the phenomena of problem solving, but there is no thesis for the phenomena. VanLehn (1989) provides a list of robust empirical findings regarding problem solving. (Space precludes us from taking on a related difficulty with problem solving research, its definition. The definition has varied from very narrow [essentially the study of solving the Tower of Hanoi] to the very broad [every goal oriented activity]. For the purpose of this paper we use VanLehn's characterization of problem solving as multistep goal-directed tasks that last a few minutes to a few hours.)

The aim of this paper is to try to present a thesis, or to at least struggle towards that aim. In doing this we do not wish to throw away the power of treating problem solving as search, instead we want to try to wield it into a form that presents a testable thesis. We do this by taking seriously another part of Maier's (1940) paper, that problem solving is not a single process, but a set of processes. This leads to multispace search theories, and we propose a specific form of a hierarchical three-space search theory of problem solving. As may already be apparent this paper is

speculative and, merely uses existing ideas, but we wish to show how these ideas can be put together in such a way that a hierarchical three-space theory falls out.

## Multispace Search

**Problem Solving as Search.** Newell and Simon (1972) proposed that for every problem there exists a problem space which is defined by three components: 1) the *initial state* of the problem; 2) a set of *operators* that can transform a problem state; 3) a *test* for whether a problem state constitutes a solution (this may be a particular *goal state* or set of goal states). Finding a solution is a process of searching the set of states logically defined by the initial state and the operators that can be applied, until a solution is found. This terminology has proved to be useful for describing a wide range of problem solving behavior. However, to encompass a wider range of phenomena, this framework has been extended in two ways. In order to include induction and problem solving within the same framework, Simon and Lea (1974) claimed that search occurs in a *dual-space*. In order to capture the influence of different representations, Hayes and Simon (1974) claimed that an *understanding process* is required as well as a search process.

**Dual-Space Search.** Simon and Lea (1974) proposed that problem solving does not necessarily consist of search of a *single* problem space. To encompass multiple spaces they generalize the components of Newell and Simon's (1972) description of problem solving in the following way: 1) the elements of a problem space are *knowledge states*; 2) operators are *generative processes* that take a knowledge state as input and produce a new knowledge state as output; 3) there are one or more *test processes* for determining solution and for comparing knowledge states; 4) there are *selection processes* for which of these generators and tests to employ, on the basis of the information contained in the knowledge states. Induction can then be related to problem solving by allowing a dual-space search to be conducted. The search for rules that describe a task is conducted in a *rule space*, the states of which are all possible rules, and the operators are processes for generating, modifying and testing rules. Testing, however, requires movement within *instance space*, consisting of all possible states of the task, and the operators are processes allowed by the task for moving between instance states. Thus the two problem spaces are conceptually distinct, but intimately related; test processes for rule space lead to the generation of instances, whereas information that results from such instances leads to movement in rule space.

Simon and Lea (1974) suggested that many induction tasks can be described in this dual-space framework. For example, in concept attainment tasks learners generate possible rules from instances. They then test or select between alternative rules by observing or creating relevant instances. Thus concept attainment can be seen as a dual-space search, in which the goal is in rule space.

Problem solving may involve search of instance space only, but it is a dual-space search if a problem solver tries to learn rules which can be generally applied to reaching

different goal states. Simon and Lea (1974) pointed out that the Tower of Hanoi problem is usually thought of as search of instance space: find a sequence of moves that transfers all disks to the goal peg. This is a single space search. But the task could be described as: find a *rule* for transferring disks from one peg to another (e.g., the first move depends on whether the number of disks is odd or even). This requires a dual-space search.

Dual-space search has been extended to scientific reasoning by Klahr and Dunbar (1988) in their Scientific Discovery as Dual Search (SDDS) model. They proposed that in scientific reasoning people have an *hypothesis space* (similar to rule space) and an *experiment space* (similar to instance space). Reasoners propose hypotheses, and test them by conducting experiments. Klahr and Dunbar (1988) and Dunbar (1993) found that subjects who tested hypotheses performed a learning task better, a finding which supports SDDS.

More evidence that dual-space search occurs in problem solving was found by Vollmeyer, Burns, and Holyoak (1996). They had participants learn to control a complex system called *biology-lab* in which they could manipulate inputs and observe the changes to the outputs. Ultimately they had to bring the system to a set of output states, but participants were not told the nature of the set of equations linking inputs to outputs. Vollmeyer et al. manipulated the goals of problem solvers by either telling them what the goal state was before they started exploring the system (a specific goal), or delaying informing them of the goal until after they had explored (a nonspecific goal). They found that a group given a specific goal learned less about the structure of the biology-lab task and transferred more poorly to a new goal, than did the nonspecific goal group. The strategies of the specific goal group indicated search of instance space (i.e., find a path to the goal), but the nonspecific goal group instead appeared to test rules (i.e., search hypothesis space).

**Understanding Processes.** Before a problem solver can attempt a problem, the problem instructions must be understood. The importance of understanding processes in natural language has been well illustrated (e.g., Bransford & Franks, 1971). Hayes and Simon (1974, 1977) explored the impact that understanding can have on problem solving. Hayes and Simon (1977) gave subjects different isomorphic versions of the Tower of Hanoi problem and found a dramatic effect on solution ease from the ease of understanding the problem description. The importance of representation in problem solving was a point emphasized long ago by the Gestaltists (e.g., Maier, 1930).

Hayes and Simon (1974) incorporated understanding into Newell and Simon's (1972) framework by proposing understanding as a subprocess that cooperates with search of the problem space. The search process is driven by the result of understanding processes, rather than the problem itself. However, it may not be that understanding processes first produce a representation of the problem, and then search takes over. The two processes may alternate or even blend together (see Hayes & Simon, 1974). That representations may be fluid and interact with attempts to

solve a problem, is a point also made by researchers working within other frameworks (e.g., Burns, 1996; Hofstadter, 1995).

## A Hierarchical Three-space Theory

**Integrating Understanding Processes.** If understanding processes create the representation of problem space, then in a dual-space search theory these processes must create the representation of not only the instance space, but of the rule space as well. Thus understanding processes define the instance states that can be searched, and do so via the candidate rules that might govern instance states. The research on functional fixedness (e.g., Maier, 1930) can be seen in terms of the problem solver's understanding processes defining the wrong rule space. Similarly, the research on how false assumptions can be a barrier to solution can be viewed in this way. Weisberg and Alba (1980) showed that problem solvers attempting the nine-dot problem could only solve it when their assumption that they could only draw lines within a restricted area was removed. In our terms, they were searching the wrong rule space. Of course, having the correct rule space does not guarantee success (as Weisberg & Alba found) as having the correct rule space to search is not equivalent to having the correct rule.

Given that representations may change during problem solving, understanding processes can be seen as conducting a form of search. VanLehn (1989) suggested that schema selection can be a form of search when a person is uncertain as to which schema to select. For example, Larkin (1983) gave expert physicists a straightforward, but difficult, physics problem to solve. Although two of the five physicists immediately selected the correct schema for solving the problem, the other three physicists tried two or more schemas. In this way, understanding processes can be seen as operators that search a space consisting of different representational states. These operators generate, modify, and test the adequacy of representations. We see representational states as encompassing more than just what type of diagram is used, additionally they reflect the problem solver's current model of how a task works. Thus we term this space *model space*.

**Model Space.** In our hierarchical three-space theory of problem solving we propose that model space provides not only the representation of instances, but also defines the rule space to be searched. Which rules appear plausible will depend on how a problem solver thinks the task works. For example, if each component of a system is thought to be independent, then rules proposing interactions will not be considered. If the model changes, then interactions may become part of states in rule space.

Current utility is the criteria for assessing one's state in model space as there is no final goal state, a better understanding of the task may always be possible. So instead of a test for "solution", there may be tests for the adequacy of a model state, that is, does this model seem to work?

Although it violates our application of the term "problem

solving" to tasks completed within a few hours, for expository purposes we will illustrate model space with the debate over competing models of light. Two models of light were proposed: a wave model and a particle model. The hypothesis that a scientist would test depended on which model the scientist believed. The wave model suggested that light is a wave, therefore a relevant hypothesis to test was whether light shows interference patterns. The particle model suggested that light is a particle, therefore a relevant hypothesis to test was whether light exerts pressure. Testing these hypotheses led to movement in the model space for light. Neither model was accepted as completely correct, instead the competing models were synthesized into a model in which light was both a particle and a wave. Although this particular movement in model space was slow, it still had the characteristics of a movement in a problem space. There were clearly defined states (initially two different models states, which expanded in number when the possibility of combinations arose), and there were processes for comparing and moving between states (driven by search of rule space). There were no processes for deciding whether the final goal state in model space had been reached, only utility. The current model of light does not rule out the possibility that a new model may emerge.

During problem solving, movement in model space may occur much faster than did movement in the model space for light. Whenever people are faced with a new task, it is necessary to form a model of that task, the current state of which may need to quickly and often be revised, just as Larkin's (1983) physicists did.

Search in rule space can drive movement in model space. For example, if the rules suggested by a model fail, then eventually the response will be to change the model. If the rules make a false claim or mandate an impossible action, then the problem solver can be said to have reached an *impasse* (Brown & VanLehn, 1980). Such impasses require repair procedures, such as when Larkin's (1983) physics experts changed their schemas when faced with a contradiction. In our terms this is movement in model space.

Success in rule space could also lead to movement in model space. While less likely to result in wholesale change to a model, success can lead to modification of the current model, such as through *elaboration*. Elaborations (see VanLehn, 1989) are assertions about the problem without having any impact on previous assertions. Simply filling in slot values in a schema is a form of elaboration, but so are new statements about the representation of the problem which may arise from the testing of rules through the generation of instances.

The hierarchical three-space model of problem solving is represented diagrammatically in Figure 1. In this model, the problem description provides the initial model state, which in turn defines the rules space consisting of all possible rules that the model suggests are plausible. The problem solver's state in rule space defines what are the relevant instances and how they should be represented in order to test rules. Instances are then generated by invoking experiments (i.e., interaction with the world) or from memory. The results of generating instances can be used to

modify rules, that is, cause movement in rule space (confirmation can be seen as a form of movement too in that the confidence in the rule state would be enhanced). Repeated failure for the rules in the rule space may lead to modification of the model, either directly (e.g., the failed rules may suggest different types of rules), or by evoking search mechanisms in model space in order to overcome the impasse. For each space, memory provides knowledge that is used by the search processes.

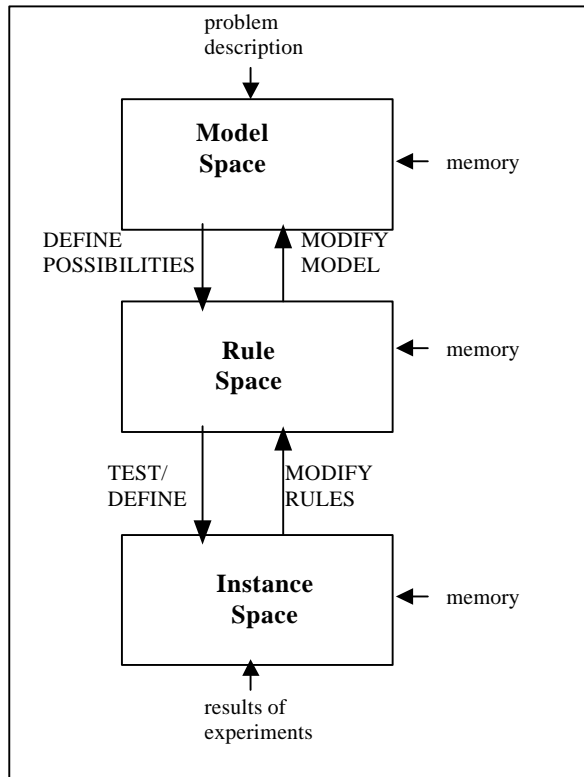


Figure 1. A hierarchical three-space theory of problem solving

We can illustrate these spaces using Vollmeyer et al's (1996) biology-lab task. Initially participants had to construct a model of what sort of system they thought they were faced with. Their model defines a rule space consisting of all possible links between inputs and outputs. If participants hit on the right model immediately, that biology-lab is a straightforward linear system, then they can solve it quickly. To test rules, instances are generated consisting of particular sets of inputs. Such a model defines a constrained, and thus quickly searchable, rule space. However, most participants start out with less precise models. For example, they have models that include the possibility of interactions, or random effects. A model including such possibility defines a larger rule space, and searching these parts of the rule space are at best a waste of time, and at worst confusing.

**Search Processes.** An advantage of treating problems solving as search of multiple spaces, is that it suggests a series of questions about the nature of the search processes.

For each space we have to ask Simon and Lea's (1974) questions: 1) what are the knowledge states, 2) what are the generative processes, 3) what are the test processes, 4) what are the selection processes?

Table 1 is a proposal for the nature of the search processes. Most of the processes invoked are processes already studied. For *instance space* the processes are those normally invoked for problem solving as search of a problem space, but Table 1 also specifies relationships between other processes, such as induction, hypothesis testing, metacognition, and analogy. Table 1 suggests a specific organization between different processes involved in problem solving and learning from problems solving. For *rule space*, induction and hypothesis testing are clearly distinguished as the generative and test processes respectively, and they are both distinguished from analogy. Table 1 also highlights processes we know little about, in particular, the selection processes for rule and model space.

It is clear from Table 1 that we understand least about model space. This is not surprising given that it encompasses the questions of "How do we form representations?" Table 1 implies though a useful way to think of research into analogy, a common topic in recent years (see Holyoak & Thagard, 1995). If analogies give people a new way of looking at a situation or problem (e.g., the water analogy for electricity) then they can be seen as a generative process in model space. Analogies of this type are therefore distinguished from induction.

Table 1: The four generalized problem space components for each of the three spaces.

	Instance space	Rule space	Model space
<b>knowledge states</b>	states of a task	hypothesized rules	possible models of the tasks
<b>generative processes</b>	operators for changing the state of the task	operators for generating rules (e.g., induction)	operators for generating new models (e.g., analogy)
<b>test processes</b>	evaluate how close current state of task is to its goal state	hypothesis testing (e.g., generate critical instance)	evaluate how well current model fits (e.g., metacognitive processes)
<b>selection processes</b>	select operator or evaluation method	decide which rule to test, or how to generate a rule	select method for evaluating or generating new model

**Formalising a Hierarchy of Spaces.** In Figure 1, the hierarchical nature of the three-space theory is made clear. We aimed to create a hierarchy because it makes the spaces clearly distinct. We agree with the proposed constraint of Baker and Dunbar (1996) that in multispace theories the spaces at different levels of abstraction (e.g., rule and instance space) should be isomorphic, whereas those at the same level (e.g., different representational forms of the same problem) should be homomorphic. Figure 1 presents the spaces as hierarchical, and we can describe them as

being hierarchical, but to truly impose this constraint we need to propose a formal definition that is hierarchical.

To give the spaces a formal definition, we start with the claim that any task can be seen as defined by a set of inputs, a set of outputs, and a set of rules relating those inputs to the outputs. Productions can have this form, so the generality of this claim is wide. In this formalism each output can be seen as a function of the inputs and constants associated with the inputs. Thus, a task with a set of  $X_1$  to  $X_M$  inputs and  $Y_1$  to  $Y_N$  outputs can be defined by the following set of very general functions:

$$\begin{aligned} Y_1 &= f_1(c_{10}, c_{11}, X_{11}, c_{12}, X_{12}, c_{13}, X_{13}, \dots, c_{1M}, X_{1M}) \\ Y_2 &= f_2(c_{20}, c_{21}, X_{21}, c_{22}, X_{22}, c_{23}, X_{23}, \dots, c_{2M}, X_{2M}) \\ &\dots \\ Y_N &= f_N(c_{N0}, c_{N1}, X_{N1}, c_{N2}, X_{N2}, c_{N3}, X_{N3}, \dots, c_{NM}, X_{NM}) \end{aligned}$$

The relationship between different hierarchical spaces can be specified in terms of the different components of these equations that a state in each space will specify. A *model state* specifies a set of functions with constants left unspecified; a *rule state* specifies a set of constants; whereas a particular set of  $X$  values (with resulting  $Y$ 's) represent *instance states*. For example, consider a task that could be described as a single output with two inputs. This would be defined by a single equation:  $Y = f(c_0, c_1, X_1, c_2, X_2)$ . A model suggesting that inputs are additive specifies the equation  $Y = c_0 + c_1X_1 + c_2X_2$ . The rule that " $X_1$  has twice the effect of  $X_2$  but there is no constant effect," is expressed by the equation:  $Y = 2X_1 + X_2$ . This hypothesized rule could be tested by generating an instance with values of 5 for  $X_1$  and 5 for  $X_2$  and testing if the resulting value of  $Y$  is 15. Biology-lab fits easily into this framework as  $X$  values can be seen as changes to inputs, constants define particular possible links, and the shape of the functions are the nature of possible rules. However our argument is that any task could be seen in these terms, so applying the hierarchy constraint when determining the exact nature of the spaces for a task can be seen as requiring a specification of how the task fits into this formalism. The mathematics of this formalism are not in themselves insightful, but fitting spaces to this formalism creates constraints on the definitions of the search spaces.

## Comparison to Other Approaches

**Other Multispace Models.** Ours is not the only work on multispace models. Another is the four-space model of Schunn and Klahr (1996) for scientific discovery. This model differs in various ways from the hierarchical three-space theory, but there is not space here to fully explore the differences. An important difference is that the four-spaces are not constrained to be hierarchical. The scope of the four-space model is not clear, but if it can be a general model of problem solving, then we would welcome it as another attempt to address the lack of theory in problem solving research. Dialectic progress requires competing alternatives.

How do multispace models in general relate to Soar (Newell, Rosenbloom, & Laird, 1989) and ACT-R (Anderson, 1993)? Soar and ACT-R are frameworks in

which detailed models of problem solving can be built. Because Soar constructs a new problem space whenever the need arises, Newell (1989) proposed that Soar could model Klahr and Dunbar's (1988) theory, so by extension it can model all multispace theories. Whereas it should be possible to build multispace models in the Soar architecture, they are not equivalent just because they both involve multiple problem spaces. The spaces in multispace models are conceptually distinct and interact in specified ways, so a compatible model built in Soar would have to incorporate these assumptions.

Anderson's (1993) ACT-R does not explicitly incorporate the idea of interactive search of multiple spaces, but there appears to be no reason why it could not model such processes. The current goal in ACT-R is critical, because subgoals encourage the firing of certain sets of productions. Such sets could be considered to define different spaces, so perhaps rapid transition between different subgoals could simulate an interactive search between spaces. The implications of such an approach are unclear.

**Situated Cognition.** We started by decrying the lack of alternatives theories in problem solving research, but there exists an approach to problem solving that does not focus on search: situated cognition. Situated cognition places a great emphasis on the context of cognition and denies (or at least de-emphasizes) that symbolic processing (such as search of a problem space) lies at the heart of cognition. The extent to which situated cognition is an antithesis to problems solving as search, is not clear. Vera and Simon (1993) tried to place situated cognition into the symbolic framework, but the replies to their article suggested that researchers taking the situated cognition approach see it as fundamentally different. However the problem with situated cognition emerging as an antithesis to the thesis of problem solving as search may be that neither the thesis nor the antithesis is clear enough to begin with.

Like any clearly stated antithesis we would welcome the emergence of a competitor such as situation cognition. Within the three-space model, in general we could try to explain the phenomena that cognition is often heavily context dependent as the claim that movement in the model space is difficult, and may usually define only a restricted rule space. Perhaps this is the general condition, and the implications of this would have to be worked out.

## Conclusions

We have argued that a hierarchical three-space theory of problem solving can be derived from existing studies and ideas about problem solving. In constructing this theory we have been guided by Schunn and Klahr's (1996) three criteria for when to propose additional problem spaces in multispace search theories. The first criterion is logical, do the spaces involve search of different goals and entities? (We would also add, do they use different operators for search?) The three spaces we propose clearly involve different kinds of states, goals, and ways of searching that space, so we think we meet this criterion. The second criterion was, do the spaces differ empirically? There is

evidence from existing literature that different factors influence behavior, so we think we can meet this criterion. Schunn and Klahr's third criterion was implementational, spaces should be able to be represented distinctly in a computational model that can perform the task. At the moment we can do more than suggest how such a model using the our theory would work, but constructing such a model is an important aim.

To test the hierarchical three-space theory we intend to examine the testable implications it has for how people may best learn from encountering novel tasks. It suggests that whether hypothesis testing will be a good strategy for learners depends on the quality of the learner's model. Learners with a poor model may be disadvantaged by being encouraged to test hypotheses. A current weakness with the theory is that we may be able to define relatively what are good and poor models in terms of some metric of the size of the rule space the model state defines, but it may be hard to define absolute model goodness. Specifying the distinction between good and poor models precisely is an important aim of future research, especially if we are to investigate the practical implications of the theory. Also required is further study into the reality and properties of the links between spaces proposed by the theory.

Have we met our aim of proposing a theory for problem solving? We are trying to develop the hierarchical three-space theory so that it can generate predictions in terms of the interactions between different spaces, and hope to make the theory a tool for organizing the different processes involved in problem solving. However, we recognize that the theory requires more development, both computationally and empirically, before it is truly more than a framework. Such attempts by problem solving researchers are necessary though because until there are such theories, problem solving will remain just a set of diverse and sometimes unrelated phenomena.

## References

- Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R. (2000). *Cognitive psychology and its implications* (5th ed.). New York: Worth.
- Baker, L. M., & Dunbar, K. (1996, July). *Problem spaces in real-world science: What are they and how do scientists search them?* Paper presented at the Eighteenth Annual Conference of the Cognitive Science Society, San Diego.
- Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistics ideas. *Cognitive Psychology*, 2, 331-350.
- Brown, J. S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science*, 4, 379-426.
- Burns, B. D. (1996). Meta-analogical transfer: Transfer between episodes of analogical reasoning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22, 1032-1048.
- Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, 17, 397-434.
- Hayes, J. R., & Simon, H. A. (1974). Understanding written problem instructions. In L. W. Gregg (Ed.), *Knowledge and cognition*. Hillsdale, NJ: Erlbaum.
- Hayes, J. R., & Simon, H. A. (1977). Psychological differences among problem isomorphs. In N. J. Castellan, D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive Theory*, vol. 2. Hillsdale, NJ: Erlbaum.
- Hegel, G. W. F. (1807/1931). *The phenomenology of mind* (J.B. Baille, Trans.). London: Allen & Unwin.
- Hofstadter, D. R. (1995). *Fluid concepts and creative analogies*. New York: Basic Books.
- Holyoak, K. J., & Thagard, P. (1995). *Mental leaps*. Cambridge, MA: MIT Press.
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, 12, 1-55.
- Larkin, J. H. (1983). The role of problem representation in physics. In D. Gentner & A. L. Stevens (Eds.), *Mental Models*. Hillsdale, NJ: Erlbaum.
- Maier, M. R. F. (1930). Reasoning in humans. I. On direction. *Journal of Comparative Psychology*, 10, 115-143.
- Maier, N. R. F. (1940). The behavior mechanisms concerned with problem solving. *Psychological Review*, 47, 43-58.
- Newell, A. (1989). How it all got put together. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing*. Hillsdale, NJ: Erlbaum.
- Newell, A., Rosenbloom, P. S., & Laird, J. E. (1989). Symbolic architectures for cognition. In M. I. Posner (Ed.), *Foundations of cognitive science*. Cambridge, MA: MIT Press.
- Newell, A., & Simon, H. A., (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Schunn, C. D., & Klahr, D. (1996). The problem of problem spaces: When and how to go beyond a 2-space model of scientific discovery. In G. W. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 25-26). Hillsdale, NJ: Erlbaum.
- Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In L. W. Gregg (Ed.), *Knowledge and cognition*. Hillsdale, NJ: Erlbaum.
- Sternberg, R. J. (1995). *In search of the human mind*. Orlando: Harcourt Brace.
- VanLehn, K. (1989). Problem solving and cognitive skill. In M. I. Posner (Ed.), *Foundations of cognitive science*. Cambridge, MA: MIT Press.
- Vera, A., & Simon, H. A. (1993). Situated action: A symbolic interpretation. *Cognitive Science*, 17, 7-48.
- Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity and systematicity of strategies on the acquisition of problem structure. *Cognitive Science*, 20, 75 - 100.
- Weisberg, R. W., & Alba, J. W. (1980). An examination of the alleged role of 'fixation' in the solution of several 'insight' problems. *Journal of Experimental Psychology: General*, 110, 169-192.

# The Representational Effect in Complex Systems: A Distributed Representation Approach

**Johnny Chuah** (chuah.5@osu.edu)

The Ohio State University  
204 Lazenby Hall, 1827 Neil Avenue, Columbus, OH 43210, USA

**Jiajie Zhang** (jiajie.zhang@uth.tmc.edu)

**Todd R. Johnson** (todd.r.johnson@uth.tmc.edu)

University of Texas at Houston  
7000 Fannin, Suite 600, Houston, TX 77030, USA

## Abstract

The representational effect refers to the phenomenon that different isomorphic representations of a common structure can generate dramatically different representational efficiencies, task difficulties, and behavioral outcomes. This paper presents a study of applying distributed representations to systematically analyze the representational effect in complex real world systems. Distributed representation is a representational system that is composed of internal and external information that is processed in a dynamic, interactive, and interwoven manner. The representational effect is observed and studied in a series of experiments involving various navigational instruments used in aviation. The cognitive task was decomposed into its components and the information distribution across internal and external representations for each component was identified. The experimental results showed that the task difficulties of different instruments correlate positively with the amount of external information for the component tasks, as predicted by the distributed representation analysis.

## Introduction

The information necessary for the performance of almost any everyday task is distributed across information perceived from the external world and information retrieved from the internal mind. These tasks are known as distributed cognitive tasks (Zhang & Norman, 1994). The external representations constructed from the information extracted from external objects (such as written symbols) and the internal representations in the mind (such as schemas) dynamically integrate and interweave to result in a rich pattern of cognitive behavior. The principle of distributed representations is that a distributed cognitive task involves a system of distributed representations that consists of internal and external representations (Zhang & Norman, 1994, 1995). The task is neither exclusively dependent on internally nor exclusively dependent on externally processed information, but rather on the interaction of the two information spaces formed by the internal and external representations.

In the aviation industry, there are a wide variety of navigational systems. Among them there is a set of very basic navigational instruments. These instruments are selectively tuned to transmitting radio stations on the ground. The received signals are then presented in a display in the cockpit for the pilot to interpret. There is only so much information that a navigation instrument needs to display: azimuth or directional information, and distance information. However, different instruments present these two pieces of information differently and result in different degrees of precision and efficiency as interpreted by the pilot.

Cockpit information displays are examples of distributed representation systems. Navigational information in a cockpit information system can and is represented through a variety of isomorphic navigation instruments. Although these instruments are isomorphic and provide the same information, they vary in their relative degrees of directness and efficiency in their representation of scale information (Narens, 1981; Stevens, 1946; Zhang, 1995). The scale information of the orientation and distance dimensions in a cockpit information display is represented across internal and external representations and can dramatically affect the representational efficiency of the display and the navigator's behavior (Zhang, 1997). This research examines the cognitive properties of the representations that such instruments produce. The specific assumption to be tested is that with the most direct system, scale information is maximally represented externally, resulting in higher efficiency, faster and more direct responses.

## Distributed representations

External representations are the representations formed from information gathered from the external environment. External representations include physical objects and/or symbols, relations and constraints between physical objects and their configurations relative to each other, and external physical rules, such as laws of physics. Through the human perceptual processes, the information necessary to form external representations is picked up by the sensory and



perceptual systems. External representations are characterized as providing information that is directly perceived and applied toward a cognitive task without being explicitly interpreted. External representations contribute information that is otherwise unavailable from representations internally generated from memory, or from representations that are internalized from perceptual information (Zhang, 1997). Perceived information from within the external environment that must be represented internally in order for cognition to operate on it is, by definition, recreated as internal representations.

Internal representations are the representations that originate from within the mind and are not initiated from the perception of external stimuli. These internal representations are in the form of, but not limited to, mental images, propositions, production rules, and schemata. Cognitive processes retrieve information from long-term memory. This information may be selectively or incidentally retrieved, and is then employed to formulate internal representations.

Internal and external representational spaces together form a distributed representational space, which is where the representation of the task (its abstract structures and properties) resides. External representations are not represented redundantly as internal representations. In combination with internal representations, external representations can directly activate and provide perceptual information necessary for responses and actions.

### Representational effect

The representational effect is the phenomenon that different isomorphic representations of a common structure can generate dramatically different representational efficiencies, task difficulties, and behavioral outcomes (Zhang & Norman, 1994). It is ubiquitous in problem solving, reasoning, and decision making across many task domains.

### Navigational Displays

The cockpit informational displays in this experimental study are navigational instruments that provide directional guidance. As the experimental task is a position-fixing task, only the instruments that have the

necessary information were provided and will be discussed here briefly. (A more in-depth review of cockpit navigational displays is provided in Zhang, 1997.). VOR (very high frequency omnidirectional range), ADF (automatic direction finder), RMI (radio magnetic indicator), and the Moving Map display are four of the more prevalent navigation systems used for such a position fixing task. The generic moving map display refers to the more advanced cathode ray tube displays found in newer airlines that provide multiple information sources over a moving map within one display.

#### VOR indicator

The VOR equipment in the aircraft receives and interprets transmitted radio signals from the ground and shows directional information of the aircraft in relation to the VOR station on the ground. The VOR indicator is usually used to show the intended course of the aircraft and the lateral position of the aircraft in relation to that intended course. The VOR indicator in Figure 1A shows a selected 315° course. The TO indication at the right of center of the display indicates that proceeding on such a course will lead the aircraft to the station. The vertical needle (CDI, course deviation indicator), when in the center as shown, indicates the aircraft is on that selected course. If the CDI pivots to the left, this will indicate to the navigator that the aircraft is off the 315° course and needs to make a correction by navigating the aircraft towards the left to get back on course. The VOR indication (course selected) is independent of the heading of the aircraft.

The VOR indicator can also be used to determine the location of the aircraft relative to the VOR station. By tuning the VOR until the CDI centers with a TO indication, the displayed course will be the magnetic bearing of the aircraft to the VOR station. Likewise, by tuning the VOR until the CDI centers with a FROM indication, the displayed course will be the magnetic bearing of the aircraft from the VOR station.

#### ADF indicator

The ADF indicator in the aircraft can also be used for directional guidance to or from the radio station, or position fixing to determine one's location. The ADF indicator

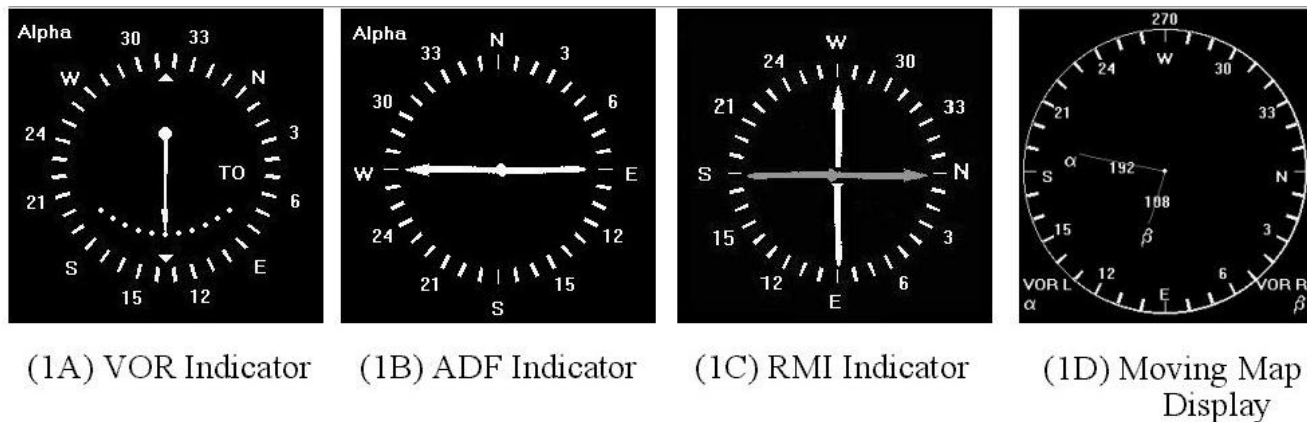


Figure 1: The four navigation instrument displays.

shows the orientation of an aircraft relative to the radio station (see Figure 1B). It only displays the relative bearing of the aircraft to the station, which is the angular distance between the lateral axis of the aircraft and the course to the station. In order to obtain a magnetic indication, which is necessary to navigate or determine one's position, the relative bearing indication must be summed with the magnetic heading of the aircraft (obtained off another instrument not shown). This sum is the magnetic bearing to the station; in order to derive the magnetic bearing from the station, the pilot would need to determine the reciprocal.

### RMI indicator

The RMI indicator is similar in its display to that of the ADF indicator. The major difference between the ADF and RMI indicators is that the ADF display is fixed and the RMI display rotates as the aircraft changes direction. The RMI display is essentially the aircraft's heading indicator with the RMI pointer(s) providing navigational information (see Figure 1C). As a consequence, the RMI provides angular distance, and orientation of the aircraft relative to the radio station as magnetic indications. It is unnecessary for the navigator to do any computations to obtain magnetic bearing information.

### Moving Map display

The primary navigational display mode of a Moving Map display shows a map of the immediate surrounding environment of the aircraft, as well as the radio stations. Magnetic bearing information is displayed alongside lines extending from the center to the radio stations. Angular distance is also provided.

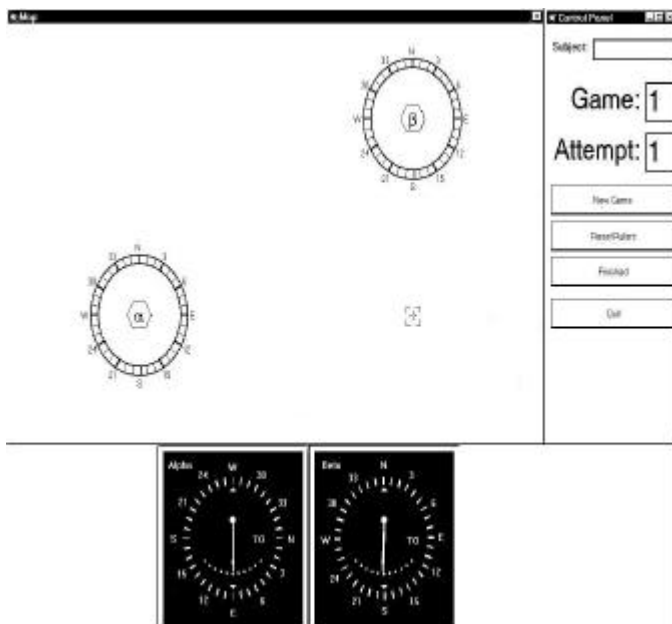


Figure 2: The experimental task display.

## Experimental Study

The experimental hypothesis is that although the four navigation instruments provide the same and all the necessary information, the different distributions of the same information across internal and external representations make some instruments harder to use while make others easier to use, with the easiest one the instrument that has most external information. Experimental participants were provided with bearing information as displayed by the instruments, and were then required to determine the current position of the aircraft on a map.

### Representational study of experimental task

A representational analysis of the experimental task identifies the abstract structures of the task and the representational properties that are responsible for the representational effect. To successfully perform the position-fixing task with the given bearing information, it is necessary to perform a triangulation using the radio stations as end points and extending from them along the bearings. The intersection of the bearings indicates the current position of the aircraft relative to the radio stations.

The four types of instruments have different representational spaces. The representational system with the largest amount of external information will be more efficient and direct (Norman, 1993; Hutchins, 1995; Zhang & Norman, 1994). Furthermore, the position-fixing task requires a triangulation method to determine the aircraft position. Both the VOR and RMI provide the necessary magnetic bearing information immediately. It is not necessary to represent the information internally. The ADF does not provide the information readily, and it is necessary to derive the magnetic bearing information through mental calculations using the heading information with the relative bearing information provided by the instrument. With the moving map display, the magnetic bearing information is also readily available. Furthermore, the information is presented in a graphical and spatial layout with the instrument displaying the position of the aircraft relative to the radio stations. There is little effort required in comparing the displayed spatial relations with the map and determining the aircraft position. The other three instruments require cognitive effort in subtending bearing lines extending from the radio stations to create an intersection in order to determine aircraft position.

Table 1: Properties of the navigation systems.

Information readily available (externally represented)	Type of navigation system			
	VOR	ADF	RMI	Mov. Map
Aircraft heading			√	√
Magnetic bearing	√		√	√
Orientation		√	√	√
Angular distance between aircraft and radio station		√	√	√
Spatial and graphical layout of information				√

The prediction is that the moving map display will outperform the other three instruments because it provides the largest amount of external information and graphically and spatially presents the information in such a manner where the operation of determining the location of the aircraft is also provided externally. The representational effect will be that the experiment participants within that navigation instrument condition will outperform the other navigation instrument conditions. Table 1 summarizes the properties of the four navigation systems. For the other instruments, RMI should be easier than ADF, which should be in turn easier than VOR.

### Method

**Subjects.** Eighty-five participants participated in the experiment for course credits in an introductory psychology course at The Ohio State University.

**Materials and Equipment.** Three Pentium computers were used with 17-inch monitors set at similar SVGA resolutions. The displayed image consisted of a large map covering most of the screen area, an instrument panel with the navigation instruments unique to each experimental condition, and a control panel that served as the experiment

interface. The map area displayed two radio stations and a square icon that represented the aircraft. The positions of the radio stations and aircraft were randomized at every trial. Figure 2 shows a screen capture of an experimental trial.

**Design and Procedure.** The experiment was a between-subject design with four conditions, one navigational instrument for each condition. Each participant had 24 trials in the experiment. For each trial, the navigation instruments were displayed, providing the necessary and essential information. The participants would then read and interpret the navigation information and, by clicking and dragging the square aircraft icon, re-position it to where they believed the actual position of the aircraft was. They would commit their decision by clicking on the OK button. If the participants were correct to within a radius of 5% of the screen diagonal dimension, they moved on to a new trial. If they were incorrect, they were given two more attempts to locate the position.

Due to the complexity of the experimental task, the instructions were carefully administered, which limited the number of participants for each experimental session to two. Participants were first given a set of written instructions, then the experimenter provided with verbal explanations and further instructions. Each participant was

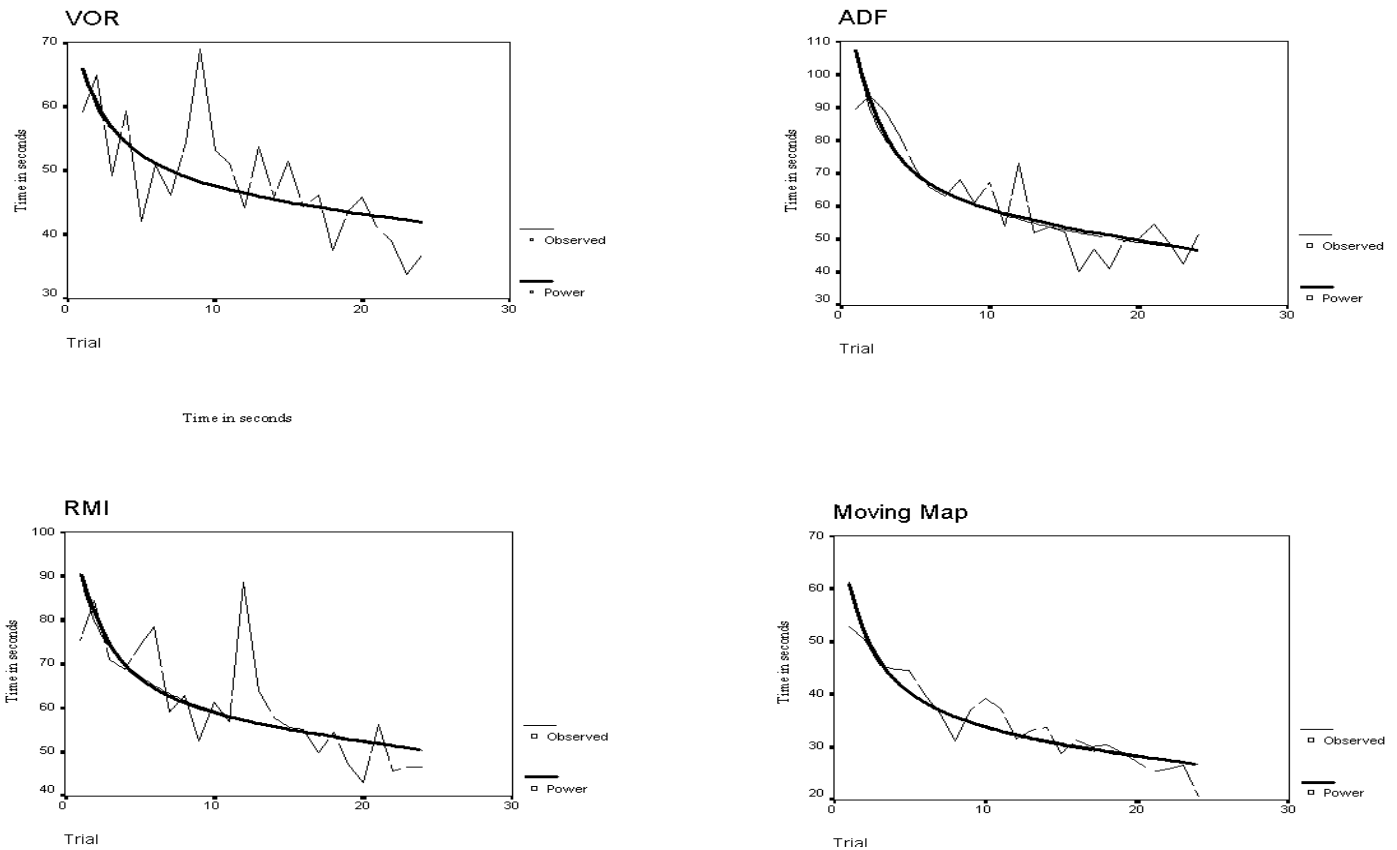


Figure 3: Performance data regressions.

given at least 3 trials to demonstrate his/her comprehension of the task prior to the start of the experiment. As the experiment required some manual dexterity to maneuver the computer pointer over the monitor screen, the computer mouse was configured for lefthanded participants when necessary.

### Results and Discussion

The performance data from all subjects were averaged by trial within each of the four conditions. Regressions were then performed for each condition.

There was an observation of a dramatic and robust power curve of learning for each condition that corresponded with the standard power law of practice. Figure 3 shows each condition with its raw averaged data and its best-fit regression. The variation in the VOR condition is the largest, as indicated by the lowest

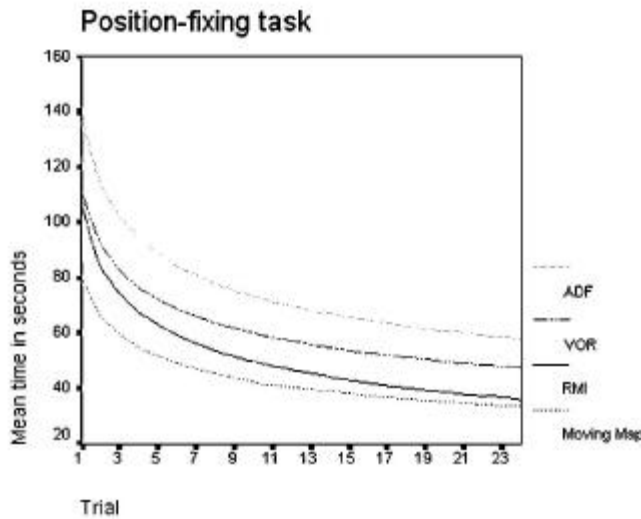


Figure 4: Regression comparison.

regression fit ( $r$ -squared value = 0.44). The other conditions have higher fits to raw data ( $r$ -squared values: ADF = 0.78, RMI = 0.57, Moving Map = 0.84).

Figure 4 shows all four regressions within one graph for easy comparison. Table 2 below summarizes the task completion times for a trial set that was used for analyses.

Table 2: Comparison of mean times.

	VOR	ADF	RMI	Mov. Map	sig.
Trial 1	92.6	104.7	78.1	62.3	0.04
Trial 6	68.1	82.1	64.1	51.1	0.05
Trial 12	54.2	99.3	60.9	40.4	0.01
Trial 18	41.8	52.9	44.9	43.0	0.60
Trial 24	42.7	63.4	41.2	28.0	0.00

The prediction was for participants in the Moving Map condition to outperform all other conditions, to be followed with the RMI and ADF conditions, and the VOR condition

participants were expected to be the slowest to complete the experimental task.

As the data show, there were significant differences between the four conditions at four of the five trials used for analyses in the ANOVA test. A posthoc analysis performed using the Tukey test for multiple comparisons revealed that there were significant differences in task completion times ( $\alpha$  level = 0.05) between the ADF and RMI conditions and between ADF and Moving Map conditions, with the ADF times higher than either of the other two conditions.

Performances levels of the four conditions got closer after 24 trials, although there was a significant difference between the ADF and Moving Map conditions. The individual power curves of learning between all four conditions resulted in this performance convergence. After 24 trials, the performance times between the conditions followed the predicted trend.

### Discussion

According to the hypothesis, the representational effect observed will favor the performance of the Moving Map over that of the RMI, ADF, and VOR. This assumption arose from the representational analysis that decomposed the cognitive task and identified the components and properties that would be responsible for such a representational effect. It was identified that the Moving Map navigation display provides all the necessary information externally and in a spatial and graphical layout and other displays provide more information that needs to be represented and computed internally, with a high cognitive cost. All the necessary information for the task is available as directly perceptible forms of external representations for the Moving Map condition. Furthermore, the information is provided in an instrument display that maps directly to the map displayed on the monitor since the instrument itself is a map. None of the instrument components has to be represented or re-represented in an internal representation, thus reducing mental workload and increasing task efficiency.

The RMI condition posted consistently faster times against the ADF condition, with significance for trials 12 and 24. The RMI displays bearing information to the user in the magnetic compass scale, as opposed to the ADF instrument that provides the information in a relative degree scale. As a result, the navigator avoids costly mental workload by obtaining more of the information from the external representational space.

The VOR task completion times were not expected to be as fast as the Moving Map display. It was anticipated that VOR times would be slightly slower than ADF times. But there were no significant differences between the VOR and ADF. The difference, if it existed, might be too small to be observed. Additionally, there is an obvious and noticeable learning process that is occurring, as the participants become more proficient and familiar with the instruments and the task itself. This may be attributed to simple skill acquisition or familiarization of the interface.

## Conclusion

The experimental results were generally consistent with the predictions of the distributed representation analysis. The prediction was that the instrument with more external information would be easier. This prediction was supported by the observed representational effect. The representational effect predicted that isomorphic representations could produce different behaviors due to the variant distributions of internal and external representational information.

The resulting behavior variance from the experiment indicates that some representations are more 'efficient' in extending the necessary information for a task. Although the different isomorphic representations result in different initial levels of performance and learning curves, performances appear to converge after a sufficient period of learning.

One argument can be made about the learning behavior: learning and practice may eliminate the representational effect after enough trials. However, further research needs to be done in more complex and dynamic settings. The current experimental task was a simple position-fixing task in a very controlled and static environment. In an unpredictable and complex environment such as that of the cockpit of an aircraft, the representational effect could be more pronounced and a possible regression to initial performance levels should be studied. Another issue that is worth of further study is whether the converged performance after learning for different representations will diverge again under extreme conditions such as high cognitive workload and time pressure.

## Acknowledgements

This research was supported by Grant N00014-96-1-0472 from the Office of Naval Research and a 1997 Summer Fellowship Research Award from the Center for Cognitive Science at The Ohio State University

## References

- Hutchins, E., (1995). How a cockpit remembers its speed. *Cognition Science*, 19, 265-288.
- Narens, L., (1981). On the scales of measurement. *Journal of Mathematical Psychology*, 24, 249-275.
- Norman, D. A., (1993). *Things that makes us smart*. Reading, MA: Addison-Wesley
- Stevens, S. S., (1946). On the theory of scales of measurement. *Science*, 103, 677-680.
- Zhang, J., & Norman, D. A. (1994). Representations in Distributed Cognitive Tasks. *Cognition Science*, 18, 87-122.
- Zhang, J., & Norman, D. A. (1995). A representational analysis of numeration systems. *Cognition*, 57, 271-295.
- Zhang, J., (1997). Distributed representation as a principle for the analysis of cockpit information displays. *International Journal of Aviation Psychology*, 7(2), 105-121.

# Precursors to Number: Making the Most of Continuous Amount

Peter Drake

pedrake@cs.indiana.edu

Department of Computer Science, Indiana University, Bloomington; Lindley Hall 215  
Bloomington, IN 47405 USA

Kelly Mix

kmix@indiana.edu

Department of Psychology, Indiana University, Bloomington; 1101 E. 10th Street  
Bloomington, IN 47405 USA

Melissa Clearfield

mclearfi@indiana.edu

Department of Psychology, Indiana University, Bloomington; 1101 E. 10th Street  
Bloomington, IN 47405 USA

May 6, 2000

## Abstract

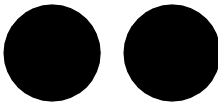
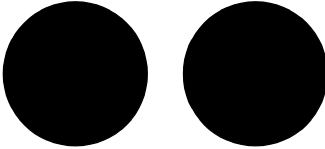


How does our understanding of number develop? There is evidence suggesting that even infants have primitive concepts of “more”, “less”, and “the same”. Some researchers have concluded that humans have an innate number sense, present from birth. In this paper, we present a two-part model which explains these results in terms of continuous amount. The first part is a quantitative model addressing the results of infant habituation studies. The second, more tentative part of the model addresses object individuation, subitizing, and number estimation.

## Amount vs Number

In this paper, we use the word *amount* to refer to the total area of the objects in view; this is a continuous quantity. *Number*, a discrete quantity, refers to how many objects are present. As shown in Table 1, these two aspects can be varied independently.

A complete model would have to take into account other features, such as total contour length (edge length), shape, and color. Except where otherwise stated, we disregard these details.

Table 1: Amount vs Number. Both pictures in each column have the same total area.

	Small Amount	Large Amount
Small Number (2)		
Large Number (3)		

## Habituation Studies on Infant Numerical Abilities

Three studies are addressed directly by the first part of our model: Starkey & Cooper (1980), Antell & Keating (1983), and Clearfield & Mix (1999). All three studies use the same habituation paradigm, described below.

An infant is shown a series of images of black circles or squares on a white background, such as those in Table 1.

The infant is shown several more images. They may differ in arrangement, but they are the same on some critical dimension, such as the number of dots. If the infant habituates (stops looking at new images as long), this is taken as evidence that the infant detected the invariant property and became bored with it.

After habituation, the infant is shown a test image which differs on the critical dimension. If the infant dishabituates

(spends significantly more time looking at the test image), he or she presumably noticed that the property changed—the test image is new and exciting. If the infant does not dishabituate, he or she presumably did not notice anything special about the test image.

In Starkey & Cooper’s study, 22-week-old infants who had been habituated on images of 2 dots dishabituated when tested on images of 3 dots (and vice versa). Infants, it appeared, can tell the difference between 2 and 3.

To discount the possibility that the infants were simply reacting to amount, Starkey & Cooper tried 4 vs 6 dots. The relative difference between the images in this condition is the same as in the 2 vs 3 condition: the larger number has 1.5 times as much area as the smaller one. If infants are using amount of area, they should be at least as likely to dishabituate in the 4 vs 6 condition.

On the other hand, older children and adults have a much easier time enumerating sets of up to 3 or 4 objects than larger sets. Enumerating large sets requires use of an explicit, learned counting procedure; smaller sets can be enumerated quickly and subconsciously, through a process called “subitizing”. The nature of subitizing remains controversial. In any case, if infants are subitizing, the 2 vs 3 condition should be more likely to produce dishabituation.

In fact, Starkey & Cooper’s subjects did *not* dishabituate in the 4 vs 6 condition. It was therefore suggested that subitizing may be innate.

Antell & Keating replicated these results in newborns (less than a week old).

Clearfield & Mix’s study reexamined the amount hypothesis. By changing the sizes of the objects (squares) in their displays, they were able to independently vary the number of objects and the total amount of area and contour length. (This study used 6- to 8-month-old infants, and only the 2 vs 3 condition was considered.)

They found that if the number remained the same, but the total area and contour length changed significantly, the infants dishabituated. Moreover, the infants *did not* dishabituate if the test image had approximately the same total area and contour length as the habituation image, *even if the number of objects was different*. In other words, infants do not appear to distinguish between 2 large objects and 3 small ones.

One problem remains: if infants are using amount to discriminate quantities, why don’t they dishabituate in the 4 vs 6 condition?

Clearfield & Mix proposed that the 4 vs 6 displays might simply contain too much visual complexity. There is no question of comparison; the infants are overwhelmed by the displays.

The current model focuses instead on the way amount is represented internally<sup>1</sup> by the infants.

If the perceived magnitude grows linearly with the actual amount, then the 4 vs 6 condition should be more likely to produce dishabituation than the 2 vs 3 condition, because the

absolute difference in amount of stuff is larger in the former condition.

Fechner (1860) suggested that the perceived amount grows as the logarithm of the actual amount. This almost does the trick, but not quite: whenever ratios are the same (as in the 2 vs 3 and 4 vs 6 conditions), differences of logarithms are equal. This would make the two conditions equally likely to produce dishabituation.

To explain the greater difficulty of the 4 vs 6 condition, we need a perception function which grows *more slowly* than a logarithm. One such function is the sigmoidal “squashing” activation function commonly used in connectionist “neural network” models (Rumelhart, McClelland, et al., 1986).

### A Quantitative Model

We presume that the photoreceptors (rods and cones) in the retina provide (indirect) input to a neuron (or, more likely, a group of neurons) which codes for the total area of the objects in view. If the total area is larger, this *area unit* is more active; if the area is smaller, it is less active.

This does not require that the image be individuated into objects or preprocessed in any other way. The area unit is actually recording the total amount of light (or lack thereof) received by the retina. For the studies mentioned in the previous section, which use simple, black objects on a white background, this is equivalent to the total area.

The activity of the area unit does not vary linearly with its input. Instead, it varies according to a function of the form:

$$f(x) = \frac{1}{1 + e^{-\alpha x}}$$

In this equation,  $x$  is the total input to the area unit, and  $\alpha$  is a parameter of the model.

The consequence of all this is that the same difference is perceived as being smaller if the absolute amounts involved are larger. This is the well-known magnitude effect, and is shown in Figure 1.

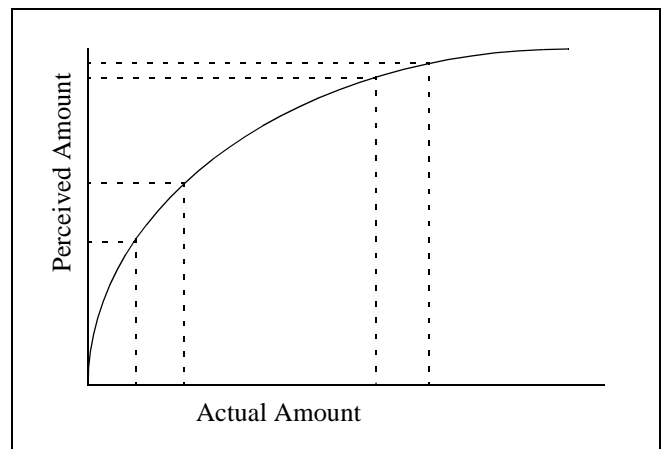


Figure 1: The magnitude effect as a result of a squashing function. Differences are perceived as smaller if the absolute magnitudes involved are larger.

<sup>1</sup> By “represented internally”, we mean “represented as a pattern of neural activation”. We do not mean to imply that infants are deliberately manipulating abstract, symbolic representations of amount.

In the context of the infant studies, our model predicts that the infant dishabituates if the perceived difference in area (between the habituation and test images) exceeds some threshold.

Just as the total activity of photoreceptors reflects the total visible area, the total activity of center-surround “edge detector” retina cells (ganglion cells) reflects the total visible contour length. A *contour unit*, analogous to the area unit, takes input from these neurons. A sufficiently large perceived difference here also induces dishabituation.

In order to make predictions, the model must be defined formally. Let  $AREA_{hab}$  be the total area (in square radians of visual angle) in the habituation images, and  $AREA_{test}$  be the total area in the test image. Similarly, let  $CONT_{hab}$  be the total contour length (in radians) in the habituation image, and  $CONT_{test}$  be the total contour length in the test image.

The model predicts that the infant will dishabituate if and only if

$$|f(AREA_{test}) - f(AREA_{hab})| > \tau_{AREA}$$

OR

$$|g(CONT_{test}) - g(CONT_{hab})| > \tau_{CONT}$$

where

$$f(x) = \frac{1}{1 + e^{-\alpha_{AREA}x}}$$

is the activation function for the area unit and

$$g(x) = \frac{1}{1 + e^{-\alpha_{CONT}x}}$$

is the activation function for the contour unit.

The model has four parameters: the thresholds  $\tau_{AREA}$  and  $\tau_{CONT}$  and the activation function sharpnesses  $\alpha_{AREA}$  and  $\alpha_{CONT}$ .

### Setting the Parameters

We begin by calculating the total area and contour length of the stimuli from each study. The distance from the infants to the screen was different in each study, so we first convert all lengths into radians of visual angle. The data are shown in Table 2.

Table 2: Data from habituation studies. Areas are in square radians of visual angle; contour lengths are in radians. The test image is assumed to be larger (in terms of area and/or contour length) than the habituation image; because of the absolute value in the formula, the model would make identical predictions for the converse condition.

Study	Dishabituation Condition				No Dishabituation Condition			
	Habituation		Test		Habituation		Test	
	Area	Contour	Area	Contour	Area	Contour	Area	Contour
Starkey & Cooper	0.00044	0.10	0.00065	0.16	0.00087	0.21	0.0013	0.31
Antell & Keating	0.0031	0.28	0.0046	0.42	0.0062	0.56	0.0093	0.84
Clearfield & Mix	0.0089	0.53	0.020	0.80	0.0089	0.53	0.0059	0.53

We were able to find parameters for which the model gives the correct predictions for all of these studies. In other words, at least one of the thresholds is exceeded in each dishabituation condition, and neither are exceeded in any no-dishabituation condition.

One satisfactory set of parameters is:

$$\alpha_{AREA} = 3000$$

$$\tau_{AREA} = 0.075$$

$$\alpha_{CONT} = 3.5$$

$$\tau_{CONT} = 0.075$$

Admittedly, there is some degree of coincidence involved

in our being able to find parameters consistent with all three studies. Variables such as lighting level, size of the card on which the object appear, and subject age may affect these values. Still, it is satisfying that none of the studies disagree qualitatively with the model, and intriguing that the same value can be used for both thresholds.

### Problems With the Quantitative Model

In addition to the variables just mentioned, we have some reservations about the quantitative model.

The model assumes that each image is registered in a single eye fixation. In fact, infants move their eyes around quite a bit while looking at an image. There are two ways this may not matter. First, if the infant is keeping a running average of the area and contour length in the image, minor eye movements should have little effect. Second, the infant



may be building an internal map of the image, and then extracting area and contour information from this “mind’s eye” view.

Another problem arises from studies on visual complexity. Karmel (1969) has given evidence that infants prefer to look at pictures with a certain amount (varying with age) of contour length. The total contour lengths in question are so huge (tens of radians) that, after passing through the activation function in our model, they would be indistinguishable. If, as our model predicts, these images are indistinguishable, how could infants have a preference? Karmel’s data come from a significantly different paradigm, and additional factors such as visual frequency may be coming into play. Still, these results will eventually have to be addressed.

### Estimation and Subitizing

The model presented thus far may go a long way to explaining infant habituation data, but it can’t be the whole story for adults. In addition to the explicit, sequential counting procedure, we are able to estimate number. This estimation ability appears to operate in parallel—unlike counting, it doesn’t take twice as long when the number of objects is twice as large. The magnitude effect appears here, too: estimation of large numbers is less accurate. Estimation is only precise within the subitizing range, up to 3 or 4 objects.

Before number can even be estimated, it is necessary to individuate objects. In this section, we propose a model of object individuation which can underlie the estimation ability.

### Temporal Synchrony

Animal studies (Eckhorn et al, 1988) have suggested an intriguing hypothesis about the visual system: in certain parts of the brain, cells which are responding to the same object fire at the same time. This is shown in Figure 2.

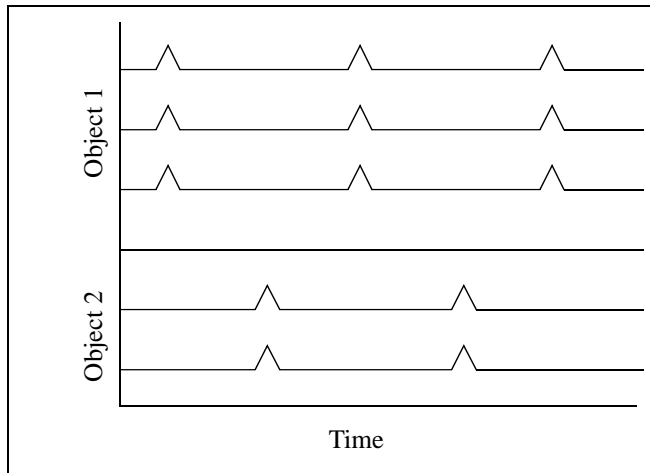


Figure 2: Temporal synchrony. Three cells are responding to the first object, two to the second. Cells responding to the same object fire at the same time.

Because the neural hardware is inherently noisy, there is a limit to the number of synchronized phases that can be kept

distinct. This has implications regarding parallel vs sequential visual search, attention, variable binding, short-term memory capacity (the magical number  $7 \pm 2$ ), and other areas of cognitive science. In the rest of this section, we explore how temporal synchrony may aid in number estimation.

### Subitizing

Temporal synchrony provides a simple explanation for the subitizing phenomenon. Suppose there is a unit which fires whenever any other unit fires. This *subitizing unit* repolarizes faster than the other units, but not so fast that it fires more than once in response to a synchronized pulse. This is shown in Figure 3.

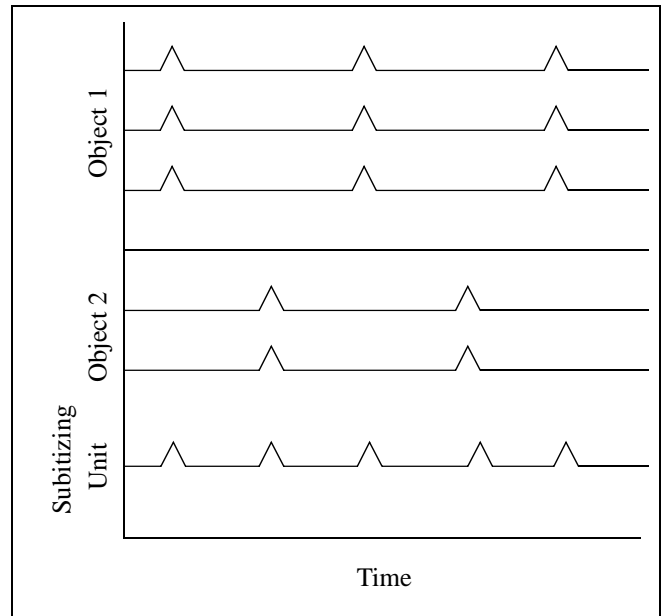


Figure 3: Subitizing with temporal synchrony. The subitizing unit fires whenever any other unit fires.

The frequency with which the subitizing unit fires encodes the number of objects visible. Beyond 3 or 4 objects, the phases begin to blur together; the subitizing unit fires at its maximum rate and the subject perceives “many” objects.

### Estimation

Numbers beyond the subitizing range can still be estimated with the help of temporal synchrony. If a unit accepts input only from others firing at a particular phase, it only receives input regarding one object. Even if there are too many objects to subitize, this can provide some useful information: the size of a typical object. If this amount is used to scale the total amount of area visible (effectively dividing the total by the size of one object), a continuous representation of the number of objects is produced. This is not a terribly accurate mechanism (it suffers from the magnitude effect), but it is much faster than counting.

## Related Work

An alternative model of nonverbal numerical abilities is the accumulator model proposed by Meck & Church (1983; Gallistel & Gelman, 1992). This model proposes an accumulator which integrates over time. As each object is perceived, the activity of the accumulator increases by a fixed amount. The accumulator's activity then serves as a representation of number.

The imprecision of larger numbers is explained as variability in the pulses passed to the accumulator. The more pulses there have been, the less accurate the resulting value in the accumulator.

Our model differs in two ways. First, we explain the lower precision of larger amounts with the squashing function. We have difficulty conceiving of a neurally plausible accumulator which is capable of both taking on very large values and providing precision for small values.

A second difference is that our model is strictly parallel, while the accumulator model is sequential: the stimuli are "fed into" the accumulator one after another. In a static image, this would require a pointing strategy, with the infant carefully "counting" each item exactly once. Since this is not a trivial task even for 3-year-olds (Fuson, 1988), it is difficult to believe that infants would have this ability. Our model does not ask so much; indeed, the quantitative model does not even require the infant to break the image down into separate objects.

## Conclusions and Future Work

We have presented a two-part model of proto-numerical abilities. The model reproduces human data on numerical perception without any explicit counting. The abilities granted by the model may provide useful grounding to children as they learn conventional counting.

The quantitative model predicts that infants will dishabituate to a sufficiently large change in either total area or total contour length. The exact meaning of "sufficiently large" depends on four parameters, and we have found values for these parameters which are consistent with several existing studies.

The quantitative model makes an interesting, counterintuitive prediction: infants will *not* dishabituate in Starkey & Cooper's 2 vs 3 condition with dots of certain sizes (e.g., extremely large ones). Conversely, the model predicts that infants *will* dishabituate in the 4 vs 6 condition for other dot sizes.

More specific predictions can be cautiously made based on the particular parameter values we found. The perceived difference between images is graphed as a function of stimulus size in Figure 4. Where this magnitude exceeds the threshold, habituation is predicted. Specific predictions are given in Table 3. We have begun empirical studies to test these predictions.

The second, more tentative part of the model accepts the temporal synchrony hypothesis of object individuation. Each visible object (or some of them, if there are too many) is bound to a particular phase. Within the subitizing range, the density of the phases indicates the number of objects. Beyond this range, the amount of area present at a particular

phase indicates the size of an individual object, which can in turn be used to estimate the number of objects.

The second part of the model makes a less surprising prediction: estimating the number of objects visible should be difficult if the objects vary greatly in size.

We are now working on a connectionist implementation of our model, based on Gasser and Colunga's (1997) Playpen model of object individuation and spatial relations.

## Acknowledgments

We wish to thank the following for their comments and encouragement: Mike Gasser, Deborah Alterman, Paul Purdom, Heather Drake, and Dan Friedman.

## References

- Antell, S. E., & Keating, D. P. (1983). Perception of numerical invariance in neonates. *Child Development, 54*, 695-701.
- Clearfield, M. W., & Mix, K. S. (1999). Number versus contour length in infants' discrimination of small visual sets. *Psychological Science, 10*, 408-411.
- Eckhorn, R., et al. (1988). Coherent Oscillations: A Mechanism of Feature Linking in the Visual Cortex? *Biological Cybernetics, 60*, 121-130.
- Fechner, G. (1860). In D. H. Howes & E. G. Boring (Eds.), *Elements of Psychophysics, volume 1*. New York: Holt, Rinehart, and Winston. Cited in Kandel, Schwartz, & Jessell, 1991.
- Fuson, K. C. (1988). *Children's Counting and Concepts of Number*. New York: Springer-Verlag.
- Gasser, M. & Colunga, E. (1997). Playpen: Toward an Architecture for Modeling the Development of Spatial Cognition. Technical Report #195, Indiana University Cognitive Science Program.
- Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition, 44*, 43-74.
- Kandel, E. R., Schwartz, J. H., & Jessell, T. M., Eds. (1991). *Principles of Neural Science, 3rd edition*. New York: Elsevier.
- Karmel, B. Z. (1969). The effect of age, complexity, and amount of contour on pattern preferences in human infants. *Journal of Experimental Child Psychology, 7*, 339-354.
- Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes, 9*, 320-334.
- Rumelhart, D. E., McClelland, J. L., et al. (1986). *Parallel Distributed Processing*. Cambridge: MIT Press.
- Starkey, P., & Cooper, R. (1980). Perception of numbers by human infants. *Science, 210*, 1033-1034.

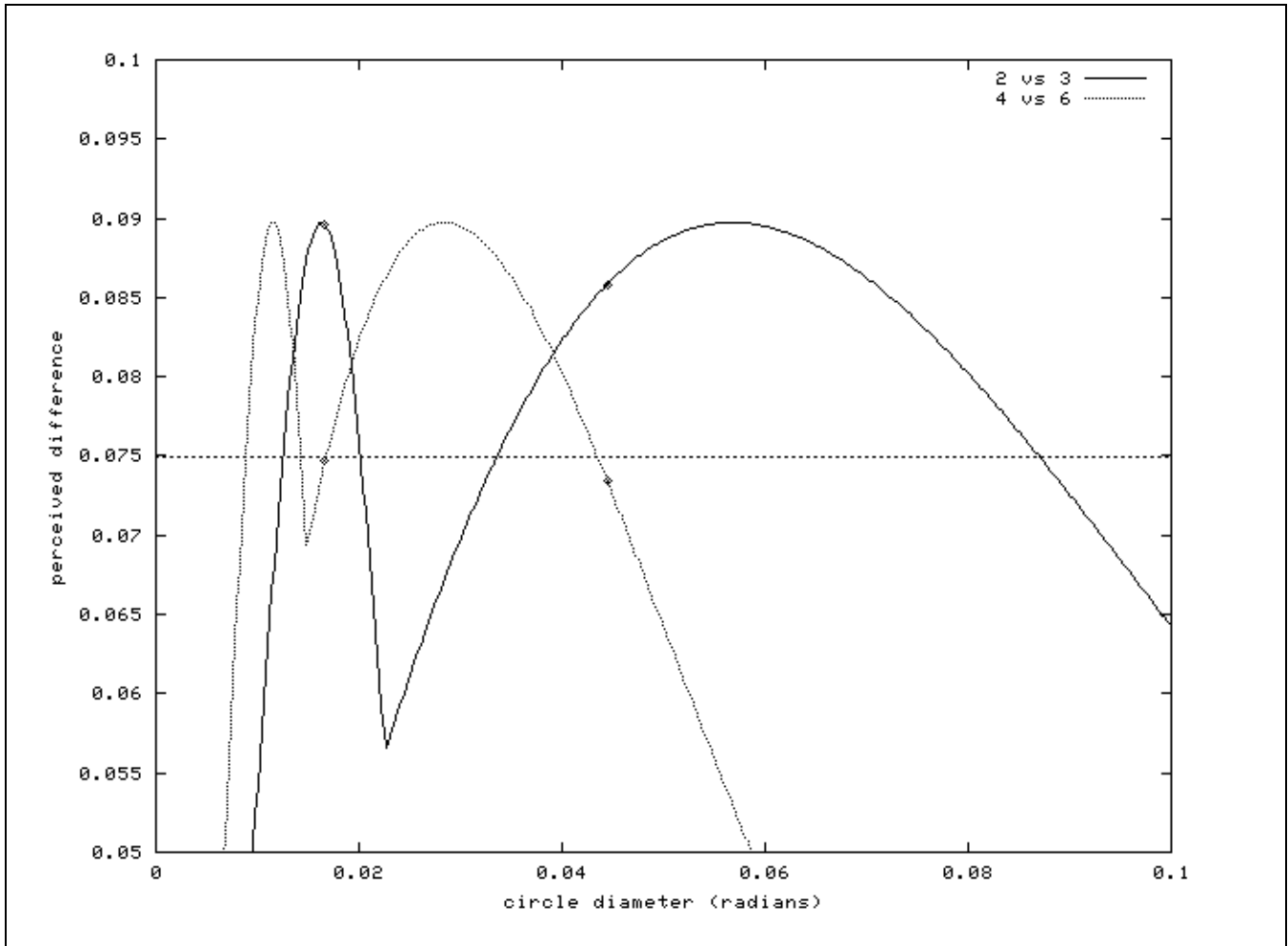


Figure 4: Perceived difference between images as a function of circular stimulus diameter. The M-shaped curves result from taking the maximum of the perceived area difference and the perceived contour difference. The marked points are the data from Starkey & Cooper (around 0.017 radians) and Antell & Keating (around 0.044 radians).

Table 3: Predictions of the model.

Circle Diameter		Dishabituation?	
radians	cm @ 60cm	2 vs 3	4 vs 6
< 0.0088	< 0.53	no	no
0.0099 - 0.012	0.53 - 0.72	no	yes
0.012 - 0.014	0.72 - 0.84	yes	yes
0.014 - 0.017 (includes Starkey & Cooper)	0.84 - 1.0	yes	no
0.017 - 0.020	1.0 - 1.2	yes	yes
0.020 - 0.034	1.2 - 2.0	no	yes
0.034 - 0.043	2.0 - 2.6	yes	yes

Table 3: Predictions of the model.

Circle Diameter		Dishabituation?	
radians	cm @ 60cm	2 vs 3	4 vs 6
0.043 - 0.087 (includes Antell & Keating)	2.6 - 5.2	yes	no
> 0.087	> 5.2	no	no

# Subjacency Constraints without Universal Grammar: Evidence from Artificial Language Learning and Connectionist Modeling

Michelle R. Ellefson (ellefson@siu.edu)

Morten H. Christiansen (morten@siu.edu)

Department of Psychology  
Southern Illinois University - Carbondale  
Carbondale, IL 62901-6502 USA

## Abstract

The acquisition and processing of language is governed by a number of universal constraints, many of which undoubtedly derive from innate properties of the human brain. However, language researchers disagree about whether these constraints are linguistic or cognitive in nature. In this paper, we suggest that the constraints on complex question formation, traditionally explained in terms of the linguistic principle of subjacency, may instead derive from limitations on sequential learning. We present results from an artificial language learning experiment in which subjects were trained either on a “natural” language involving no subjacency violations, or an “unnatural” language that incorporated a limited number of subjacency violations. Although two-thirds of the sentence types were the same across both languages, the natural language was acquired significantly better than its unnatural counterpart. The presence of the unnatural subjacency items negatively affected the learning of the unnatural language as a whole. Connectionist simulations using simple recurrent networks, trained on the same stimuli, replicated these results. This suggests that sequential constraints on learning can explain why subjacency violations are avoided: they make language more difficult to learn. Thus, the constraints on complex question formation may be better explained in terms of innate *cognitive* constraints, rather than linguistic constraints deriving from an innate Universal Grammar.

## Introduction

One aspect of language that any comprehensive theory of language must explain is the existence of linguistic universals. The notion of language universals refers to the observation that although the space of logically possible linguistic subpatterns is vast; the languages of the world only take up a small part of it. That is, there are certain universal tendencies in how languages are structured and used. Theories of language evolution seek to explain how these constraints may have evolved in the hominid lineage. Some theories suggest that the evolution of a Chomskyan Universal Grammar (UG) underlies these universal constraints (e.g., Pinker & Bloom, 1990). More recently, an alternative perspective is gaining ground. This approach advocates a refocus in evolutionary thinking; stressing the adaptation of linguistic structures to the human brain rather than vice versa (e.g., Christiansen, 1994; Kirby, 1998). Accordingly, language has evolved to fit sequential learning and processing mechanisms existing prior to the appearance of language.

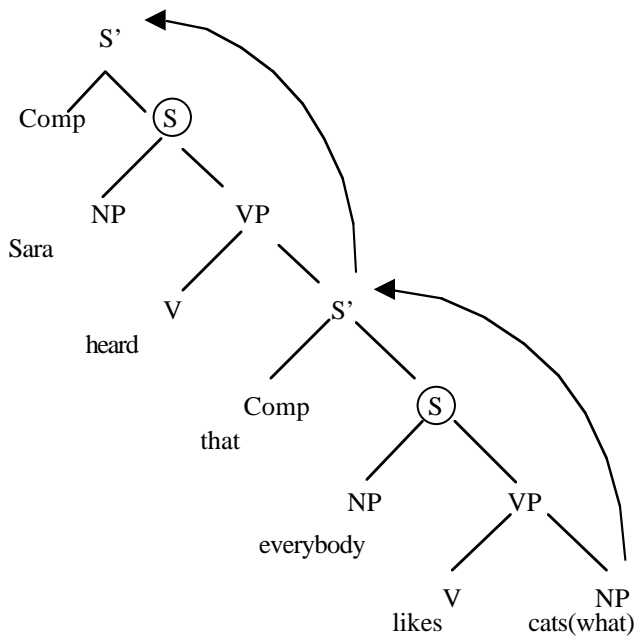
These mechanisms presumably also underwent changes after the emergence of language, but the selective pressures are likely to have come not only from language but also from other kinds of complex hierarchical processing, such as the need for increasingly complex manual combinations following tool sophistication. Thus, many language universals may reflect non-linguistic, cognitive constraints on learning and processing of sequential structure rather than innate UG.

This perspective on language evolution also has important implications for current theories of language acquisition and processing. It suggests that many of the cognitive constraints that have shaped the evolution of language are still at play in our current language ability. If this is correct, it should be possible to uncover the source of some linguistic universal in human performance on sequential learning tasks. Christiansen (2000; Christiansen & Devlin, 1997) has previously explored this possibility in terms of a sequential learning explanation of basic word order universals. He presented converging evidence from theoretical considerations regarding rule interactions, connectionist simulations, typological language analyses, and artificial language learning in normal adults and aphasic patients, corroborating the idea of cognitive constraints on basic word order universals.

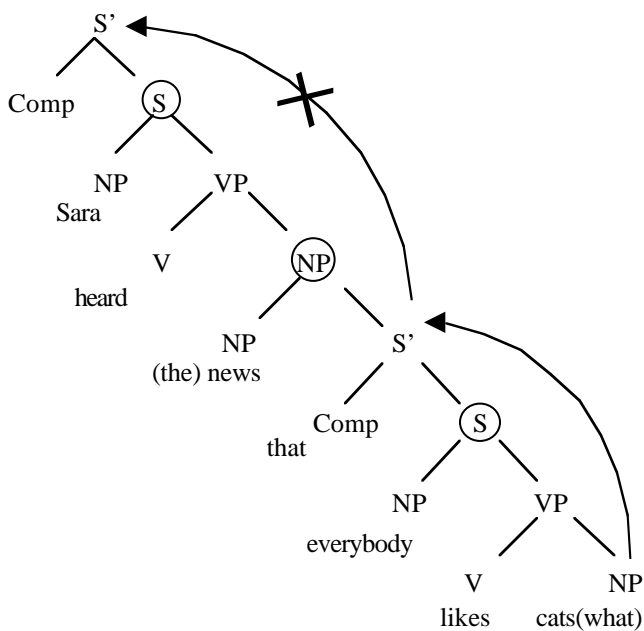
In this paper, we take a similar evolutionary approach to one of the classic linguistic universals: subjacency. We first briefly discuss some of the linguistic data that have given rise to the subjacency principle. Next, we present an artificial language learning experiment that investigates our hypothesis that limitations on sequential learning rather than an innate subjacency principle provide the appropriate constraints on complex question formation. Finally, we report on a set of connectionist simulations in which networks are trained on the same material as the humans, and with very similar results. Taken together, the results from the artificial language learning experiment and the connectionist simulations support our idea that subjacency violations are avoided, not because of an innate subjacency principle, but because of cognitive constraints on sequential learning.

## Why Subjacency?

According to Pinker and Bloom (1990), subjacency is one of the classic examples of an arbitrary linguistic constraint that makes sense only from a linguistic perspective. Informally, the subjacency principle involves the assumption of certain



- 2a. Sara heard that everybody likes cats.
2. What (did) Sara hear that everybody likes?



1. Sara heard (the) news that everybody likes cats.
3. \* What (did) Sara hear (the) news that everybody likes?

principles governing the grammaticality of sentences. "Subjacency, in effect, keeps rules from relating elements that are 'too far apart from each other', where the distance apart is defined in terms of the number of designated nodes that are between them" (Newmeyer, 1991, p. 12). Consider the following sentences:

1. Sara heard (the) news that everybody likes cats.  
N V Wh N V N
2. What (did) Sara hear that everybody likes?  
Wh N V Comp N V
3. \*What (did) Sara hear (the) news that everybody likes?  
Wh N V N Comp N V

According to the subjacency principle, sentence 3 is ungrammatical because too many boundary nodes are placed between the noun phrase complement (NP-Comp) and its respective 'gaps'.

The subjacency principle, in effect, places certain restrictions on the ordering of words in complex questions. The movement of *wh*-items (*what* in Figure 1) is limited as far as the number of so-called bounding nodes that it may cross during its upward movement. In Figure 1, these bounding nodes are the S and NP's that are circled. Put informally, as a *wh*-item moves up the tree it can use comps as temporary "landing sites" from which to launch the next move. The subjacency principle states that during any move only a single bounding node may be crossed. Sentence 2 is therefore grammatical because only one bounding node is crossed for each of the two moves to the top comp node. Sentence 3 is ungrammatical, however, because the *wh*-item has to cross two bounding nodes—NP and S—between the temporary comp landing site and the topmost comp.

Not only do subjacency violations occur in NP-complements, but they can also occur in Wh-phrase complements (Wh-Comp). Consider the following examples:

4. Sara asked why everyone likes cats.  
N V N Comp N V N
5. Who (did) Sara ask why everyone likes cats?  
Wh N V Wh N V N
6. \*What (did) Sara ask why everyone likes?  
Wh N V Wh N V

According to the subjacency principle, sentence 6 is ungrammatical because the interrogative pronoun has moved across too many bounding nodes (as was the case in 3).

In the remainder of this paper, we explore an alternative explanation of the restrictions on complex question formation. This alternative explanation suggests that subjacency violations are avoided, not because of a biological adaptation incorporating the subjacency principle, but because language *itself* has undergone adaptations to root out such violations in response to non-linguistic constraints on sequential learning

Figure 1. Syntactic trees showing grammatical (2) and ungrammatical (3) Wh-movement.

**Table 1. The Structure of the Natural and Unnatural Languages (with Examples)**

NAT		UNNAT	
Sentence	Letter String Example	Sentence	Letter String Example
1. N V N	Z V X	1. N V N	Z V X
2. Wh N V	Q Z M	2. Wh N V	Q Z M
3. N V N comp N V N	Q X M S X V	3. N V N comp N V N	Q X M S X V
4. N V Wh N V N	X M Q X M X	4. N V Wh N V N	X M Q X M X
5. Wh N V comp N V	Q X V S Z M	5*. Wh N V N comp N V	Q X V X S Z M
6. Wh N V Wh N V N	Q Z V Q Z V Z	6*. Wh N V Wh N V	Q Z V Q Z V

Note: Nouns (N) = {Z, X}; Verbs (V) = {V, M}; comp = S; Wh = Q.

### Artificial Language Experiment

Artificial language learning has been shown to be an effective tool in the understanding of the acquisition of language (e.g., Gomez & Gerken, 1999; Saffran, Aslin, & Newport, 1996). More recently, artificial language learning has been used to explore how languages themselves may have evolved in the human species (Christiansen, 2000).

### Subjects

Sixty undergraduates were recruited from an introductory psychology class at Southern Illinois University, and earned course credit for their participation.

### Materials

We created two artificial languages, natural (NAT) and unnatural (UNNAT). Each artificial language consisted of a set of letter strings. The letters in the strings each represented a specific grammatical class (see Table 1). The letters Z and X represented nouns. V and M stood for verbs. The letter S designated a complementizer. Interrogative pronouns were denoted by the letter Q. These strings were constructed based on the sentence structure of the six examples discussed above. Unique letter strings were created for training and testing sessions.

**Training Stimuli** Twenty letter strings, 10 of each for NAT and UNNAT, were created to represent grammatical and ungrammatical complex question formation structures (SUB). The grammatical SUB items used for the NAT training, while the ungrammatical SUB items were used for UNNAT training. Examples of SUB letter strings for both conditions can be seen in Table 1 as sentences 5 and 6.

An additional 20 general training items were constructed to represent general grammatical structures (GEN) that do not involve subadjacency. These items were the same for both languages. Examples of GEN letter strings for both conditions are sentences 1 through 4 in Table 1. In summary, 10

SUB and 20 GEN training strings were created for each language.

**Test Stimuli** An additional set of novel letter strings was created for the test session. For each language there were 30 grammatical items and 30 ungrammatical items. Twenty-eight novel SUBs were constructed. For these unique SUB letter strings there were 14 each, of grammatical and ungrammatical complement structures. Grammaticality in both languages was based on what the grammar for that condition specified as legal sentences (Table 1)—not by what may be a grammatical/ungrammatical sentence in English. Thus, for the UNNAT language, the ungrammatical SUBs (from the viewpoint of English) were scored as grammatical and the grammatical SUBs (from the viewpoint of English) were scored as ungrammatical. Grammaticality in the NAT language corresponded to English, with grammatical SUBs scored as grammatical and ungrammatical SUBs scored as ungrammatical. Testing in both groups also included 16 novel grammatical GEN items and 16 novel ungrammatical GEN items in which one of the letters, except those in the first and last position, were changed.

Previous artificial language learning research has established that distributional “surface” information, computed over fragments consisting of two or three consecutive letters (bigrams/trigrams), may affect how well a language is learned. In order to ensure that the NAT language was not more “regular” than the UNNAT language, in terms of distributional information, and therefore potentially easier to learn, we controlled our stimuli for five different kinds of fragment information.

1) *Associative chunk strength* is measured as the sum of the frequency of occurrence in the training items of each of the fragments in a test item, weighted by the number of fragments in that item (Knowlton & Squire, 1994). E.g., the associative chunk strength of the item ZVX would be calculated as the sum of the frequencies of the fragments ZV, VX and ZVX divided by 3. Two-tailed t-tests indicated that there were no differences between the languages in associative chunk strength for the grammatical ( $t < 1$ ) and the ungrammatical ( $t < 1$ ) items.

2) *Anchor strength* is measured as the relative frequency of initial and final fragments in similar anchor positions in the training items (Knowlton & Squire, 1994). E.g., the anchor strength of the item QXMSXV is calculated as the sum of the frequencies of the fragments QX and QXM in initial positions in the training items and of the fragments XV and SXV in final positions in the training items. Again, there were no differences between the two languages in the anchor strength of the grammatical ( $t(58)=1.75, p>.085$ ) or the ungrammatical items ( $t<1$ ).

3) *Novelty* is measured as the number of fragments that did not appear in any training item (Redington & Chater, 1996). E.g., if the fragments XVS and VS from the item QXVSZM never occurred in a training item, then the test item would receive a novelty score of 2. Here there is a significant difference between the novelty scores for the grammatical test items in the NAT language (.43) and the UNNAT language (0) ( $t(58)=3.50, p<.001$ ). However, given that items with novel fragments will seem less familiar they are more likely to not to be accepted as grammatical, making it more difficult to correctly classify the test items from the NAT language. Thus this difference provides a bias against our hypothesis that the NAT language should be easier to learn. There were no differences between the ungrammatical items ( $t<1$ ).

4) *Novel fragment position* is measured as the number of fragments that occur in novel absolute positions where they did not occur in any training item (Johnstone & Shanks, 1999). E.g., if the fragment VQZ from the item QZVQZV never occurred in this absolute position in any of the training items then this item would be assigned a novel fragment position score of 1. There were no differences between the novel fragment scores for the grammatical ( $t(58)=1.54, p>.13$ ) or ungrammatical items ( $t<1$ ) across the two languages.

5) *Global similarity* is measured as the number of letters that a test item is different from the nearest training item (Vokey & Brooks, 1992). E.g., if the test item QZM has QZV as its closest training item then it would be assigned a global similarity score of 1. There were no differences between the two languages for the grammatical ( $t=0$ ) and ungrammatical ( $t<1$ ) items.

## Procedures

Subjects were randomly assigned to one of three conditions (NAT, UNNAT, and CONTROL). NAT and UNNAT were trained using the natural and unnatural languages, respectively. The CONTROL group completed only the test session. During training, individual letter strings were presented briefly on a computer. After each presentation, participants were prompted to enter the letter string using the keyboard. Training consisted of 2 blocks of the 30 items, presented randomly. During the test session, participants decided if the test items were created by the same (grammatical) or different (ungrammatical) rules as the training items. Testing consisted of 2 blocks of 60 items, again presented randomly.

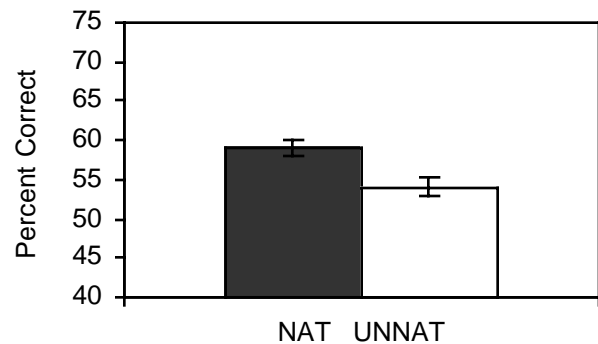


Figure 2. Overall correct classification for NAT and UNNAT languages.

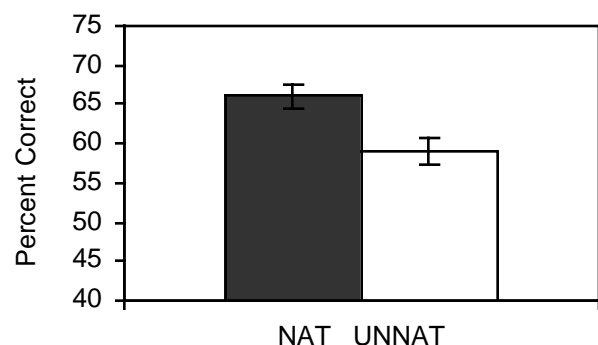


Figure 3. Correct classification of GEN items for NAT and UNNAT languages.

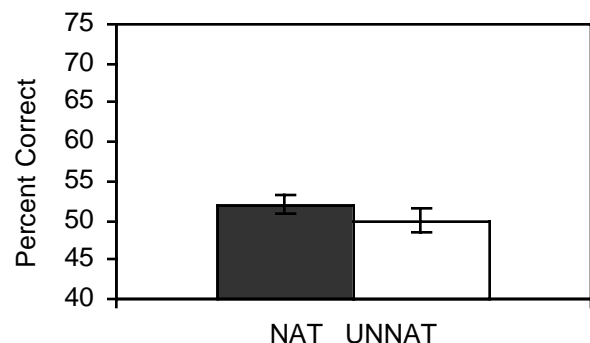


Figure 4. Correct classification of SUB items for NAT and UNNAT languages.

## Results and Discussion

**Control Group** Since the test items were the same for all groups, but scored differently depending on training condition, the control data was scored from the viewpoint of both the natural and unnatural languages. Differences between correct and incorrect classification from both language perspectives were non-significant with all  $t$ -values  $<1$  (range of correct classification: 49.5%–50.5%). Thus, there was no inherent bias in the test stimuli toward either language.

**Experimental Group** An overall t-test indicated that NAT (59%) learned the language significantly better than UNNAT (54%) (Figure 2;  $t(38)=3.27, p<.01$ ). This result indicates that the UNNAT was more difficult to learn than the NAT. Both groups were able to differentiate the grammatical and ungrammatical items (NAT:  $t(38)=4.67, p<.001$ ; UNNAT:  $t(38)=2.07, p<.05$ ). NAT correctly classified 70% of the grammatical and 51% of the ungrammatical items. UNNAT correctly classified 61% of the grammatical and 47% of the ungrammatical items. NAT (66%) exceeded UNNAT (59%) at classifying the common GEN items (Figure 3;  $t(38)=2.80, p<.01$ ). Although marginal, NAT (52%) was also better than UNNAT (50%) at classifying SUB items (Figure 4;  $t(38)=1.86, p=.071$ ). Note that the presence of the SUB items affected the learning of the GEN items. Even though both groups were tested on exactly the same GEN items, the UNNAT performed significantly worse on these items. Thus, the presence of the subjacency violations in the UNNAT language affected the learning of the *language as a whole*, not just the SUB items. From the viewpoint of language evolution, languages such as UNNAT would loose out in competition with other languages such as NAT because the latter is easier to learn.

### Connectionist Simulations

In principle, one could object that the reason why we found differences between the NAT and the UNNAT groups is because the NAT group is in some way tapping into an innately specified subjacency principle when learning the language. Another possible objection is that the NAT language follows the general pattern of English whereas the UNNAT language does not, and that our human results could potentially reflect an “English effect”. To counter these possible objections and to support our suggestion that the difference in learnability between the two languages is brought about by constraints arising from sequential learning, we present a set of connectionist simulations of our human data.

### Networks

For the simulations, we used simple recurrent networks (SRNs; Elman, 1991) because they have been successfully applied in the modeling of both non-linguistic sequential learning (e.g., Christiansen & Devlin, 1997; Cleeremans, 1993) and language processing (e.g., Christiansen, 1994; Elman, 1991). SRNs are standard feed-forward neural networks equipped with an extra layer of so-called context units. The SRNs used in our simulations had 7 input/output units (corresponding to each of the 6 letters plus an end of sentence marker) as well as 8 hidden units and 8 context units. At a particular time step  $t$ , an input pattern is propagated through the hidden unit layer to the output layer. At the next time step,  $t+1$ , the activation of the hidden unit layer at time  $t$  is copied back to the context layer and paired with the current input. This means that the current state of

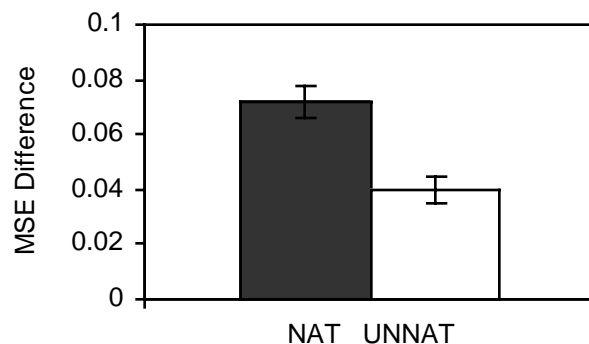


Figure 5. MSE differences for grammatical (low error) and ungrammatical (high error) items for NAT and UNNAT networks.

the hidden units can influence the processing of subsequent inputs, providing an ability to deal with integrated sequences of input presented sequentially.

### Materials

For the simulations we used the same training and test items as in the artificial language learning experiment.

### Procedures

Forty networks, with different initial weight randomizations (within  $\pm .5$ ), were trained to predict the next consonant in a sequence. The networks were randomly assigned to the NAT and UNNAT training conditions, and given 20 passes through a random ordering of the 30 training items appropriate for a given condition. The learning rate was set to .1 and the momentum to .95. After training, the networks were tested separately on the 30 grammatical and 30 ungrammatical test items (again, according to their respective grammar).

Following successful training, an SRN will tend to output a probability distribution of possible next items given the previous sentential context. Performance was measured in terms of how well the networks were able to approximate the correct probability distribution given previous context. The results are reported in terms of the Mean Squared Error (MSE) between network predictions for a test set and the empirically derived, full conditional probabilities given the training set (Elman, 1991). This error measure provides an indication of how well the network has acquired the grammatical regularities underlying a particular language, and thus allows for a direct comparison with our human data.

### Results and Discussion

The results show that the NAT networks had a significantly lower MSE (.185; SD: .021) than the UNNAT networks (.206; SD: .023) on the grammatical items ( $t(38)=2.85, p<.01$ ). On the ungrammatical items, the NAT nets had a slightly higher error (.258; SD: .036) compared with the UNNAT nets (.246; SD: .034), but this difference was not significant ( $t<1$ ). This pattern resembles the performance of the human subjects where the NAT group was 11% better



than the UNNAT group at classifying the grammatical items, though this difference only approached significance ( $t(38)=1.10, p=.279$ ). The difference was only <3% in favor of the NAT group for the ungrammatical items ( $t=1$ ). Also similarly to the human subjects, there was a significant difference between the MSE on the grammatical and the ungrammatical items for both the NAT nets ( $t(38)=7.69, p<.001$ ) and the UNNAT nets ( $t(38)=4.33, p<.001$ ). It is plausible to assume that the greater the difference between the MSE on the grammatical (low error) and the ungrammatical (higher error) items, the easier it should be to distinguish between the two types of items. As illustrated in Figure 5, the NAT networks have a significantly better basis for making such distinctions than the UNNAT networks (.072 vs. .040;  $t(38)=4.31, p<.001$ ). Thus, the simulation results closely mimic the behavioral results, even though the SRNs clearly do not have a built-in subadjacency principle nor do they have prior knowledge of English. The results therefore corroborate our suggestion that constraints on the learning and processing of sequential structure can explain why subadjacency violations tend to be avoided: they were weeded out because they made the sequential structure of language too difficult to learn.

### Conclusion

In this paper, we have provided evidence in favor of an alternative account of the universal constraints on complex question formation. The artificial language learning results show that not only are constructions involving subadjacency violations hard to learn in and by themselves, but their presence also makes the language as a whole harder to learn. The connectionist simulations further corroborated these results, emphasizing that the observed learning difficulties in relation to the unnatural language arise from non-linguistic constraints on sequential learning. These results, together with the results on word order universals (Christiansen, 2000; Christiansen & Devlin, 1997), suggest that constraints arising from general cognitive processes, such as sequential learning and processing, are likely to play a larger role in sentence processing than has traditionally been assumed. This means that what we observe today as linguistic universals may be stable states that have emerged through an extended process of linguistic evolution. When language itself is viewed as a dynamic system sensitive to adaptive pressures, natural selection will favor combinations of linguistic constructions that can be acquired relatively easily given existing learning and processing mechanisms. Consequently, difficult to learn language fragments, such as our unnatural language, will tend to disappear. Furthermore, if we assume that the production system is based conservatively on a processing system acquired in the service of comprehension, then this system would be unlikely to produce subadjacency violations because they would not be represented there in the first place. In conclusion, rather than having an innate UG principle to rule out subadjacency violations, we suggest they

may have been eliminated altogether through an evolutionary process of linguistic adaptation constrained by prior cognitive limitations on sequential learning and processing.

### Acknowledgments

We would like to thank Takashi Furuhashi, Lori Smorzynski, and Brad Appelhans for their help with data collection.

### References

- Christiansen, M. H. (1994). *Infinite languages, finite minds: Connectionism, learning and linguistic structure*. Unpublished doctoral dissertation, Centre for Cognitive Science, University of Edinburgh, U. K.
- Christiansen, M. H. (2000). Using artificial language learning to study language evolution: Exploring the emergence of word order universals. In J. L. Dessalles & L. Ghadapour (Eds.), *The Evolution of Language: 3rd International Conference* (pp. 45-48). Paris, France: Ecole Nationale Supérieure des Telecommunications.
- Christiansen, M.H. & Devlin, J.T. (1997). Recursive Inconsistencies Are Hard to Learn: A Connectionist Perspective on Universal Word Order Correlations. In *Proceedings of the 19th Annual Cognitive Science Society Conference* (pp. 113-118). Mahwah, NJ: Lawrence Erlbaum Associates.
- Cleeremans, A. (1993). *Mechanisms of implicit learning: Connectionist models of sequence processing*. Cambridge, MA: MIT Press.
- Elman, J.L. (1991). Distributed representation, simple recurrent networks, and grammatical structure. *Machine Learning*, 7, 195-225.
- Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, 70, 109-135.
- Johnstone, T., & Shanks, D. R. (1999). Two mechanisms in implicit artificial grammar learning? Comment on Meulemans and Van der Linden (1997). *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 25(2), 524-531.
- Kirby, S. (1998). *Language evolution without natural selection: From vocabulary to syntax in a population of learners*. Edinburgh Occasional Paper in Linguistics, EOPL-98-1.
- Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 20(1), 79-91.
- Newmeyer, F. (1991). Functional explanation in linguistics and the origins of language. *Language and Communication*, 11(1/2), 3-28.
- Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Brain and Behavioral Sciences*, 13(4), 707-727.
- Redington, M., & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of Experimental Psychology: General*. 125(2), 123-138.
- Saffran, J. R., Aslin, R. N., Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926-1928.

# Reflective Introspective Reasoning Through CBR

Susan Eileen Fox  
fox@macalester.edu  
Macalester College  
1600 Grand Avenue  
Saint Paul, MN 55105 USA

## Abstract

In recent years, “introspective reasoning” systems have been developed to model the ability to reason about one’s own reasoning performance. This research examines “reflective” introspective reasoning: introspecting about the introspective reasoning process, itself. We introduce a reflective introspective reasoning system that uses case-based reasoning (CBR) as its central reasoning method. We examine the advantages of such a system, and attempt to classify the reasoning failures within introspective system that indicate a need to reflect higher.

## Introduction

In artificial intelligence, meta-reasoning systems have been used for a variety of purposes: to predict the behavior of other agents, to guide the acquisition and application of domain knowledge, and to “learn” by adjusting the system’s own reasoning processes in response to experience. Meta-reasoning systems model the ability humans have to reason about our own and others’ reasoning performance. Systems that apply meta-knowledge and meta-reasoning to improvement of their own reasoning processes are called “introspective learning” systems.

The ability to reason introspectively requires knowledge of one’s own reasoning methods, and observation, evaluation, and alteration of those methods when needed. A similar ability has been documented in studies of human reasoning behavior, and modeled with a variety of artificial intelligence techniques. Introspective reasoning systems often lack a critical component of human meta-reasoning: the ability to introspect about our introspections themselves. Introspective reasoning is often applied only to reasoning in service of some underlying task, such as navigation planning. An introspective reasoner that can apply its reasoning to its own introspective processes and repeat this reflection upon itself indefinitely is called “reflective” (Ibrahim, 1992).

We have developed a reflective introspective reasoner that learns by altering its reasoning methods at all levels. RILS<sup>1</sup> analyzes both its task-level planning process and its introspective reasoning process, using a unified reasoning method, case-based reasoning (CBR), for all its tasks. Re-using CBR simplifies the introspective model of the system’s reasoning processes and facilitates reflection.

A key issue for a reflective reasoner is deciding when to reflect. RILS uses failure-driven learning; reflective introspection occurs when an impasse is reached.

In the next section, we provide some background on “reflection,” introspective learning in humans, and other introspective reasoning systems. We then describe the RILS approach to reflective introspection, and we classify the kinds of reasoning failures RILS uses to trigger reflective introspective reasoning.

## Background

### Reflection

The term reflection, as it is used in this paper, refers to systems which can shift the focus of their processing from the current task to the problem-solving task itself, and can repeat this shift of processing indefinitely (Ibrahim, 1992). As a system shifts to a higher-level task, it constructs a “reflective tower” of reasoning processes. Each process analyzes and alters the one beneath it, which is suspended until the higher-level process completes its task.

While not a requirement, Reflection is easiest to achieve if the same reasoning method is used at all reasoning levels. In order to introspect, the system must have a model of its own reasoning processes. The model becomes more complex as the number of different reasoning methods grows. Simplicity in the model enables the system to meaningfully alter more features of its processing, leading to greater flexibility.

The strength of a reflective system is its flexibility: every aspect of the system is open to adaptation and improvement, as the system responds and learns from its experiences. The drawback to a reflective system is the potential of the system to reflect infinitely without making forward progress at any level. The system, when in doubt, must choose not to reflect and continue processing. Many reflective systems only shift attention to a higher level when an explicit failure, particularly a catastrophic failure occurs.

### Human Introspective Reasoning Behavior

Several studies have found evidence that humans engage in introspective learning behavior: altering their reasoning strategies as their experience grows.

Chi & Glaser (1980) found differences between the reasoning strategies of experts and novices: experts approach problems in a more playful way, spend more time

---

<sup>1</sup>Reflective Introspective Learning System

analyzing of a problem before attempting to solve it, and understand better the important features of a problem. They suggest these strategies are learned as part of the process of becoming an expert.

Flavell, Friedrichs, & Hoyt (1970) found that older children were better than younger children at monitoring how well they had performed a given task and judging when they had completed it. The development demonstrated here indicates an awareness of one’s own reasoning processes, as well as learning from that awareness to perform better. Kreutzer, Leonard, & Flavell (1975) found that children improved their understanding of their own memory processes as they became older: asked to describe strategies for remembering things, older children tended to describe many strategies and their outcomes, younger children very few.

Kruger & Dunning (1999) found that competence at a task was related to the ability to accurately judge one’s own competence. This suggests that introspective skills are integrally intertwined with particular domain skills.

In all these cases humans show an improved performance on a domain task when they also show evidence of playful, introspective learning behaviors.

### Introspective Reasoning Systems

A variety of approaches to introspective reasoning systems exist. A few have examined reflective introspection.

SOAR is a rule-based system which does deliberately address reflection (Rosenbloom, Laird, & Newell, 1993). SOAR’s rule base may contain rules which control the rule selection process, among others. SOAR can learn new behaviors by creating new meta-rules. Its meta-rules cannot affect all portions of its reasoning process.

The Massive Memory Architecture performs both introspective reasoning and case-based reasoning in a task-driven manner (Arcos & Plaza, 1993). Autognostic uses a “Structure-Behavior-Function” model to represent reasoning processes (Stroulia & Goel, 1995). RAPTER uses model-based reasoning: an explicit model of “assertions” describing the ideal reasoning behavior is used to diagnose failures (Freed & Collins, 1994). These systems do not include reflective capabilities.

Meta-AQUA maintains reasoning trace templates (Meta-XPs) which describe the patterns of reasoning that indicate reasoning failures (Cox, 1996). Meta-AQUA’s Meta-XPs could be applied to the introspective process itself, but reflection is not the focus of the project.

IULIAN integrates introspective learning with the overall domain task in a way that permits reflection (Oehlmann, Edwards, & Sleeman, 1994). IULIAN uses case-based planning to generate both domain introspective plans, but, as in SOAR, its introspective plans have incomplete access to the mechanisms which use them.

The ROBBIE system (Fox & Leake, 1995), the precursor to RILS, is related to the model-based reasoning systems described above. It contains an explicit collection of assertions which describe the ideal reasoning process of its case-based planner. ROBBIE’s uses its model to perform detection, diagnosis, and repair of reason-

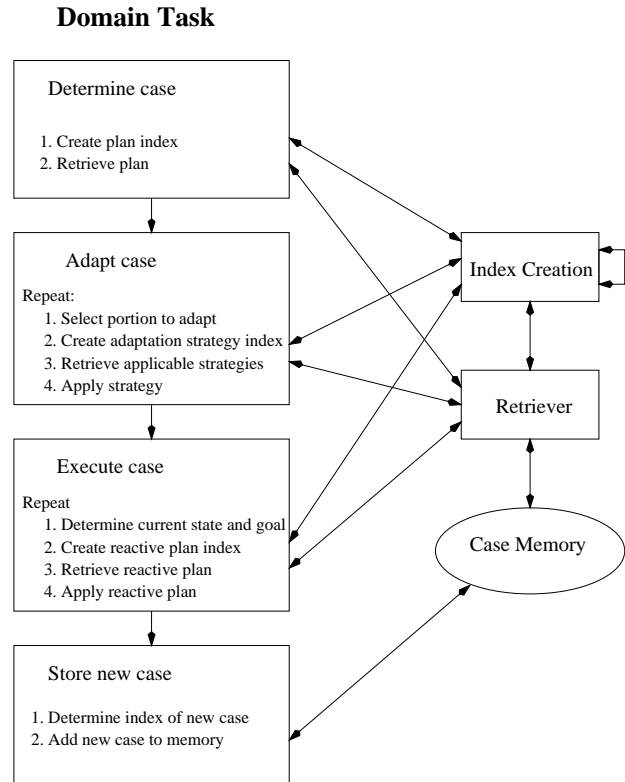


Figure 1: The domain task architecture of RILS: reuse of CBR

ing failures. It is not reflective; its introspective model describes only its underlying planning system. ROBBIE does reuse its CBR processes for multiple planning tasks. RILS was developed from ROBBIE as a *case-based* introspective learner that retains an explicit model of its reasoning embodied in a set of introspective cases.

### The Reflective Introspective Learning System

RILS performs route planning for a simulated robot, supported by case-based and introspective learning. Case-based reasoning is the central reasoning method for all RILS tasks. The case memory stores cases for creating and executing route plans, and also stores “assertion cases” that, taken together, comprise its introspective model. The model captures the reasoning processes of both planning and introspective tasks.

#### RILS’ Domain Task

RILS operates in the same domain as the ROBBIE system (Fox & Leake, 1995); it navigates a simulated robot through a domain of streets, using case-based planning to create high-level plans and case-based reactive planning to interactively execute the plans in its simulated domain. Figure 2 shows a sample of RILS’ domain.

RILS reuses its case-based index creation, retrieval, and case memory for multiple domain tasks: creation of an index, selection of a high-level plan, adaptation of the plan, and selection of reactive planlets to execute

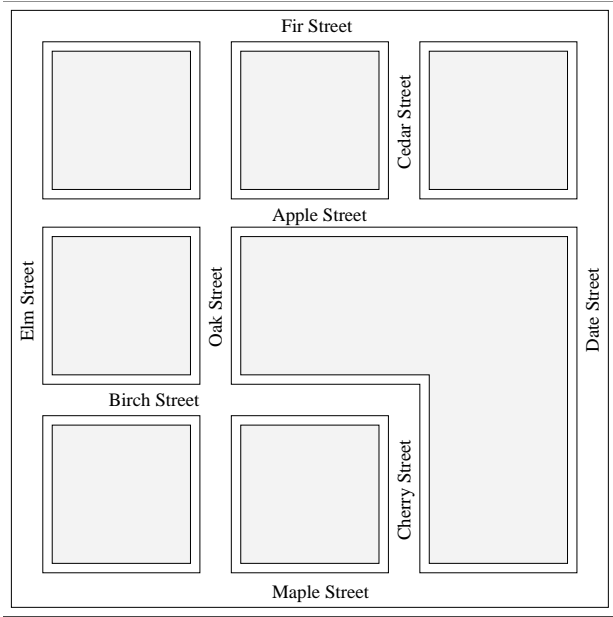


Figure 2: A sample of RILS' domain

its goals. Learning occurs at the domain level through the storage of new high-level plans once they have been successfully executed, and through the addition of index creation rules which are learned through introspective reasoning. The introspective reasoner recognizes and repairs those times when the plan case retrieval retrieves a suboptimal match to a given situation.

The following example demonstrates how RILS' operates at the domain level, and how introspective learning occurs to the domain reasoning. Suppose that RILS has as a new goal to move from the corner of Maple and Elm Streets to the corner of Cherry and Maple Streets. It has in its case memory two potential matches to this problem: a plan from the corner of Maple and Elm to the corner of Birch and Oak, and a plan from the corner of Apple and Elm to the corner of Apple and Oak. Without any learned indexing rules, RILS prefers the first plan to the second one, because it shares the same starting location as its current problem. RILS adapts the selected plan and successfully executes it to arrive at its goal.

The process of executing the plan also streamlines it, so that the plan RILS stores back into its case memory involves merely turning east and moving along Maple to Cherry Street. The storage of new plan cases is the most basic level of learning in RILS.

The introspective reasoning system monitors the domain-level reasoning process, and detects a reasoning failure: the plan being stored into memory is more similar to an unretrieved case (from Apple and Elm to Apple and Oak) than to the retrieved one. This triggers introspective reasoning to diagnose and repair the domain level reasoning. The ultimate repair is to add an indexing rule to detect and prefer case matches where the general direction of movement is the same (i.e., "move straight east"). We discuss the introspective task in more detail in the next section.

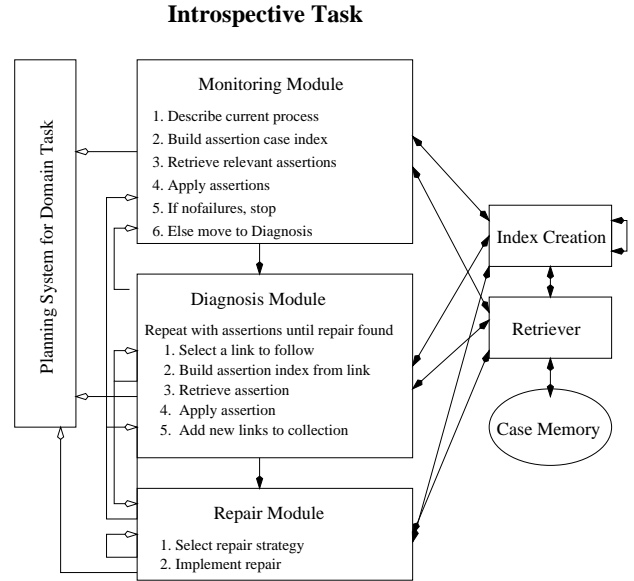


Figure 3: The introspective task architecture of RILS: monitoring applies only to domain-level reasoning; diagnosis and repair apply to both domain and introspective modules.

### RILS' Introspective Task

The goal of the introspective reasoning system is to detect reasoning failures in all portions of the system itself and, where possible, to correct the system's processing to avoid repeating the reasoning failure in the future. A "reasoning failure" is any situation which is not predicted by the system itself, whether it is a failure or a success in terms of the domain task. For example, an unanticipated opportunity to achieve a goal is as much of a reasoning failure as a failure to achieve a goal.

In order to reason about its own reasoning, RILS has a model of its ideal reasoning behavior. Figure 1, which shows RILS' domain reasoning, also represents abstractly one portion of the introspective model. Figure 3 shows RILS' introspective reasoning, represents the rest of its introspective model, as well. For each component of the reasoning process, RILS has a collection of "assertions" at multiple levels of abstraction which describe in detail RILS' expectations about its own performance.

RILS actively monitors each step in the domain reasoning to verify that the actual behavior corresponds to its model. When an assertion fails to be true of the system's actual behavior, RILS suspends the lower-level reasoning task and takes over, attempting to diagnose and repair the detected failure. The diagnosis module searches through its model, examining those assertions which are causally related to the detected failure, until it finds a related assertion which has a repair recommended for it. Once a repair is performed, control returns to the planning process.

While RILS may actively monitor its underlying reasoning task, it cannot actively monitor its introspective reasoning process. To do so would lead immediately to an infinite reflective tower: RILS would monitor<sub>1</sub> its

monitoring, then have to monitor<sub>2</sub> the monitoring<sub>1</sub>, and so forth. Instead, RILS waits for “impasses” in its introspective process: places where the reasoning process cannot continue, including unexpected catastrophic failures at the domain task level. If RILS discovers explicit evidence that the introspective reasoning process itself is flawed, it suspends the introspective process and reflectively applies its diagnosis and repair processes to the introspective task. These impasses are detailed in the next section.

It seems intuitive that RILS should actively monitor its domain-level reasoning and passively monitor its introspective reasoning. Route planning and execution interact with a highly complex, poorly understood world, where reasoning failures are common. RILS’ understanding of the state of the world at any time is limited. On the other hand, the introspective reasoner has a much more restricted domain: the reasoning of the system itself, and must be assumed to be more reliable.

Each component of RILS’ introspective reasoning is implemented using the same case memory and case retrieval mechanisms as are used in the planning task. The introspective model, rather than being a monolithic collection of assertions, is represented by “assertion” cases stored separately in the case memory. When monitoring the planner’s reasoning, RILS constructs an index representing the current point in the reasoning process, and uses it to retrieve those assertion cases which are applicable to that point. When diagnosing a reasoning failure, RILS retrieves cases based on causal links which are stored in each assertion case. RILS’ repair module retrieves and applies repair plans from the case memory.

Assertion cases must contain sufficient information to retrieve them when needed. Consider the sample assertion case show in figure 4: the assertion case says that the diagnosis module will only consider assertions causally relevant to the current problem. Each assertion case contains an assertion (and information to help apply it), links to other causally-related assertions, applicable repair strategies, and statistics on the assertion’s use and success. Assertion cases are retrieved to support monitoring and diagnosis of reasoning failures. As part of the ROBBIE process, a general vocabulary was developed which describes a wide range of assertions about reasoning processes: the generality of this vocabulary is demonstrated by the ease with which the introspective model was transformed to include a reflective component.

We illustrate the reflective aspect of RILS with the following example, extending the example in the previous section. Suppose that RILS experiences the reasoning failure described above: it selects an inappropriate case, diagnoses the failure, and learns a new indexing rule to avoid the faulty selection in the future. All this requires introspection only about the planner. RILS performs reflective introspection when the introspective reasoner itself is faulty. In this case, suppose that the repair module incorrectly instantiated the new indexing rule, so that it will not be retrieved when it is applicable. Some time in the future, RILS plans a route from the corner of

```
(diagnose-spec2
 (assertion diagnosis specific 2 during)
 (and (contains-part assert-case links)
      (member-of-structure
       (part-value assert-case links)
       checked-assertions))
 (variables assert-case checked-assertions)
 (links (abstr (diagnosis general 2))
        (prev (diagnosis specific 1))
        (next (diagnosis specific 3)))
 (repair)
 (statistics (uses 12)
            (failures 0)))
```

Figure 4: An assertion case for the diagnosis component: “Every assertion retrieved during diagnosis will have a link to one already under consideration”

Birch and Elm Streets to the corner of Birch and Oak. It should apply the new indexing rule, but fails to find it. It therefore retrieves an inappropriate case in exactly the same manner as described above. When the domain-level reasoning failure is discovered and diagnosed, RILS notices that this is exactly the same error as it supposedly corrected earlier. This is the evidence RILS needs in order to invoke reflective introspection. It will suspend the introspective task, and apply the diagnosis and repair modules to the introspective task itself.

Diagnosing and repairing this type of introspective reasoning failure is difficult, because of the long time lag between the actual failure and its detection. RILS keeps a history of its past reasoning decisions in order to be able to diagnose an unbounded distance into its past reasoning. At present this history is used to attempt a diagnosis on introspective reasoning failures: repairing such failures is still work in progress. In the next section we classify the kinds of impasse situations and reasoning failures RILS uses to determine opportunities for reflective introspection.

### Reflective reasoning failures in RILS

RILS responds reflectively only to explicit failures which indicate a flaw in the introspective reasoner. Because each module reuses CBR for its reasoning, the sorts of failures RILS must look for are similar for each module. The underlying cause of each failure type differs depending on the module in question: we will examine each module in turn below. The basic categories of failures are:

1. Case memory lacks a required case;
2. RILS fails to retrieve a relevant case;
3. RILS retrieves an irrelevant case; and
4. RILS improperly applies a retrieved case.

Currently, RILS detects these reasoning failures and attempts to reflectively introspect about them. It does not yet repair its introspective process, though work on that aspect is underway.

## The Monitoring Module

The monitoring process examines only the route planner. It retrieves those assertions which are currently relevant and checks that no assertion is violated.

If an assertion case is missing from the case memory, that implies that the introspective model is incomplete, and hence inaccurate. This is a troubling failure, because RILS alters its reasoning processes based on the assumption that its introspective model captures the ideal behavior of the system. If a case is missing, then it is possible that RILS would alter itself incorrectly. This is not a failure type that we anticipate RILS handling in any deep manner: It might be able to conjecture that a case is missing and then let a human user/programmer assist in correcting the situation.

Failing to retrieve a relevant case is an easier problem to detect, though detection may be delayed some period from the occurrence of the failure. A failure to retrieve reflects some flaw in the index of the assertion case, or a flawing the index creation and retrieval mechanism. RILS can examine its case memory for cases that are referred to by other assertion cases in memory but that have not been retrieved along with them. RILS keeps statistics on the application of assertion cases to help with this analysis. Correcting this failure requires the alteration of the indexing and retrieval methods for assertion cases; we expect RILS to incorporate this repair in the future.

Because the monitoring module knows the context of the planner at a given moment, recognizing an irrelevant assertion case is easy to detect on the spot: one case in which reflection can occur at the moment of the reasoning failure. As before, this indicates either a flaw in the index of the particular case retrieved, which is easy to check and correct, or a flaw in how the system creates indices and retrieves assertion cases. RILS can alter the indexing rules for plan cases; we should be able to extend this to altering the indexing of assertion cases as well.

A misapplication of a retrieved case could be detected if the lack of it leads to an unexpected catastrophic failure of some sort. RILS examines catastrophic failures at the route planning level to determine if the monitoring process failed to detect a problem before the catastrophic failure occurred, and considers the possibility that an assertion was misapplied.

## The Diagnosis Module

The diagnosis module retrieves assertion cases and evaluates them in much the same way as the monitoring module. Therefore, many of the comments made about monitoring also apply to diagnosis. The main difference is in how the diagnosis module chooses which cases to retrieve: it starts with an assertion case which is known to be a failure, and then performs a heuristic-guided breadth-first search through those assertions which are causally related to the detected failure. It uses the causal links each assertion contains.

For reflective diagnosis, it may be initially unclear which assertion has failed, RILS creates a set of potential failed assertions and searches in parallel starting from

each possibility.

A missing assertion case is just as much of a problem for diagnosis as for monitoring. RILS could distinguish between diagnosis and monitoring by which module was most recently in use, but would have just as much difficulty recognizing the lack and determining a repair.

Failing to retrieve a relevant case for the diagnosis module could be due to two different failures. Like the monitoring module, diagnosis could miss a case due to indexing problems. Because diagnosis is performing highly targeted retrievals, however, this is a problem likely to be discovered as it happens. Alternatively, if assertion A should contain a link to a causally related assertion B, but lacks that link, then B could be overlooked. This problem must be detected by examining the cases in memory and their usage statistics. Retrieving an irrelevant case is, again, easy to detect.

Circumstances that indicate a problem with the diagnosis module include times when the diagnosis process fails to find any applicable repair. Alternatively, the diagnosis module may attempt to evaluate an assertion whose value depends on one that was overlooked or misapplied, and may be unable to complete its evaluation.

## The Repair Module

The repair component retrieves and uses repair cases which describe how to change the system to correct a diagnosed reasoning failure.

A missing repair case is, as for assertion cases, difficult to detect, unless a repair is referred to in an assertion case and does not exist in the case memory.

Failing to retrieve a relevant repair or retrieving an irrelevant one may be detected immediately, as the repair module performs very targeted case retrieval. It would be repaired, as above, by either altering the indexing of the repair case, or altering the index and retrieval methods of the system.

Misapplying a repair strategy is a difficult failure to detect. A repair could be executed, and might not correct the experienced failure. As in the example in the previous section, RILS may detect this type of failure if it faces a similar situation again and generates an identical repair.

## Conclusions

RILS demonstrates the power of case-based reasoning to serve as the central reasoning method of a system performing a wide range of reasoning tasks. By having one approach to reasoning, RILS has a fairly simple, powerful model of its reasoning processes. It uses that model to introspect about its reasoning process: learning from failures of its domain-level reasoning, detecting opportunities to reflectively introspect and, eventually, learn by improving its introspective reasoning processes.

A system that combines task reasoning with a reflective introspection capability should be able to respond with flexibility to a complex environment, by re-tooling its knowledge and processes at all levels of abstraction. This seems to be a talent which humans possess, as we

improve our reasoning strategies with experience. RILS provides one possible abstract model of this process.

Reflection permits a system to adapt itself to its surroundings, but poses a hazard if left unchecked. We have demonstrated here a “failure-driven” approach to controlling reflection: only when clear evidence exists that the introspective reasoning is flawed will RILS choose to reflect. Our future work on RILS will complete the reflective skills it has by giving it the tools to repair, not just diagnose, its introspective reasoning processes,

## References

- Arcos, J. & Plaza, E. (1993). A reflective architecture for integrated memory-based learning and reasoning. In Wess, S., Altoff, K., & Richter, M. (Eds.), *Topics in Case-Based Reasoning*. Springer-Verlag, Kaiserslautern, Germany.
- Chi, M. & Glaser, R. (1980). The measurement of expertise: a development of knowledge and skill as a basis for assessing achievement. In Baker, E. & Quellmalz, E. (Eds.), *Educational testing and evaluation: Design, analysis and policy*. Sage Publications, Beverly Hill, CA.
- Cox, M. (1996). *Introspective multistrategy learning: Constructing a learning strategy under reasoning failure*. Ph.D. thesis, College of Computing, Georgia Institute of Technology. Technical Report GIT-CC-96-06.
- Flavell, J., Friedrichs, A., & Hoyt, J. (1970). Developmental changes in memorization processes. *Cognitive Psychology*, 1, 324–340.
- Fox, S. & Leake, D. (1995). An introspective reasoning method for index refinement. In *Proceedings of 14th international Joint Conference on Artificial Intelligence*. IJCAI.
- Freed, M. & Collins, G. (1994). Adapting routines to improve task coordination. In *Proceedings of the 1994 Conference on AI Planning Systems*, pp. 255–259.
- Ibrahim, M. (1992). Reflection in object-oriented programming. *International Journal on Artificial Intelligence Tools*, 1(1), 117–136.
- Kreutzer, M., Leonard, M., & Flavell, J. (1975). An interview study of children’s knowledge about memory. *Monographs of the Society for Research in Child Development*, 40. (1, Serial No. 159).
- Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one’s own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77 (6).
- Oehlmann, R., Edwards, P., & Sleeman, D. (1994). Changing the viewpoint: re-indexing by introspective questioning. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 675–680. Lawrence Erlbaum Associates.
- Rosenbloom, P., Laird, J., & Newell, A. (1993). *Meta Levels in Soar*, Vol. I, chap. 26. The MIT Press.
- Stroulia, E. & Goel, A. (1995). Functional representation and reasoning in reflective systems. *Applied Artificial Intelligence: An International Journal, Special Issue on Functional Reasoning*, 9(1), 101–124.

# The Chinese Room: Just Say “No!”

Robert M. French

Quantitative Psychology and Cognitive Science

University of Liège

4000 Liège, Belgium

email: [rfrench@ulg.ac.be](mailto:rfrench@ulg.ac.be)

## Abstract

It is time to view John Searle’s Chinese Room thought experiment in a new light. The main focus of attention has always been on showing what is wrong (or right) with the argument, with the tacit assumption being that somehow there could be such a Room. In this article I argue that the debate should not focus on the question “If a person in the Room answered all the questions in perfect Chinese, while not understanding a word of Chinese, what would the implications of this be for strong AI?” Rather, the question should be, “Does the very idea of such a Room and a person in the Room who is able to answer questions in perfect Chinese while not understanding any Chinese make any sense at all?” And I believe that the answer, in parallel with recent arguments that claim that it would be impossible for a machine to pass the Turing Test unless it had experienced the world as we humans have, is no.

## Introduction

Alan Turing’s (1950) classic article on the Imitation Game provided an elegant operational definition of intelligence. His article is now exactly fifty years old and ranks, without question, as one of the most important scientific/philosophical papers of the twentieth century. The essence of the test proposed by Turing was that the ability to perfectly simulate unrestricted human conversation would constitute a sufficient criterion for intelligence. This way of defining intelligence, for better or for worse, was largely adopted as of the mid-1950’s, implicitly if not explicitly, as the overarching goal of the nascent field of artificial intelligence (AI).

Thirty years after Turing’s article appeared, John Searle (1980) put a new spin on Turing’s original arguments. He developed a thought experiment, now called “The Chinese Room,” which was a reformulation of Turing’s original test and, in so doing, produced what is arguably the second most widely read and hotly discussed paper in artificial intelligence. While Turing was optimistic about the possibility of creating intelligent programs in the foreseeable future, Searle concluded his article on precisely the opposite note: “...no [computer] program, by itself, is sufficient for intentionality.” In short, Searle purported to have shown that real (human-like) intelligence was impossible for any program implemented on a computer. In the present article I will begin by briefly presenting Searle’s well-known transformation of the

Turing’s Test. Unlike other critics of the Chinese Room argument, however, I will not take issue with Searle’s argument per se. Rather, I will focus on the argument’s central premise and will argue that the correct approach to the whole argument is simply to refuse to go beyond this premise, for it is, as I hope to show, untenable.

## The Chinese Room

Instead of Turing’s Imitation Game in which a computer in one room and a person in a separate room both attempt to convince an interrogator that they are human, Searle asks us to begin by imagining a closed room in which there is an English-speaker who knows no Chinese whatsoever. This room is full of symbolic rules specifying inputs and outputs, but, importantly, there are no translations in English to indicate to the person in the room the meaning of any Chinese symbol or string of symbols. A native Chinese person outside the room writes questions — *any questions* — in Chinese on a piece of paper and sends them into the room. The English-speaker receives each question inside the Room then matches the symbols in the question with symbols in the rule-base. (This does not have to be a direct table matching of the string of symbols in the question with symbols in the rule base, but can include any type of look-up program, regardless of its structural complexity.) The English-speaker is blindly led through the maze of rules to a string of symbols that constitutes an answer to the question. He copies this answer on a piece of paper and sends it out of the room. The Chinese person on the outside of the room would see a perfect response, even though the English-speaker understood no Chinese whatsoever. The Chinese person would therefore be fooled into believing that the person inside the room understood perfect Chinese.

Searle then compares the person in the room to a computer program and the symbolic rules that fill the room to the knowledge databases used by the computer program. In Searle’s thought experiment the person who is answering the questions in perfect written Chinese still has no knowledge of Chinese. Searle then applies the conclusion of his thought experiment to the general question of machine intelligence. He concludes that a computer program, however perfectly it managed to communicate in writing, thereby fooling all human questioners, would still not understand what it was writing, any more than the person in the Chinese Room



understood any Chinese. Ergo, computer programs capable of true understanding are impossible.

### Searle's Central Premise

But this reasoning is based on a central premise that needs close scrutiny.

Let us begin with a simple example. If someone began a line of reasoning thus: "Just for the sake of argument, let's assume that cows are as big as the moon," you would most likely reply, "Stop right there, I'm not interested in hearing the rest of your argument because cows are demonstrably NOT as big as the moon." You would be justified in not allowing the person to continue to his conclusions because, as logical as any of his subsequent reasoning might be, any conclusion arising from his absurd premise would be unjustified.

But when are we justified in accepting demonstrably false premises for the sake of argument? If a discussion began by supposing that the work week was 30 hours long, instead of 40, it would be ridiculous to reply, "But the work week is demonstrably NOT 30 hours long, therefore I am not interested in hearing the rest of your argument." On the other hand, if a discussion began by assuming that Lee Harvey Oswald was an ice-cream cone — however *logically* possible this might be — one would certainly be justified in evoking the I-don't-want-to-hear-anymore response. Space prevents us from attempting to delineate these two types of counterfactual premises, but suffice it to say that the mere logical possibility of a premise is not necessarily enough for it to serve as the basis of an argument in which we hope to derive truths about the real world, especially if we can demonstrate the nomological impossibility of the premise. Dennett (1996) makes a similar point regarding Davidson's (1986) Swampman argument.

In this light, let us consider the central premise on which Searle's argument hangs — namely, that there could be such a thing as a "Chinese Room" in which an English-only person *could actually fool* a native-Chinese questioner. I hope to show that this premise is no more plausible than the existence of lunar-sized cows and, as a result, we have no business allowing ourselves to be drawn into the rest of Searle's argument, any more than when we were asked to accept that all cows were the size of the moon.

Ironically, the arguments in the present paper *support* Searle's point that symbolic AI is not sufficient to produce human-like intelligence, but do so not by comparing the person in the Chinese Room to a computer program, but rather by showing that the Chinese Room itself would be an impossibility for a symbol-based AI paradigm.

## Subcognitive Questioning and the Turing Test

To understand why such a Room would be impossible, which would mean that the person in the Room could never fool the outside-the-Room questioner, we must look at an argument concerning the Turing Test first put forward by French (1988, 1990, 2000a). French's claim is that no machine that had not experienced life as we humans had could ever hope to pass the Turing Test. His demonstration involves showing just how hard it would be for a computer to consistently reply in a human-like manner to what he called "subcognitive" questions. Since Searle's Chinese Room argument is simply a reformulation of the Turing Test, we would expect to be able to apply these arguments to the Chinese Room as well, something which we will do this later in this paper.

It is important to spend a moment reviewing the nature and the power of "subcognitive" questions. These are questions that are explicitly designed to provide a window on low-level (i.e., unconscious) cognitive or physical structure. By "low-level cognitive structure", we mean the subconscious associative network in human minds that consists of highly overlapping activatable representations of experience (French, 1990). Creating these questions and, especially, gathering the answers to them require a bit of preparation on the part of the Interrogator who will be administering the Turing Test.

The Interrogator in the Turing Test (or the Questioner in the Chinese Room) begins by preparing a long list of these questions — the Subcognitive Question List. To get answers to these questions, she ventures out into an English-language population and selects a representative sample of individuals from that population. She asks each person surveyed all the questions on her Subcognitive Question List and records their answers. The questions along with the statistical range of answers to these questions will be the basis for her Human Subcognitive Profile. Here are some of the questions on her list (French, 1988, 1990).

### Questions using neologisms:

"On a scale of 0 (completely implausible) to 10 (completely plausible):

- Rate *Flugblogs* as a name Kellogg's would give to a new breakfast cereal.
- Rate *Flugblogs* as the name of start-up computer company
- Rate *Flugblogs* as the name of big, air-filled bags worn on the feet and used to walk across swamps.
- Rate *Flugly* as the name a child might give to a favorite teddy bear.
- Rate *Flugly* as the surname of a bank accountant in a W. C. Fields movie.

- Rate *Flugly* as the surname of a glamorous female movie star.

“Would you like it if someone called you a *trubhead*? (0= not at all, ..., 10 = very much)”

“Which word do you find prettier: *blutch* or *farfaletta*?”

Note that the words *flugbogs*, *flugly*, *trubhead*, *blutch* and *farfaletta* are made-up. They will not be found in any dictionary and, yet, because of the uncountable influences, experiences and associations of a lifetime of hearing and using English, we are able to make judgments about these neologisms. And, most importantly, while these judgments may vary between individuals, their variation is not random. For example, the average rating of *Flugly* as the surname of a glamorous actress will most certainly fall below the average rating of *Flugly* as the name for a child’s teddy bear. Why? Because English speakers, all of us, have grown up surrounded by roughly the same sea of sounds and associations that have gradually formed our impressions of the prettiness (or ugliness) of particular words or sounds. And while not all of these associations are identical, of course, they are similar enough to be able to make predictions about how, on average, English-speaking people will react to certain words and sounds. This is precisely why Hollywood movie moguls gave the name “Cary Grant” to a suave and handsome actor born “Archibald Alexander Leach” and why “Henry Deutchendorf, Jr.” was re-baptised “John Denver.”

Questions using categories:

- Rate *banana splits* as *medicine*.
- Rate *purses* as *weapons*.
- Rate *pens* as *weapons*.
- Rate *dry leaves* as *hiding places*.

No dictionary definition of “dry leaves” will include in its definition “hiding place,” and, yet, everyone who was ever a child where trees shed their leaves in the fall knows that that piles of dry leaves make wonderful hiding places. But how could this information, and an infinite amount of information just like it that is based on our having experienced the world in a particular way, ever be explicitly programmed into a computer?

Questions relying on human physical sensations:

- Does holding a gulp of Coca-Cola in your mouth feel more like having pins-and-needles in your foot or having cold water poured on your head?
- Put your palms together, fingers outstretched and pressed together. Fold down your two middle fingers till the middle knuckles touch. Move the other four pairs of fingers (i.e., your two index

fingers, your two thumbs, etc.). What happens to your other fingers? (Try it!)

We can imagine many more questions that would be designed to test not only for subcognitive associations, but for internal physical structure. These would include questions whose answers would arise, for example, from the spacing of a human’s eyes, would be the results of little self-experiments involving tactile sensations on their bodies or sensations after running in place, and so on.

People’s answers to subcognitive questions are the product of a lifetime of experiencing the world with our human bodies, our human behaviors (whether culturally or genetically engendered), our human desires and needs, etc. (See Harnad (1989) for a discussion of the closely related *symbol grounding problem*.)

I have asked people the question about Coca-Cola and pins-and-needles many times and they overwhelmingly respond that holding a soft-drink in their mouth feels more like having pins and needles in their foot than having cold water poured on them. Answering this question is dead easy for people who have a head and mouth, have drunk soft-drinks, have had cold water poured on their head, and have feet that occasionally fall asleep. But think of what it would take for a machine that had none of these to answer this question. How could the answer to this question be explicitly programmed into the machine? Perhaps (after reading this article) a programmer could put the question explicitly into a vast CYC-like computer database (Lenat & Guha, 1990), but there are infinitely many questions of this sort and to program them all in would be impossible. A program that could answer questions like these in a human-like enough manner to pass a Turing Test would have had to have experienced the world in a way that was very similar to the way in which we had experienced the world. This would mean, among many other things, that it would have to have a body very much like ours with hands like ours, with eyes where we had eyes, etc. For example, if an otherwise perfectly intelligent robot had its eyes on its knees, this would result in detectably non-human associations for such activities as, say, praying in church, falling when riding a bicycle, playing soccer, or wearing pants.

The moral of the story is that it doesn’t matter if we humans are confronted with made-up words or conceptual juxtapositions that never normally occur (e.g., *dry leaves* and *hiding place*), we can still respond and, moreover, our responses will show statistical regularities over the population. Thus, by surveying the population at large with an extensive set of these questions, we draw up a Human Subcognitive Profile for the population. It is precisely this subcognitive profile that could not be reproduced by a machine that had not experienced the world as the members of the sampled human population had. The Subcognitive Question List that was used to produce the Human

Subcognitive Profile gives the well-prepared Interrogator a sure-fire tool for eliminating machines from a Turing test in which humans are also participating. The Interrogator would come to the Turing Test and ask both candidates the questions on her Subcognitive Question List. The candidate most closely matching the average answer profile from the human population will be the human.

### The English Room

Now let us see how this technique can be gainfully applied to Searle's Chinese Room thought experiment. We will start by modifying Searle's original *Gedankenexperiment* by switching the languages around. This, of course, has no real bearing on the argument itself, but it will make our argument easier to follow. We will assume that inside the Room there is a Chinese person (let's call him Wu) who understands not a word of written English and outside the Room is a native speaker/writer of English (Sue). Sue sends into the Room questions written in English and Wu must produce the answers to these questions in English. Now, it turns out that Sue is not your average naive questioner, but has read many articles on the Turing Test, knows about subcognitive questions and is thoroughly familiar with John Searle's argument. She also suspects that the person inside the (English) Room might not actually be able to read English and she sets out to prove her hunch.

Sue will not only send into the Room questions like, "What is the capital of Cambodia?", "Who painted *The Mona Lisa*?" or "Can fleas fly?" but will also ask a large number of "subcognitive questions." Because the Room, like the computer in the Turing Test, had not experienced the world as we had and because it would be impossible to explicitly write down all of the rules necessary to answer subcognitive questions in general, the answers to the full range of subcognitive questions could not be contained in the lists of symbolic rules in the Room. Consequently, the person in the Room would be revealed not to speak English for exactly the same reason that the machine in the Turing Test would be revealed not to be a person.

Take the simple example of non-existent words like *blutch* or *trubhead*. These words are neologisms and would certainly be nowhere to be found in the symbolic rules in the English Room. Somehow, the Room would have to contain, in some symbolic form, information not only about all words, but also non-words as well. But the Room, if it is to be compared with a real computer, cannot be infinitely large, nor can we assume infinite fast search of the rule base (see Hofstadter & Dennett, 1981, for a discussion of this point). So, we have two closely related problems: First, and most crucially, *how* could the rules have gotten into the Room in the first place (a point that Searle simply ignores)? And secondly, the number of explicit

symbolic rules would require essentially an infinite amount of space. And while rooms in thought experiments can perhaps be infinitely large, the computers that they are compared to cannot be.

In other words, the moral of the story here, as it was for the machine trying to pass the Turing Test, is that no matter how many symbolic rules were in the English Room they would not be sufficient for someone who did not understand written English to fool a determined English questioner. And this is where the story should rightfully end. Searle has no business taking his argument any further — and, ironically, *he doesn't need to*, since the necessary inadequacy of an such a Room, regardless of how many symbolic rules it contains, proves his point about the impossibility of achieving artificial intelligence in a traditional symbol-based framework. So, when Searle asks us to accept that the English-only human in his Chinese Room could reply in perfect written Chinese to questions written in Chinese, we must say, "That's strictly impossible, so stop right there."

### Shift in Perception of the Turing Test

Let us once again return to the Turing Test to better understand the present argument.

It is easy to forget just how high the optimism once ran for the rapid achievement of artificial intelligence. In 1958 when computers were still in their infancy and even high-level programming languages had only just been invented, Simon and Newell, two of the founders of the field of artificial intelligence, wrote, "...there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until — in a visible future — the range of problems they can handle will be coextensive with the range to which the human mind has been applied." (Simon & Newell, 1958). Marvin Minsky, then head of the MIT AI Laboratory, wrote in 1967, "Within a generation the problem of creating 'artificial intelligence' will be substantially solved" (Minsky, 1967).

During this period of initial optimism, the vast majority of the authors writing about the Turing Test tacitly accepted Turing's premise that a machine might actually be able to be built that could pass the Test in the foreseeable future. The debate in the early days of AI, therefore, centered almost exclusively around the validity of Turing's operational definition of intelligence — namely, did passing the Turing Test constitute a sufficient condition for intelligence or did it not? But researchers' views on the possibility of achieving artificial intelligence shifted radically between the mid-1960's and the early 1980's. By 1982, for example, Minsky's position regarding achieving artificial intelligence had undergone a radical shift from one of unbounded optimism 15 years earlier to a far more sober assessment of the situation: "The AI

problem is one of the hardest ever undertaken by science” (Kolata, 1982). The perception of the Turing Test underwent a parallel shift. At least in part because of the great difficulties being experienced by AI, there was a growing realization of just how hard it would be for a machine to ever pass the Turing Test. Thus, instead of discussing whether or not a machine that had passed the Turing Test was really intelligent, the discussion shifted to the question of whether it would even be possible for any machine to pass such a test (Dennett, 1985; French, 1988, 1990; Crockett 1994; Harnad, 1989; for a review, see French, 2000b).

### **The Need for a Corresponding Shift in the Perception of the Chinese Room**

A shift in emphasis identical to the one that has occurred for the Turing Test is now needed for Searle’s Chinese Room thought experiment. Searle’s article was published in pre-connectionist 1980, when traditional symbolic AI was still the dominant paradigm in the field. Many of the major difficulties facing symbolic AI had come to light, but in 1980 there was still little emphasis on the “sub-symbolic” side of things.

But the growing difficulties that symbolic AI had in dealing with “sub-symbolic cognition” were responsible, at least in part, for the widespread appeal of the connectionist movement of the mid-1980’s. While several of the commentaries of Searle’s original article (Searle, 1980) briefly touch on the difficulties involved in actually creating a Chinese Room, none of them focus outright on the impossibility of the Chinese Room as described by Searle and reject the rest of the argument because of its impossible premise. But this rejection corresponds precisely to rejecting the idea that a machine (that had not experienced the world as we humans have) could ever pass the Turing Test, an idea that many people now accept. We are arguing for a parallel shift in emphasis for the Chinese Room *Gedankenexperiment*.

### **Can the “Robot Reply” Help?**

It is necessary to explore for a moment the possibility that one could somehow fill the Chinese Room with all of the appropriate rules that would allow the non-Chinese-reading person to fool a no-holds-barred Chinese questioner. Where could rules come from that would allow the person in the Chinese Room to answer all of the in-coming questions in Chinese perfectly? One possible reply is a version of the Robot Reply (Searle, 1980). Since the rules couldn’t have been symbolic and couldn’t have been explicitly programmed in for the reasons outlined above (also see French, 1988, 1990), perhaps they could have been the product of a Robot that had experienced and interacted with the world as we humans would have, all the while

generating rules that would be put in the Chinese Room.

This is much closer to what would be required to have the appropriate “rules,” but still leaves open the question of how you could ever come up with such a Robot. The Robot would have to be able to interact seamlessly with the world, exactly as a Chinese person would, in order to have been able to produce all the “rules” (high-level and subcognitive) that would later allow the person in the Room to fool the Well-Prepared Questioner. But then we are back to square one, for creating such a robot amounts to creating a robot that would pass the Turing Test.

### **The Chinese Room: a Simple Refutation**

It must be reiterated that when Searle is attacking the “strong AI” claim that machines processing strings of symbols are capable of doing what we humans call thinking, he is explicitly talking about programs implemented on *computers*. It is important not to ignore the fact, as some authors unfortunately have (e.g., Block, 1981), that computers are *real machines* of finite size and speed; they have neither infinite storage capacity nor infinite processing speed.

Now consider the standard Chinese Room, i.e., the one in which the person inside the Room has no knowledge of Chinese and the Questioner outside the Room is Chinese. Assume that the last character of the following question is distorted in an extremely phallic way, but in a way that nonetheless leaves the character completely readable to any reader of Chinese: “Would the last character of this sentence embarrass a very shy young woman?” In order to answer this question correctly — a trivially easy task for anyone who actually reads Chinese — the Chinese Room would have to contain rules that would not only allow the person to respond perfectly to all strings of Chinese characters that formed comprehensible questions, but also to the infinitely many possible legible *distortions* of those strings of characters. Combinatorial explosion brings the house down around the Chinese Room. (Remember, we are talking about real computers that can store a finite amount of information and must retrieve it in a finite amount of time.)

One might be tempted to reply, “The solution is to eliminate all distortions. Only standard fonts of Chinese characters are permitted.” But, of course, there are hundreds, probably thousands, of different fonts of characters in Chinese (Hofstadter, 1985) and it is completely unclear what would constitute “standard fonts.” In any event, one can sidestep even this problem.

Consider an equivalent situation in English. It makes perfect sense to ask, “Which letter could be most easily distorted to look like a cloud: an ‘O’ or an ‘X’?” An overwhelming majority of people would, of course, reply “O”, even though clouds, superficially and

theoretically, have virtually nothing in common with the letter “O”. But how could the symbolic rules in Searle’s Room possibly serve to answer this perfectly legitimate question? A theory of clouds contained in the rules certainly wouldn’t be of any help, because that would be about storms, wind, rain and meteorology. A theory or database of cloud forms would be of scant help either, since clouds are anything but two dimensional, much less round. Perhaps only if the machine/Room had grown up scrawling vaguely circular shapes on paper and calling them clouds in kindergarten and elementary school, then maybe it would be able to answer this question. But short of having had that experience, I see little hope of an a priori theory of correspondence between clouds and letters that would be of any help.

### Conclusion

The time has come to view John Searle’s Chinese Room thought experiment in a new light. Up until now, the main focus of attention has been on showing what is wrong (or right) with the argument, with the tacit assumption being that somehow there could be such a Room. This parallels the first forty years of discussions on the Turing Test, where virtually all discussion centered on the sufficiency of the Test as a criterion for machine intelligence, rather than whether any machine could ever actually pass it. However, as the overwhelming difficulties of AI gradually became apparent, the debate on the Turing Test shifted to whether or not any machine that had not experience the world as we had could ever actually pass the Turing Test. It is time for an equivalent shift in attention for Searle’s Chinese Room. The question should not be, “If a person in the Room answered all the questions in perfect Chinese, while not understanding a word of Chinese, what would the implications of this be for strong AI?” Rather, the question should be, “Does the very idea of such a Room and a person actually be able to answer questions in perfect Chinese while not understanding any Chinese make any sense at all?” And I believe that the answer, in parallel with the impossibility of a machine passing the Turing Test, is no.

### Acknowledgments

The present paper was supported in part by research grant IUAP P4/19 from the Belgian government.

### References

Block, N. (1981) Psychologism and behaviourism. *Philosophical Review*, 90, 5-43  
 Crockett, L. (1994) *The Turing Test and the Frame Problem: AI’s Mistaken Understanding of Intelligence*. Ablex

Davidson, D. (1990) Turing’s test. In Karim A. Said et al. (eds.), *Modelling the Mind*. Oxford University Press, 1-11.  
 Davidson, D. (1986). Knowing One’s Own Mind, *Proceedings and Addresses of the American Philosophical Association*, 60 (3), p.443.  
 Dennett, D. (1985) Can machines think? In *How We Know*. (ed.) M. Shafto. Harper & Row  
 Dennett, D. (1996). Cow-Sharks, Magnets, and Swampman. *Mind and Language*, 11(1):76-77.  
 French, R. M. (1988). Subcognitive Probing: Hard Questions for the Turing Test. *Proceedings of the Tenth Annual Cognitive Science Society Conference*, Hillsdale, NJ: LEA. 361-367.  
 French, R. M. (1990). Subcognition and the Limits of the Turing Test. *Mind*, 99(393), 53-65. Reprinted in: P. Millican & A. Clark (eds.). *Machines and Thought: The Legacy of Alan Turing* Oxford, UK: Clarendon Press, 1996.  
 French, R. M. (2000a). Peeking Behind the Screen: The Unsuspected Power of the Standard Turing Test. *Journal of Experimental and Theoretical Artificial Intelligence*. (in press).  
 French, R. M. (2000b). The Turing Test: The First Fifty Years. *Trends in Cognitive Sciences*, 4(3), 115-122.  
 Harnad, S. (1989) Minds, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence*, 1, 5-25  
 Hofstadter, D. (1985). Variations on a Theme as the Crux of Creativity. In *Metamagical Themas*. New York, NY: Basic Books. p. 244.  
 Hofstadter, D. & Dennett, D. (1981). *The Mind’s I*. New York, NY: Basic Books.  
 Kolata, G. (1982) How can computers get common sense? *Science*, 217, p. 1237  
 Lenat, D. & Guha, R. (1990). *Building Large Knowledge-Based Systems*. Reading, Mass.: Addison-Wesley.  
 Minsky, M. (1967) *Computation: Finite and Infinite Machines*. Prentice-Hall, p. 2  
 Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3, 414-424.  
 Simon, H. and Newell, A. (1958) Heuristic problem solving: The next advance in operations research. *Operations Research*, 6

# The Influence of Source and Cost of Information Access on Correct and Errorful Interactive Behavior

Wayne D. Gray & Wai-Tat Fu  
Human Factors & Applied Cognition  
George Mason University  
Fairfax, VA 22030 USA  
+1 703 993 1357  
gray@gmu.edu

## ABSTRACT

Routine interactive behavior reveals patterns of interaction among the cognitive, perceptual, and motor elements of embodied cognition and the task and artifact used to perform the task. Such interactions are difficult to study, in part, because they require collecting a large quantity of mostly correct behavior. The current study varied conditions that were designed to affect the ease and likelihood that information would be stored in-the-world versus in-the-head. The data are examined to determine how subtle differences in the source and cost of information access may lead to different patterns of correct and errorful behavior.

## INTRODUCTION

Interactive behavior emerges out of the constraints and opportunities provided by the interaction of embodied cognition (Kieras & Meyer, 1997) with task goals and the artifact used to perform the task (the ETA,  $\eta$ , triad). The interactions among the components of the ETA triad that determine interactive behavior may be extremely subtle with small changes in costs leading to large shifts in performance. For example, changing information gathering from an eye movement to a mouse movement influenced the decision-making strategies adopted in a classic decision-making paradigm (Lohse & Johnson, 1996). When the cost of making a move in solving simple puzzles increased from one keystroke to several (O'Hara & Payne, 1998; O'Hara & Payne, 1999; Svendsen, 1991) the strategy used to solve the puzzles shifted from one in which search was "reactive and display-based" to one in which search was more plan-based. The subtlety of change in response to minor variations in interface design should not be underestimated. For example, by increasing the cost of information acquisition from a simple saccade to a head movement, Ballard (Ballard, Hayhoe, & Pelz, 1995) induced a shift from a memoryless strategy to one that required holding information in working memory.

In the work reported here, we were interested in how the requirement to access information *in-the-world* versus *in-the-head* would influence routine interactive behavior. Almost by definition, most routine interactive behaviors are successfully executed. Hence, our focus is not on outcome measures of success, but on process measures of performance. Two important sources of clues regarding

process are patterns of information access and errors that are made, detected, and corrected during performance.

Unfortunately, errors in routine interactive behavior are relatively rare and collecting enough such errors to discover underlying patterns requires collecting a large quantity of correct interactive behavior. For example, Gray (in press) found only 96 keypress errors in a data set of 1,946 keypresses collected from 9 people as they programmed 56 shows on a simulated VCR.<sup>1</sup> For this reason, we collected massive amounts of data under a variety of conditions that were designed to vary the ease and likelihood that show information would be stored in-the-world versus in-the-head. The raw data were analyzed to yield three categories of information; patterns of information access during performance, types of erroneous goals attempted (*push errors*), and correct goals that were abandoned prematurely (*premature pops*). These categories were then interrogated to determine how subtle differences in information access may lead to different patterns of correct and errorful behavior.

The next section introduces the model and the approach on which the determination and classification of errors was based. We then present the methods and procedures used in the current study. The empirical results are discussed in three sections. The first provides an overview of performance, the second discusses the fit of the data to model, while the third presents error data. We conclude with a summary and discussion of how varying the cost of information access during routine performance influences correct as well as errorful behavior.

---

<sup>1</sup> Participants used a mouse to interact with the simulation. The actual VCR was operated by pressing and sliding various physical buttons. Hence, neither the simulated nor the actual VCR required key presses. Few task analysis methods analyze behavior down to the level of physical actions (see, e.g., the survey of task analysis methods reported by Kirwan & Ainsworth, 1992). Throughout this paper, our use of the terms "keypress" reflects the fact that by including mouse clicks (or button presses) in the analysis, the task analysis is at the "keystroke level." This usage of the term "keystroke level" follows the distinction made by Card, Moran, and Newell (1983).

## CONSTRAINED INTERACTIVE BEHAVIOR IN UNDERCONSTRAINED INTERFACES

Task goals for programming a VCR include setting a program's day-of-week, start time, channel, and end time (see Figure 1). Unfortunately, programming an actual VCR entails mapping these simple task goals into a variety of device specific goals. The result is a task-to-device rule hierarchy such as is shown in Figure 2.

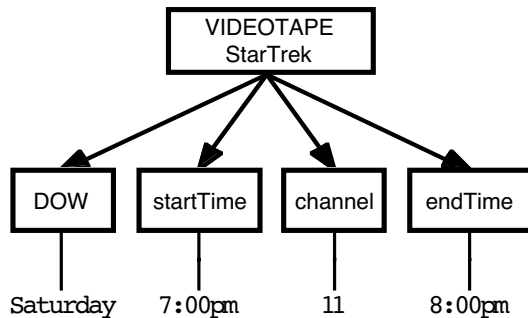


Figure 1: Task goals for programming a VCR.

The controversial part of what is being asserted is not that there is such a mapping, but that, in many cases, there is one least-effort mapping that, if left to themselves, most users will discover and use. If this least-effort mapping is taught, most users will conform to it despite a plethora of alternatives. The task-to-device rule hierarchy is encouraged, not mandated, by soft constraints derived from principles of cognitive least-effort (described in Gray, in press).

For many interactive devices, the sequence and methods of operation are highly constrained by design. For example, if your task goal is to take \$100 out of your checking account using an ATM, you must find an ATM; insert your card; key in your pin number; press fast cash; take the money; and then take the card. For any one ATM, there is not much variability in the set of methods or their sequence.

In contrast, if you are programming the VCR simulated in our study, the device does not prevent you from clicking on the start mode button, setting the start hour, clicking on the end mode button, setting the end hour, clicking on start mode button (again), setting the channel to 10, setting the day of week to Saturday, going back and setting the channel to 11, clicking on the clock set mode button, clicking on PROG REC, clicking on end mode (again), setting the 10min, setting the min, clicking on start mode (yet again), setting the 10min, setting the min, and finally, clicking on the clock set mode button (again).

Although somebody could program the VCR in this way, in fact, nobody does. In the study reported by Gray (in press), out of 9 participants who were not taught how to program the VCR, but discovered the methods by themselves, seven adopted the task-to-device rule

hierarchy of Figure 2 and two adopted minor variants. In the studies reported below, of the 72 participants shown Figure 2 as the experimenter programmed the first show, all but two used the task-to-device rule hierarchy to program the next four shows. Although extreme variation was possible, little variation was found.

The task-to-device rule hierarchy shown in Figure 2 was derived (Gray, in press) from three sources. The first was a simple task analysis of the methods available for programming shows on the simulated VCR. The second was an analysis of participant behavior during the instructionless learning phase of the study. The third was the analyses of the unsuccessful trials – those that ended without the VCR being successfully programmed. The resultant task-to-device rule hierarchy was used to analyze the 56 trials which were successfully programmed. By definition, any errors made on these *okay* trials were detected and corrected by the participants before telling the experimenter that they were done programming the VCR.

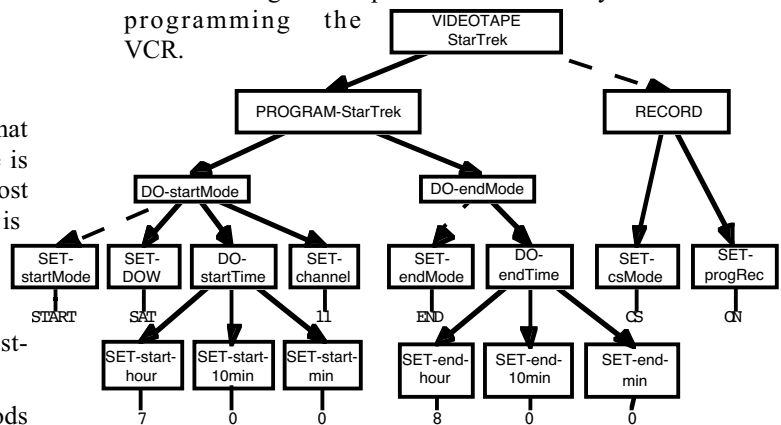


Figure 2: A mapping of the task goals from Figure 1 onto the device. This task-to-device rule hierarchy is largely determined by soft constraints. (Subgoals are represented by boxed nodes. Leaf nodes are unboxed and may represent multiple keystrokes. The dashed line leading from DO-startMode and DO-endMode indicate that subgoals SET-startMode and SET-endMode must be performed before the others. Contrariwise, the dashed line from VIDEOTAPE to RECORD indicates that RECORD must be performed last. With those three exceptions, the subgoals of a goal may be performed in any order.)

## EXPERIMENT

The current study used a new simulation of the VCR task adopted by Gray (in press). One of our goals for the current study was to determine whether new groups of participants in slightly different task conditions would conform to the task-to-device rule hierarchy shown in Figure 2. Another goal was to verify and extend the error taxonomy.

Although these goals are important, they are not the main goals of this paper. Rather, our main goal is to explore how correct as well as errorful interactive behavior is

affected by changing the cost of information access. For the *control* group, the show's start time, end time, day-of-week, and channel were clearly visible to participants.

The *gray-box* condition was designed to increase the effort required to obtain show information. For the control condition, information access required an eye movement to the show information window. In contrast, for the gray-box condition, the labels in the show information window were visible but the fields were covered by gray boxes. For example, to see the channel field, the participant had to move the cursor to and click on the gray box covering that field. The value stayed visible as long as the cursor remained in the field.

The *memory-test* condition encouraged the storage of show information in long-term memory. For each trial, clicking on the START button removed the show information window and opened a memory test window. The memory test required the participant to select the show's start and end hour, 10min, min, as well as day-of-week and channel from a series of pop-up menus. Prior to programming the show, the participant iterated between the show information window and the memory test until the test was passed.

When the VCR was being programmed, we encouraged the memory-test condition to retrieve show information from memory by discouraging the use of the show information window. As per the gray-box condition, the fields of the information window were covered by gray boxes. In addition, moving the cursor out of the VCR window caused the VCR to be covered by a black box. The black box stayed until the participant moved the cursor back to and clicked on the VCR window. Hence, for the memory-test condition, when a participant moved to and clicked on a gray box, the corresponding setting of the VCR (indeed, all settings of the VCR) was covered by the black box.

## Method

The experiment used VCR 2.0, a simulation of a commercial VCR built in Macintosh Common Lisp. All keypresses on any button object in VCR 2.0 were time stamped to the nearest tick (16.667 msec) and saved to a log file along with a complete record of the information in the VCR's displays (e.g., mode, time, day-of-week, channel, and so on).

## Participants

Sixty-four George Mason University undergraduates participated in the experiment for course credit. Participants were randomly assigned to conditions and were run individually. Each session took approximately 30 min.

## Procedure

The study began with the task-to-device rule hierarchy (Figure 2) in front of the participant. The experimenter programmed the first trial of show-0. As the show was programmed, the experimenter pointed to the figure, relating each step of programming to a node in the figure. After the first trial, the experimenter watched as the participant programmed show-0 to criterion. At that point, the experimenter left the room while the participant programmed shows 1 through 4 to the criterion of two successive correct trials. (As show-0 was an instruction and practice show, it is excluded from the analyses reported below.)

For all conditions each trial began with the VCR covered by a black box and a clearly visible information window that contained the current show's name, start time, end time, day of week, and channel. This information could be freely studied before the trial began. The information window also contained the START button. Clicking on the START button began the trial, changed the label from START to STOP, and either removed the black box that had covered the VCR (for control and gray-box) or opened the memory test window (for the memory-test condition).

At the end of each trial, the participant was given feedback as to how long the trial took and as to whether the show had been programmed correctly. If the show was not programmed correctly, the participant was provided feedback on the first error that the software found. The order in which errors were checked was: clock time, start time, end time, day of week, channel, and program record.

## OVERVIEW OF PERFORMANCE

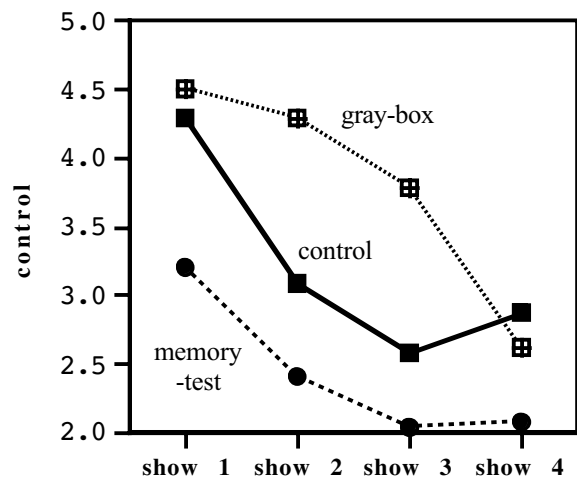


Figure 3: Mean trials to criterion.

## Trials-to-criterion

A two-way analysis of variance (ANOVA) was conducted on the number of trials to reach the criterion of two successive correct shows. Condition (control, gray-box,



memory-test) was a between-subjects factor and show (1-4) was within-subjects. The main effect of condition was significant,  $F(2, 69) = 4.478$ ,  $p = .015$  ( $MSE = 10.035$ ), as was the main effect of show,  $F(3, 207) = 5.896$ ,  $p = .0007$  ( $MSE = 5.053$ ). The interaction of condition by show was not significant ( $F < 1$ ) (see Figure 3).

Planned comparisons by condition yielded a significant difference between gray-box and memory-test ( $p = .0002$ ) as well as between control and memory-test ( $p = .0370$ ). The difference between the control and gray-box condition was not significant.

### Checking the Show Information Window

In all conditions, participants were free to study the show information before each trial began. During the trial we expected the greatest reliance on the show information window for the control condition, less reliance for the gray-box condition, and the least reliance for the memory-test condition. Unfortunately, as we did not collect eye movement data, any discussion of what the control group did is speculation. However, we do have data that supports our interpretation of the tradeoff between information in-the-head versus in-the-world for the other two conditions.

For trials that were successfully programmed (for which all errors were detected and corrected before the participant clicked on the STOP button), the gray-box condition checked information 293 times (a mean of 1.31 checks per show). In contrast, participants in the memory-test condition checked an information field 10 times (0.05 checks per participant per show). This contrast suggests that the memory-test group almost exclusively relied on memory as their source of show information.

For the gray-box condition, 149 of the information checks were made immediately prior to the use of the information (e.g., checking the day of week field and then setting day of week). In contrast, only 33 checks were made on an information field immediately after the corresponding VCR display was set.

These patterns of checking suggest that the gray-box participants did not memorize show information to the degree forced on the memory-test condition. However, the low number of information checks per show (a mean of 1.31 fields checked per show) suggests that the perceptual-motor strategy was the backup strategy, not the primary strategy for this group. Furthermore, the 149:33 (or 4.5 to 1) disparity between information acquisition checks versus information verification checks suggests a trust in working memory that the trials-to-criterion data indicates was not justified.

These data are consistent with the notion that the cognitive system minimizes local effort, not necessarily total effort (see also Gray, in press). For the gray-box condition, the failure to verify saved several seconds worth of effort during a good trial, but may have resulted

in more trials ending in error and, when compared to the memory-test condition, more trials needed to reach criterion. A similar conclusion is suggested by some of the error data that we review below.

### FIT OF DATA TO MODEL

A goal and subgoal analysis was conducted on trials that ended successfully. This restriction meant that any errors made during the trial had to be detected and corrected before the participant pressed the STOP button.

For these analyses, ACT-PRO (Fu & Gray, 1999) was used to parse the log file into goals, subgoals, and operators. Each deviation from the task-to-device rule hierarchy shown in Figure 2 was noted and classified by ACT-PRO. (The classification categories used here are an expansion of those reported by Gray, in press).

Over the course of the study 36,877 keypresses were collected. ACT-PRO parsed these into 12,704 goals and subgoals. Of this number, 98.4% (12,560) are goals that are captured by the task-to-device rule hierarchy.

Of the uncaptured goals and subgoals, 56 can be readily interpreted as the participant returning to a mode to double-check a setting. These additions increase the percentage of goals and subgoals accounted for to 98.8%.

The remaining 148 goals can be examined to determine if they represent errors or are simply alternatives to the task-to-device rule hierarchy used by the model. Of these potential errors, 16 represented alternative ways of correctly programming the VCR. These alternatives were manifested by five participants. Only two of these five participants used the alternative on a majority of trials. Hence, although there may be hundreds of ways of segmenting and sequencing the task of programming this VCR, the model shown in Figure 2 accounts for the vast majority of correct behavior shown by the overwhelming majority of participants.

### ACCOUNTING FOR ERROR

The taxonomy developed by Gray (in press) relied on model-tracing (Anderson, 1993) to identify deviations from the task-to-device rule hierarchy as *push errors* or *pop errors*. Any key that is pressed at a time or place where the model would not press it is a push error. Any goal or subgoal that is abandoned, or popped, before the model would end it is a pop error.

#### Push Errors

As discussed above, ACT-PRO classified 148 goal pushes as violations of the model's task-to-device rule hierarchy. After we subtract those behaviors that can be interpreted as alternative rule-hierarchies we are left with a data set of 132 push errors. In this paper, space constraints force us to limit our discussion to the 31 erroneous attempts to

increment rather than decrement (or vice versa) the channel setting.

Except for channel, each of the other settings had only one button. For day-of-week, hour, 10min, or min this button would only increment, never decrement the setting. In contrast, channel had two buttons; one to increase the displayed setting and one to decrease it. Hence, whereas if an erroneous attempt to decrement the day-of-week, hour, 10min, or min, was detected and corrected by the participant, it would have gone unnoted by the experimenter. In contrast, any goal to decrement the displayed channel setting when it should have been incremented (or vice versa) would be obvious from the log file. (Note that the target channel setting was higher than the default setting for two shows and lower than the default for the other two shows.)

An ANOVA of errors by conditions for incrementing versus decrementing the channel revealed a marginally significant effect,  $F(2, 69) = 2.787, p = .069, MSE = .683$ . The mean per trial error rate was higher for memory-test (0.750) than for gray-box (0.333) and lowest for control (0.208). Planned comparisons showed that the difference between memory-test and control was significant ( $p = .027$ ) while the difference between memory-test and gray-box was marginally significant ( $p = .087$ ).

While programming, participants in the memory-test condition checked show information a total of 10 times. The reliance on information in-the-head versus in-the-world resulted in an increase in errors. However, the information was well-learned and participants soon retrieved the correct information and set the channel to the correct setting. The transient nature of this error suggests a momentary fluctuation in strength of the memory trace due to noise (Altmann & Gray, 1999; Anderson & Lebière, 1998).

### Pop Errors

By the analysis introduced by Gray (in press), not only can pushing a goal be an error, but popping can be errorful as well. Popping a goal before its target setting has been reached is a *premature pop*. The data set collected by Gray (in press) was too small to distinguish between various types of premature pops. However, the 182 premature pops collected in the current study is an order of magnitude larger than that previously obtained. This set permits us to distinguish between three types of premature pops.

Local premature pops (pp-local) entail beginning to program a VCR setting but stopping before the target setting is achieved. For example, if the target day-of-week is Saturday and the current day-of-week is Tuesday, pressing the DOW key twice and then going off and doing something else would be classified as a pp-local. Time premature pops (pp-time) entail completing one or two of the DO-startTime or DO-endTime subgoals (see Figure 2)

but abandoning the goal before the remaining subgoals are completed. Similarly, mode premature pops (pp-mode) entail popping the DO-startMode or Do-endMode goal before all of their subgoals are completed.

Across the three types of premature pops a repeated measures ANOVA showed no main effect of condition ( $F < 1$ ), a significant effect for type of premature pop [ $F(2, 138) = 12.868, p < .0001, MSE = .041$ ] as well as a significant interaction of condition by type [ $F(4, 138) = 2.989, p = .021$ ]. As Figure 4 shows, the gray-box condition made the most pp-local errors with the memory-test condition making the least. This pattern was reversed for pp-mode errors.

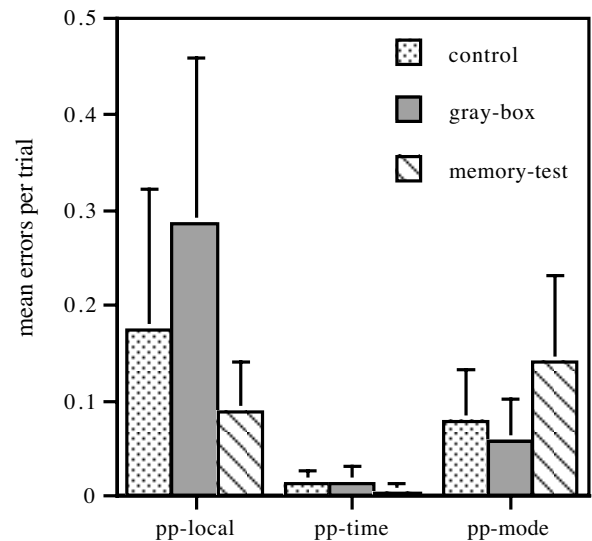


Figure 4: Premature pops by condition. (Error bars show the 95% confidence interval of the SEM.)

The higher pp-local error rate for gray-box is consistent with their pattern of fewer checks to verify show information. These errors – which were caught – as well as the errors that were not caught (i.e., those that led to greater trials-to-criterion for this group) may have resulted from the low rate of verification shown by this group.

Likewise, the higher rate of mode errors for memory-test may be the result of their reliance on memory. Although the gray boxes covered up the values of the information fields, they did not cover the labels for those fields. Hence, the labels may have served as a type of goal posting. The control and gray-box conditions would have been reminded of the goals for the current mode every time they glanced at the show information window.

### SUMMARY AND CONCLUSIONS

The most striking aspect of the between group differences in errors and performance is that all were avoidable. All performance differences can be traced to differences in willingness to either memorize or visually access show information. For each trial, the memory-test group had

quick and reliable access to show information in memory. The other groups made more undiscovered errors that resulted in more trials-to-criterion. Apparently verification is lower cost – and hence more likely – if based on knowledge in-the-head rather than accessing knowledge in-the-world.

On trials for which any error made was eventually detected and corrected, we found an interaction between group and type of premature pop. The gray-box condition was more likely to abandon the current key (pp-local) before completing a setting, whereas the memory-test condition was more likely to switch modes before all subgoals were completed (pp-mode). The pattern for pp-local errors is consistent with that for trials-to-criterion. In both cases, errors were made because the gray-box group was unwilling to invest in the time and effort needed to obtain reliable information.

Our interpretation of pp-mode errors suggested an advantage to relying on information in-the-world rather than in-the-head. Both the control and gray-box conditions accessed the show information throughout performance. In addition to obtaining the value of the information fields, accessing the show information window may have served as a type of goal posting to remind participants what settings they had programmed and what remained to be done. In contrast, the memory-test condition would have had to keep a corresponding checklist in-the-head. Unlike the show information that they memorized, the state of this mental checklist was dynamic and changed throughout task performance.

We interpreted the push error that we analyzed as evidence for fluctuations in the strength of items encoded in long-term memory. The fact that the misretrieved settings were detected and corrected without recourse to the show information window is consistent with the ACT-R assumption of transient fluctuations in strength (Altmann & Gray, 1999; Anderson & Lebière, 1998).

The study of routine interactive behavior is not itself routine. To study how small changes in artifact design affect performance, massive amounts of correct behavior must be collected. The analysis of routine interactive behavior enhances our understanding of how the cognitive, perceptual, and motor elements of embodied cognition interact with task and artifact to affect correct and errorful performance. This report suggests that small changes in the cost of information access may result in differences in the trials needed to reach criterion and the patterns of errors made.

## ACKNOWLEDGEMENTS

The work reported was supported by a grant from the National Science Foundation (IRI-9618833) as well as by the Air Force Office of Scientific Research AFOSR#F49620-97-1-0353.

## REFERENCES

- Altmann, E. M., & Gray, W. D. (1999). Preparing to forget: Memory and functional decay in serial attention. *Manuscript submitted for publication.*
- Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., & Lebière, C. (Eds.). (1998). *Atomic components of thought*. Hillsdale, NJ: Erlbaum.
- Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 7(1), 66-80.
- Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Fu, W.-T., & Gray, W. D. (1999). ACT-PRO: Action protocol tracer -- a tool for analyzing simple, rule-based tasks. Proceedings of the *Sixth ACT-R Workshop* (pp. ). Fairfax, VA: ARCH Lab.
- Gray, W. D. (in press). The nature and processing of errors in interactive behavior. *Cognitive Science*.
- Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction*, 12(4), 391-438.
- Kirwan, B., & Ainsworth, L. K. (Eds.). (1992). *A guide to task analysis*. Washington, DC: Taylor & Francis.
- Lohse, G. L., & Johnson, E. J. (1996). A comparison of two process tracing methods for choice tasks. *Organizational Behavior and Human Decision Processes*, 68(1), 28-43.
- O'Hara, K. P., & Payne, S. J. (1998). The effects of operator implementation cost on planfulness of problem solving and learning. *Cognitive Psychology*, 35, 34-70.
- O'Hara, K. P., & Payne, S. J. (1999). Planning and the user interface: The effects of lockout time and error recovery cost. *International Journal of Human-Computer Studies*, 50(1), 41-59.
- Svendsen, G. B. (1991). The influence of interface style on problem solving. *International Journal of Man-Machine Studies*, 35(3), 379-397.

# Practices of Questioning and Explaining in Learning to Model

**James G. Greeno (greeno@csl.stanford.edu)**

School of Education, Stanford University  
485 Lasuen Mall, Stanford CA 94305 USA

**Melissa C. Sommerfeld (msomme@leland.stanford.edu)**

Cognition and Learning Laboratory, Stanford University  
CERAS 105 Stanford CA 94305 USA

**Muffie Wiebe (wiebe@leland.stanford.edu)**

Cognition and Learning Laboratory, Stanford University  
CERAS 105 Stanford CA 94305 USA

## Abstract

Conceptual learning in mathematics involves more than getting to the right answer. Recent efforts in math education reform have focused on having students work in groups on open-ended projects that are based in realistic contexts. We extend previous analyses with hypotheses about conceptual and interactional aspects of understanding and solving problems by groups. The conceptual hypothesis focuses on integration of information in the group's *situation model* and its *problem model*. The interactional hypotheses involve patterns of interaction that make it easy or hard for the group to open up a discussion of assumptions in its reasoning, and that make the group accountable to a wider audience for explaining relations between the situations and mathematical operations involved in their solutions. Regarding educational practice, these findings highlight a way in which student groups must coordinate their conceptual and interactional work to arrive at satisfactory solutions to the problems posed. The present study suggests the importance of students in these environments not only connecting the contextual situation and the attending mathematics, but also reconsidering the situation in light of their new mathematical understanding (bringing the mathematics back into play in their understanding of the situation). Interactional patterns in a group make this relatively easier or harder, and this must be accounted for in implementing new curricula and conducting teacher education.

## Introduction

When a group works together on a problem involving mathematics, how does that work get done? How does the group arrive at its understanding of the problem on which they are working? How do they go about conducting the work? What happens when someone questions what another member of the group is doing or proposing to do? The person questioned may offer an explanation that justifies the claim or action. Alternatively, the group might collectively take up the question, and construct a new understanding. Or the question could be ignored, deflected, or dismissed.

When students are involved in solving open-ended mathematics problems without one correct answer, it's not immediately obvious when an error has been made. So when a mistake *is* recognized by a member of a group, how does that happen? This paper reports results from an analysis of a student group engaged in a mathematical

modeling unit. We present two episodes in which the group's problem solving involved mistaken assumptions. In one episode the group identified and corrected the mistake, but in the other they did not. In this paper we illustrate how a mistake was recognized and resolved, and how that differed from more common instances in which mistakes are not corrected. Implications for cognitive theory and for the design of learning environments are discussed.

## An Interactional Aspect of Reasoning

The interactional aspect of activity that we focus on involves explanatory practices. We develop our analysis using the schema of conversational contributions provided by Clark and Schaefer (1989) and adapted for analyzing discourse in problem solving by Greeno and Engle (1995). In this schema, each contribution to the process of understanding and solving a problem includes a presentation of information or action and an acceptance, resulting in grounding the contribution in the participants' mutual understanding.

In much ordinary discourse, explanations occur mainly when someone questions or disagrees with something that someone else says or does (McLaughlin, Cocy & Reed, 1992). In Clark and Schaefer's scheme, occasions for explanation often arise when one participant presents some information and another other participant responds with a question, a challenge, or an alternative. When this happens, the group can take up the question, challenge, or the suggested alternatives. This "taking up" involves a kind of negotiation in which the question or challenge may be resolved, one of the alternatives may be chosen, or the group may leave the issue with the understanding that their uncertainty or disagreement remains. In such a negotiation, explanations may occur frequently between group members.

The discourse patterns of different groups or of groups in different situations vary in how open they are to presenting and considering questions, challenges, or alternatives. A presentation provides a possible continuation of a trajectory in the activity. A question, challenge, or alternative proposal presents a potential diversion from that trajectory, and the resolution may bring about a change in the group's direction. It is reasonable—perhaps necessary—for groups to maintain some level of inertia in their interactions in order to enable them to function productively. The amount of that inertia can vary depending on which participant has the floor. It is easier to challenge and

question some participants than others, and some participants are more likely than others to discuss an issue rather than to forge ahead, ignoring the intrusion, or to simply give a justificatory explanation.

Mathematics educators such as Lampert (e.g., 1990) and Cobb (e.g., Cobb Yackel & Wood, 1992) have emphasized the importance of students' developing practices of explanatory discourse that support valid reasoning and understanding in mathematics. Cobb and Yackel (1996) distinguished between *social norms*, *sociomathematical norms*, and *mathematical practices* in the activity patterns of mathematics classrooms. Social norms, or *participation structures* (Erickson, 1986; Lampert, 1990), include the extent to which participants expect each other to provide explanations and conduct their conversations so that it is easy to present questions, challenges, or alternative proposals and have them taken up. Sociomathematical norms include what counts as acceptable and valued mathematical explanation. Mathematical practices include methods that are established as acceptable without need for explanation.

### **A Conceptual Aspect of Reasoning**

The conceptual activity that we focus on involves the coordination of cognitive resources from different conceptual domains in activities of reasoning. Like heterogeneous reasoning (e.g., Stenning & Sommerfeld, this volume), the cognitive process that we consider involves reasoning that is informed by different kinds of information sources. In heterogeneous reasoning, the sources are different representations (e.g., a diagram and a set of logical formulas). In the reasoning that we observed, the information sources were from different conceptual domains—one primarily involving ecology and the other primarily involving mathematics.

In professional practices that use mathematics, such as architecture or scientific research (e.g., Hall & Stevens, 1996), the integration of information drawn from mathematics and another domain is often seamless; often one cannot be understood without the other. This ubiquitous, implicit coordination of mathematics with another conceptual domain informs professionals' evaluations of their work, including identification of mistakes.

In school problem solving, the coordination of mathematics with other domains is often more problematic. In a study of primary-grade students solving word arithmetic problems, Kintsch and Greeno (1985) hypothesized two forms of understanding that they called *situation models* (following Kintsch & van Dijk, 1978), and *problem models*. According to this hypothesis, solving a mathematical problem includes understanding properties and relations of objects and events in the problem (the situation model), and using that information to construct an understanding in mathematical terms (the problem model). The problem model is often supported by material representations such as equations, which aid students in carrying out appropriate mathematical procedures. Nathan, Kintsch, & Young (1992)

hypothesized that difficulty in forming a model of a situation and coordinating that situation with a problem model is a significant impediment to students' success in learning algebra. They designed an interactive computer system that supports students' construction of algebra problem situations and makes relations needed for the problem model salient. Use of this program facilitated students' problem solving and learning.

### **Learning Environment**

The curriculum materials in this study came from the Middle school Mathematics through Applications Project (MMAP), which was organized as a design experiment (Brown, 1992; Collins, 1992). The design team, housed at the Institute for Research on Learning at Stanford, included teachers and curriculum developers as well as cognitive science researchers. The team developed a middle-school mathematics curriculum in which students work in groups to solve extensive design problems using mathematics (Goldman, Moschkovich, & MMAP, 1995). Students work in interactive learning environments that are middle-school aged versions of design work in architecture, population biology, cryptography, or cartography. The purpose of the curriculum is to have students use math to address problems situated in non-mathematical contexts, often with the assistance of computer applications.

The data we analyzed came from an 8th grade MMAP classroom in the San Francisco Bay Area. They were collected by Rogers Hall and his colleagues (Hall, 1999; Hall, this volume). In the approximately 30-day unit we discuss, called Guppies, students created mathematical models of biological population growth. For their study, Hall and colleagues had collaborated with the teacher in designing a revision of the unit that had been taught and observed earlier. This revision included further emphasis on how to construct mathematical models of population growth and about the exponential functions that underlie them. In addition, they included more explicit attention to the relation between assumptions about guppies' reproduction and parameters of the mathematical model. Our analysis in this paper focuses on one group of students (Manuel, Lisa, Kera & Ned) whose improvement on pre/post assessments placed them about midway in learning of the focus groups videotaped by Hall.

### **Analysis**

We examine two episodes from a videotape record of one student group. These episodes were chosen because each of them included a proposal for a move in the problem space that was incorrect. However, in one case the group identified the error and corrected it, while in the other the group did not identify the error, but instead proceeded using a flawed piece of information. We explain this difference between the successful episode and the unsuccessful one with two hypotheses about collaborative understanding and problem solving in interaction.

First, we hypothesize that the interaction of the group included a kind of threshold for taking up questions, challenges, and alternative proposals that could change the course of activity. Specifically, we hypothesize that detection of a misalignment between the group's situation model and its problem model could occur by a participant's questioning of an operation that was proposed or performed. Both episodes began with an operation initiated by one participant. In the first episode, another participant presented a new interpretation of the situation (which we hypothesize was based on her understanding of the situation model), which illustrated that the current operation was incorrect. In the second episode this participant questioned the initiated operation and expressed a lack of understanding of it. The first of these episodes resulted in the group changing their mathematical approach to accurately reflect their new understanding. In the second episode, the group did not change the operation it was carrying out. We hypothesize, then, that at least in these two cases, presenting a persuasive interpretation based on a situation model was sufficient to bring about a change, while merely expressing uncertainty and lack of understanding of the operation was not.

Second, we hypothesize that in both episodes the initial error involved an inadequate alignment between what the participants understood about the world of guppies (their situation model) and the mathematics (their problem model). That is, the students either did not attend to all the relevant details in the text of the problem when formulating their situation model, or they did not attend to all of the details of the situation model when formulating the problem model. This is consistent with the findings of Nathan, Kintsch & Young (1992), who suggested that when numbers are abstracted from their context, it is possible for students to perform operations that aren't faithful to the situations they are meant to represent (cf Hall et al., 1989). However, such mistakes can be recognized when the context the math is supposed to represent is considered, often using a simulation of some type. In the first episode such a simulation occurred, which led to a reconceptualization of both the problem and situation models. In the second episode however, the problem and situation models were not integrated, and the mistake was not recognized. It is important to note that we do not believe that situation models and problem models are static states, but that they ideally develop in coordination with each other in a recursive process. We suggest that although mathematics problems can be completed successfully without coordination of problem and situation models, integration of the two can highlight when mistakes have been made, leading to more successful problem solving.

### Episode 1 - Pretest

Our first episode comes from the pretest in which the students were trying to answer the question: "Given an initial population of twenty mice who reproduce every season, how large will the population be at the end of two years?" The students had decided that each mouse couple

would have four babies during each breeding season, and that the mice would reproduce eight times in four years. They were asked to show their solutions, which the students did by drawing a graph that depicted the size of the mouse population after each breeding season. Manuel had proposed that the vertical axis would need to extend to 340 mice. When questioned by Lisa and Kera, he explained this conclusion by repeating the mathematical procedures he used: there would be 40 mice born each season, resulting in 320 births, which would be added to the initial 20. When they began to construct the graph, the following exchange occurred:

180 **Manuel:** SO there's sixty... so let's see the first season is over here [*making a mark on the graph*]

185 **Lisa:** Wait a minute

186 **M:** and then sixty plus... is going to be a hundred

189 **L:** Wait a minute, it's forty, and then it's like [*put pencil down and placed fingertips together*] OK. It's forty, right?... And then you have to pair those up [*pressed palms of hands together*] and then they have kids [*spread the palms of her hands apart*]

195 **Kera:** pair the-

196 **M:** oh yeah [*laughing*]

202 **K:** ...my gosh, that's a lot of nasty mice.

Manuel's participation at the beginning of this episode was consistent with the group's usual pattern, in which Manuel initiated actions and responded to questions by explaining why his proposals were satisfactory. Lisa indicated a question ("Wait a minute") then, when Manuel proposed adding 40 to the first data point to infer the next data point, Lisa took the floor, capturing attention with a gesture along with her speech, presenting a reasoned explanation for a different operation that would take account of the 40 mice that were in the population after the first season when they calculated the number born in the second season.

As the students began to graph their results, Lisa realized that the ending population had been miscalculated, and she interrupted the trajectory of the group with a suggestion that was recognized, acknowledged and finally implemented. Lisa's suggestion (line 189) recalled the context of the problem—how mice reproduce—which enabled the students to evaluate the mathematical model they were creating. The linear model that they had previously created had made sense to all three of the members until Lisa simulated a model of the situation they were supposed to be addressing. Thinking about the population growth in those terms enabled her to recognize the error of adding the same number of newborn mice every season. This served to relate the problem model back to the situation model, as the students were forced to think incrementally about the growth of their population.

We interpret this conceptually as follows. First, Manuel's proposal that there would be 320 births involved applying a familiar schema of mathematical practice. A process that increases a quantity may do so by producing a constant quantity during each of several intervals. Inferences about

this kind of process can be made using a schema with three variables: an amount per unit, a number of units, and a total amount. We hypothesize that Manuel applied this schema, using the number of mice the group had calculated for the first year as the amount per unit (births per season) and the number of seasons as the number of units. When Manuel specified and represented 60 as the number mice after the first season, Lisa related this to a situation model in which the number of births in each season depends on the number of mice in the population that season. In that model, the number of births during the second season had to be calculated by considering how many mouse couples there were during the second season. This contradicted Manuel's method and required a change in the problem model, one in which there was a separate calculation of the number of births for each season, rather than a single calculation of the total number of births during eight seasons

### **Episode 2 - Birthrate worksheet**

Our second example is more typical of the type of error recognition that we observed in this group. This episode comes from the middle of the unit. Students were trying to create their own model of guppy population growth. They were told that a population of ten guppies would be coming from Venezuela, and the students' task was to determine how large a tank they would need to hold the guppy population at the end of two years. In order to solve the problem the students needed to determine the specific composition of their original ten guppies (gender and age) and tabulate their birth rate- a complex equation that used multiple variables (see also Stenning & Sommerfeld, this volume).

The interaction in this episode was different than the first episode: although in this group ideas and questions were almost always acknowledged and attended to, one member's suggestion did not always serve to stop or change the trajectory of the group. In this episode Manuel proposed a "shortcut" through the mathematics of a worksheet (line 444). Apparently realizing that this suggestion did not fit into the expectations of the worksheet, Lisa questioned Manuel many times and attempted to stop the group. In this case however, Lisa's interjections (lines 451, 497, 501 and 505) did not successfully redirect the group.

444 **M**: ...If four percent of the frys survive, why don't we just forget about the fry survival and just put that amount for the, for how much are born.

447 **L**: 'cause the number born are not how much survived.

448 **M**: Yes. Yes, the ones who survive are the ones we count, not the ones who are dead, because we don't make room for the ones that are dead.

451 **L**: ...I'm kinda confused

462 **M**: ...why don't we just put four percent on the guppies' birth, because that's how many are going to survive.

497 **L**: but what's that four percent?

498 **K**: the ones that survive

501 **L**: yeah, I know, but how many... of the guppies are four percent?

503 **M**: we don't know, we'll let that mechanical thing work and tell us

505 **L**: wait, are you answering assumption-

506 **M**: let's just try it out.

The group was working on a worksheet for calculating a value of the birth rate to enter into the computer model. The worksheet included four steps. First, the students made assumptions about the gender and age distribution of an initial population of ten guppies. Second, they were to calculate the total number of guppy fry that would be born according to data provided about the number of fry per female of each age in the population. Third, they were to apply a percentage of infant mortality, due to the fact that about 95% of newborn guppies are eaten by their mothers. Fourth, they were to calculate an effective birth rate by dividing the number of surviving fry by 10, the size of the initial population, and converting this to a percentage.

Manuel proposed that the survival rate (which he incorrectly remembered as 4% instead of 5%) could be entered in the computer model as the birth rate. Lisa questioned this, ("how many ... of the guppies are four percent?") but Manuel did not take up the question.

We interpret this episode using a hypothesis about a problem model that was based on an incomplete use of a situation model. The computer program and the worksheet required an entry labeled "birth rate," intended to be expressed as a percentage of the population in the previous season. The group understood that the value of this parameter should reflect the loss of most of the guppies that had been born. The percentage of surviving guppy fry — 4% — fit these specifications, and Manuel proposed to use that as the birth rate. Lisa's questions about this operation were analogous to the challenge she presented in Episode 1. The correct value should have taken into account the number or percentage of guppy fry born in relation to the previous total population, and then take 5% of that (or 4%, on Manuel's misremembered figure). If Lisa's questions had specified the neglected quantity in this case, as she did in Episode 1, she might have succeeded in having her alternative taken up and considered.

In this interaction, Lisa's question was not sufficient to force the group to recognize the mathematical error they had made. In the earlier episode, Lisa stopped the group with a suggestion that simulated the situation model they were supposed to be working from, enabling the group to identify an error in their mathematical reasoning. In this case, Lisa attempted to stop the trajectory of the group without either making a specific suggestion about the relation of the problem model to the situation model, or proposing a new situation model. The other members of the group were not forced to think differently about what they are saying, and consequently, no change was made.

## Comparison

We chose these two episodes because they present a useful contrast for thinking about group problem solving. In both episodes, the group was working along a mathematically incorrect path, and one student questioned that path; but in one case the group corrected itself, and in the other it did not. Specifically, both episodes began with a proposal by one student (Manuel), which involved a mathematical shortcut. These shortcuts appear to have made some sense to the other students in the group, but neither proposal would have led to a successful solution to the problem at hand. What factors may have been involved in the first episode becoming a successful problem solving effort, while the second did not?

We hypothesize that the principal conceptual difference between these two interactions lay in the students relating the situation model and the problem model. The curriculum was designed so that students necessarily developed a situation model about guppies, and used that situation model to construct their mathematical problem model. When they constructed problem models that neglected significant aspects of the situations, incorrect assumptions and conclusions could be easily missed.

The patterns of interaction between the two episodes also differed. In this group, one student (Manuel) consistently took the function of directing the process, initiating and performing operations. Two other students (Lisa and Kera) frequently asked questions or expressed uncertainty. Generally Manuel responded to these by justifying the operation he had initiated or performed. In the first episode (the pretest problem), Lisa not only questioned Manuel's line of thinking, she presented a definite alternative to Manuel's operation. It appears that this met the threshold required for Manuel and Kera to attend to Lisa's idea and to accept it. In the second episode (the birthrate worksheet), Lisa questioned Manuel's shortcut and referred to a critical property of the situation. But, apparently, raising a question rather than proposing an alternative was insufficient to meet the threshold needed to open up a negotiation of how they should proceed.

## Discussion

The hypotheses arising from analysis of these two episodes point to potential contributions both in fundamental cognitive science and the design of learning environments.

First, consistent with Nathan, Kintsch, and Young (1992), we see that students working on contextual problems get into conceptual difficulty when they do not adequately align their situation model with their domain-specific problem model. However, we extend that finding to suggest that for problem solving in real-world contexts it is also important to realign the problem model to the situation model, checking for sensibility in the integrated understanding of the context-mathematics relationship. In this way the problem and situation models develop in coordination with each other and are constantly changing in response to one another. The details of this mapping

between situation model and problem model and back again are subject to further study. From an educational standpoint, it seems that it is important to do more than provide students with a contextual situation from which they can extrapolate a problem model. Another important step is for students to connect the numbers back to the situation model, for it is all too easy to get lost in the abstract world of numbers and forget about their meanings. (See also, Stenning and Sommerfeld, this volume.)

Links between a situation model and a problem model could also be accomplished through the use of material representations. In this unit the students are given worksheets and a computer program in an effort to help them understand the components of making a model, and to guide their understanding of how math can work to create that model. In its current state, the MMAP technology presents a problem model in the form of a network of problem quantities. However, it does not have provisions to facilitate relating the math back to a situation model. One way that the technology might be changed is to present a mathematical representation (as it currently does) alongside a simulation of the components that are taking place. Simulations of that sort might serve to create more links from the problem model back to the situation model, forcing students to think situationally about the math they are producing, making it more likely that they will notice their own mistakes if the simulation doesn't work as they expected.

Still another way to increase students' attention to links between situation models and problem models would be to develop a socio-mathematical norm (Cobb & Yackel, 1996) in which students expect to be accountable for explanations that justify mathematical operations and representations in terms of properties and relations of quantities in situations.

Additionally, interactional patterns create thresholds for questioning, which affect how a suggestion is taken up or explained. Although in this group the students seemed to feel that a mutual understanding was important in order to proceed, oftentimes that understanding was not a consensus. One potential way that such a pattern may be altered is through the larger classroom learning practices, including aspects of reasoning and explaining for which students are accountable to each other and to the teacher. Accountability is provided through discourse activities at different levels, as Hall and Rubin (1998) discussed in their analysis of Magdalene Lampert's teaching. Lampert had an explicit policy that any member of a group could be called on to provide an explanation for any of the group's results. This policy made each group of students accountable for achieving mutual understanding so that the individual members could satisfy the expectation that they would be able to understand the group's results.

Introducing this order of accountability into classroom practices can serve the conceptual linking as well. It is necessary to include some sort of provision for making sure that students understand that finding the "correct" mathematical answer is only part of their responsibility. They also should be responsible for relating what they



found back to the model that they were trying to create and to share that information with their peers. In another example involving FCL classrooms (Brown and Campione, 1994), at the end of a unit students are made accountable for what they have learned by sharing their findings with the rest of their class. Therefore, assumptions that are made throughout the unit need to be accounted for and explained. In the course of doing the birthrate sheet, Manuel made many assumptions about the number of guppies, their survival rate, and how that affected the birth rate. If the group had felt some accountability to a “larger audience,” that is, if they had to present their findings to the class and explain the elements of their model, they might have been less likely to take logical leaps without trying to understand how they related to the model being created.

### Acknowledgements

This research was conducted under a grant from the Spencer Foundation. We are very grateful to Rogers Hall and his associates, Anthony Torralba and Susan John, in the Math at Work research group at UC Berkeley for sharing the videotapes with us. Ongoing conversations with them continue to inform these analyses. We also thank Randi Engle, Keith Stenning, and Rogers Hall for their comments and assistance on this paper.

### References

- Aczel, A. D. (1996). *Fermat's last theorem*. New York: Dell.
- Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences*, 2, 141-178.
- Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, 13, 259-294.
- Cobb, P., Yackel, E., & Wood, T. (1992). Interaction and learning in mathematics classroom situations. *Educational Studies in Mathematics*, 23, 99-122.
- Cobb, P., & Yackel, E. (1996). Constructivist, emergent, and sociocultural perspectives in the context of developmental research. *Educational Psychologist*, 31, 175-190.
- Collins, A. (1992). Toward a design science of education. In E. Scanlon & T. O'Shea (Eds.), *New directions in educational technology* (pp. 15-22). Berlin: Springer.
- Goldman, S., Moschkovich, J., The Middle-school Mathematics through Applications Project Team (1995). In J. L. Schnase & E. L. Cunnius (Eds.), *Proceedings of CSCL '95: The First International Conference on Computer Support for Collaborative Learning*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Greeno, J. G., & Engle, R. (1995) Combining analyses of cognitive processes, meanings, and social participation: Understanding symbolic representations. In J. D. Moore & J. Fain (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 591-596). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hall, R. (1999, Sept. 30). Case studies of math at work: Exploring design-oriented mathematical practices in school and work settings. Final report to the National Science Foundation.
- Hall, R., Kibler, D., Wenger, E., & Truxaw, C. (1989). Exploring the episodic structure of algebra story problem solving. *Cognition and Instruction*, 6, 223-283.
- Hall, R., & Rubin, A. (1998). There's five little notches in here: Dilemmas in teaching and learning the conventional structure of rate. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 189-235). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hall, R., & Stevens, R. (1996). Teaching/learning events in the workplace: a comparative analysis of their organizational and interactional structure. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 160-165). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, 92, 163-182.
- Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363-394.
- Lampert, M. (1990). When the problem is not the question and the solution is not the answer. *Mathematical knowing and teaching. American Education Research Journal*, 27, 29-63.
- McLaughlin, M. L., Cody, M. J., & Reed, S. J. (Eds.), *Explaining one's self to others: Reason-giving in a social context*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Nathan, M. J., Kintsch, W., & Young, E. (1992). A theory of algebra-word-problem comprehension and its implications for the design of learning environments. *Cognition and Instruction*, 9, 329-389.
- Voigt, J. (1995). Thematic patterns of interaction and sociomathematical norms. In P. Cobb and H. Bauersfield, (eds.), *The emergence of mathematic meaning*. Hillsdale, NJ: Erlbaum.

# Work at the Interface between Representing and Represented Worlds in Middle School Mathematics Design Projects

Rogers Hall (rhall@socrates.berkeley.edu)

Graduate School of Education  
University of California, Berkeley, CA 94720

## Introduction

A central process in scientific or mathematical thinking involves being able simultaneously to look *at* and *through* the interface between representing and represented worlds (Gravemeijer, 1994; Latour, 1999). This is particularly true of thinking practices in which people construct and then explore models to gain access to situations that do not yet exist or that occur across scales of time and space that prevent direct observation. While this flexible use of modeling is central to many disciplines, pedagogy has until recently focused primarily on the notational structure of formal systems of representation. This approach can trap learners in the situation of looking *at* complex representational systems, without being able to look *through* them to construct and explore represented worlds (Greeno & Hall, 1997).

This paper adds to a line of work in cognitive studies of mathematics education that examines how learners work at the interface between representing and represented worlds to make inferences, identify and recover from conceptual errors, and manage calculation (Cobb, Yackel, & McClain, 1999; Hall, 1996; Nathan, Kintsch, & Young, 1992; Nemirovsky, in press). Empirical materials are drawn from group modeling efforts in project-based middle school mathematics classrooms (Hall, 1999). In particular, I focus on how a group of students develops an increasingly sophisticated capacity for working with the concept of predation, treated as a functional relation between animal populations (i.e., predator and prey) that can be implemented in particular computational media. In my analysis, these media and other resources available through talk and embodied action develop into systems of activity (Goodwin, 1994) that make up conceptual understanding. From this perspective, concepts and their implementation in diverse representational technologies are inseparable.

## Background to the Student Design Project

Data for this paper (also for papers by Greeno, Sommerfeld, & Wiebe and Stenning & Sommerfeld, this volume) come from studies conducted in middle school mathematics classrooms where students worked on design projects. These projects were supported by curriculum units developed to embed important mathematical concepts in realistic applications (Greeno & MMAP, 1998). In the study reported in this paper, students were asked to act as biological consultants who would devise a proposal for

preserving and then returning a population of guppies to a Venezuelan stream environment. As adopted for use in their classroom, the project lasted approximately four weeks and included the following: task memos directing the activities of student groups, worksheets and supporting case material for the contexts of design problems, a software tool (HabiTech™) that allowed students to model and investigate structures and processes in population biology (see Figure 2), and a set of extension scenarios asking students to model hypothetical events within the Venezuelan stream environment (e.g., harvesting by farmers or the introduction of a predatory fish).

Just before and after working on this project, student groups attempted a 20-minute design challenge in which they were asked to model the relation between mice and cat populations living in a barn over a period of several years. Both the design challenge and students' daily activities during the longer unit were filmed, and various design documents (intermediate and final) were collected. An analysis contrasting pre/post unit performance of groups on the design challenge provides evidence *that* students' understandings of population biology and associated mathematical concepts changed over time. In this paper, I focus on *how* these understandings changed, by following the work of a typical group (labeled the MLKN group) with a particular concept through the longitudinal record of their daily activities on the unit. Data materials are divided into six segments: Segments 1 and 6 come from the pre- and post test design challenge (respectively); Segments 2 through 5 come from the longitudinal record of group work during the unit.

## Evidence of Conceptual Change from a Pre/Post Design Challenge

An utterance-level comparison of the MLKN group's performance on a pre and post-unit design challenge showed that they, like the majority of groups in their classroom (5 of 7 groups), were able to construct and explore a more complete functional model of population growth and predation at the end of the four week unit. At the pretest challenge, this group failed to mention deaths for either population, they did not link together overlapping timelines for otherwise correct models of mouse and cat births, and they made no mention or use of the concept of predation until questioned about it.

As evident in the following exchange<sup>1</sup>, which was recorded at the end of the pre-test, members of this group did understand the qualitative effect of predation, in the sense that cats eat mice and so “reduce” their population. But they had no way to implement this understanding as a functional relation linking together their isolated, hand-calculated models (i.e., for mice and cat populations).

### Segment 1: Predation at the Pre-test (4/21/98)

- 1 Rogers: So if the mice are eating grain...
- 2 Manuel: Uh, huh.
- 3 Rogers: What are the cats eating?
- 4 Lisa: [Mice.
- 5 Manuel: [Mice.
- 6 Rogers: What does that do to the mouse population?
- 7 Manuel: Reduce them.
- 8 Rogers: Ok, [so, as you were doing the mice calculations]
- 9 Manuel: [Ah! Oh::]  
(Lisa and Kera look at Manuel)
- 10 Rogers: Sounds like you were just kinda goin with, four per litter for the mice and letting them... go=
- 11 Kera: =Go, ok.
- 12 Rogers: So they're gonna be getting rubbed out by the cats, right?
- 13 Manuel: [Uh huh.
- 14 Kera: [Right.

The absence of predation as a functionally explicit concept strikes Manuel first (line 9), then he and Kera agree that their models allow cats to grow without bound. As they go on to acknowledge (not shown), this is something that violates the entire premise of the design challenge, and they are eager to get another chance at this kind of problem.

At the post test design challenge, the MLKN group's understanding of population concepts was still unstable and dependent on particular means of implementation (see papers by Greeno et al. and Stenning & Sommerfeld, this volume), but they were also able to implement and explain a functional model of predation. For example, as Manuel struggled to combine timelines for mice and cat populations into an integrated model, Lisa recalled their earlier use of a “Special 2 thing” (i.e., a user-defined function) to model the predation of guppies by wolf fish during the classroom design project. This recalled use of a special function provided a starting point for a fully explicit implementation of predation on the post test.

In the following exchange, recorded near the end of the MLKN post test, Lisa asks Kera for an update on what they

are doing, while Manuel and Ned (silent) work to repair an error with their combined timeline. As Kera explains, they started the combined model with too many mice, generated in an earlier model of mice living alone.

### Segment 6: Predation at the Post Test (5/26/98)

- 1 Lisa: ((to Kera)) Could you run that by me?
- 2 Kera: Um, we ran the model for two years. But we forgot that one year, the cats were living with them. So then they were dying [(inaudible).
- 3 Manuel: [Forty eight. ((resets Moose/Mice<sup>2</sup> to 48))
- 4 Lisa: ((looking at interface)) Uh huh.
- 5 Kera: [(Not in this year.)
- 6 Manuel: [Ok, so now... bring that... to negative. ((relinks Special 2 to Moose/Mice negative pole)) And we started with, how many? ((scrolls down to check Wolves/Cats)) Six, ok. Here we go. Now build... to two thousand and four. ((resets timeline)) Two thousand and four... Now, to the end. ((runs To End))
- 7 HT: ((huge negative value for Moose/Mice population))
- 8 Manuel: Oo::
- 9 Lisa: So how many... [That's only]
- 10 Manuel: [After], after two thousand and four there's negative [mice.
- 11 Lisa: [Can we bring in some dogs there! ((laughing))
- 12 Manuel: Ok:: ((laughing, opens graph window)) Kaboom.
- 13 HT: ((huge negative decline for Moose/Mice))
- 14 Lisa: Oh gosh!

At the end of this design challenge, I (as a research interviewer) asked the group exactly when mice die off. Their first idea was to narrow the timeline, a simplification that increased the resolution of their graph in both axes for time and population abundance. They eventually used this more fine-grained graph and a table of linked values to find that, in their implementation of predation, cats consumed all the mice after only one month.

Comparing pre and post test performances (Segments 1 and 6), it is clear that the concept of predation—along with technical means for implementing, using, and interrogating this concept—changed within the working capacity of the MLKN group. While they neither mention nor implement predation on the pre-test, at the post test this group makes several important advances: (1) They combine partial results from an investigation of mice to model the introduction of cats; (2) They define a predation function that explicitly links cat and mouse populations; (3) They display, investigate, and explain a resulting crisis in the mouse

<sup>1</sup> Transcript conventions include: ((activity descriptions)) appear in double parens, (uncertain hearings) in single parens; [overlapping [onset of talk is shown with left brackets; dynamic computer responses are transcribed as turns at talk.

<sup>2</sup> HabiTech™ provides named population nodes for Caribou, Wolves, Moose, and Guppies. Using Moose for Mice presented students with no particular difficulty.

population; and (4) They notice that cats will, in turn, face a related crisis brought about by a lack of food.

Particularly important for an analysis of work at the interface between representing and represented worlds, these students appear to be able to move fluidly between their roles as middle school collaborators (e.g., Lisa asks for and Kera provides an explanation), technical designers (e.g., Manuel and Ned implement the network, but Kera follows and can explain their implementation), and observers/consultants for a Venezuelan stream environment (e.g., Lisa's proposal that they add dogs to the environment). How students move between these figured worlds (Holland, Lachicotte, Skinner, & Cain, 1998) in a way that helps to develop and explore functionally explicit population models is a question for longitudinal analysis.

### A "Net Wall" Solution to Predation

In this and the following section, I analyze several selections from MLKN's work during the unit on population modeling. First (Segments 2 and 3), I examine their elaborated response as fictional consultants to Venezuelan farmers, in the form of a "net wall" that will serve as a mechanical barrier to predatory fish. The MLKN group sees this as a solution to the problem of losing all the guppies, which farmers need to control mosquito growth, to an exotic population of upstream predators (i.e., the wolf fish). Second (Segments 4 and 5), I examine their computational implementation of predation more closely, asking how their experiences during the unit may have contributed to a more sophisticated performance on the post test design challenge.

After successfully modeling the growth of a guppy population in captivity, the MLKN group chose an extension scenario in which predatory wolf fish were released up stream from the guppies' pond, and farmers later noticed that these guppies were disappearing. The group predicted that the guppy population would flourish in the stream environment before the arrival of wolf fish, then die out as guppies were eaten by newly arriving predators.

Engaging their fictional role as consultants to Venezuelan farmers, the MLKN group began working on solutions that would preserve the guppy population, eventually settling on Kera's proposal for a mechanical "net wall." In the following exchange, Kera reprises the idea of a net in which mesh openings capture wolf fish but allow guppies to swim through. By installing this net at the upstream boundary of the pond, she proposes they can catch and remove wolf fish before they reach the farmers' guppies (see Figure 1).



Figure 1. Kera (middle) animates the path of a guppy swimming through a hole in the net during turn 1 of Segment 2: ((R hand forms opening; L hand, fingers wiggling, traces through)). Lisa (right) forms her own version of the net with her hands as Manuel (left) looks on.

### Segment 2: Blocking the Arrival of Wolf Fish (5/18/98)

- 1 Kera: =Ok, this is (the idea with) the guppies. ((R hand forms opening; L hand, fingers wiggling, traces through)) And it goes sh:::, straight through the net. ((R hand holds opening; L hand traces in and sticks)) And the big fishes go... and they get caught=
- 2 Lisa: =Caught, yeh. And then, they=
- 3 Kera: =(hands grab at center then rise) You pull::: it up and then you take it out.
- 4 Lisa: But why should we pull it out?
- 5 Manuel: No::: [The stream is like fi::ve fee::t deep.
- 6 Lisa: [Do you know...?
- 7 Manuel: No not even five feet, three feet[... deep.]
- 8 Lisa: [Ok, ok, ok, come on.]
- 9 Manuel: You can just pick em out.
- 10 Lisa: So, yeh yeh, so, so it should be like... no no, we can't HIRE anyone to pick it out. It should just like, flow::: naturally. Stuff like that, you know? You know, cause see the [guppies
- 11 Kera: [You gotta pull it out!
- 12 Lisa: No... they won't be CAUGHT in there, cause they're like, HUGE, ok? The hole will be this big.

In this first selection of work from the classroom, several phenomena are important for understanding how students shift between representing and represented worlds. First, a world of Venezuelan streams, farmers, and interacting fish populations is densely inhabited by members of the MLKN group. They literally construct the stream, fish, and a mechanical barrier in the gestural stage between Kera and Lisa, as Manuel reaches in from "downstream." Fish, the stream, and human actors are all

animated (Goffman, 1979; Ochs, Jacoby, & Gonzales, 1994) within this shared space.

Second, while the technical details of the “net wall” barrier are still underway, the importance of isolating guppies from these predators is clearly their emergent goal. Animated from the perspective of a consultant to Venezuelan farmers, this is a response to the consequences of predation, now articulated with the developing notion of a habitat that has semi-permeable boundaries.

The importance of predation in MLKN’s consulting proposal becomes clear later during this class meeting, when the group calls me over to discuss the boundaries of the stream environment. When I ask about the effect of their “net wall” on a graph of the guppy population they had drawn earlier, Kera starts a conditional response (Segment 3).

### Segment 3: The Graphical Shape of Predation (5/18/98)

- 1 Rogers: The graph of the guppy population. Manuel thinks its gonna continue to... [be wavy] and you all think its gonna go down and then [come back up.
- 2 Manuel: [Be wavy.]
- 3 Lisa: No we=
- 4 Kera: =It depends. ((points to drawing of stream in notebook)) Are there still, like... wolf fish in here that are eating the guppies?
- 5 Rogers: Um[:: you can
- 6 Kera: [If there is, ((traces upward path)) then its gonna go a little wavy. But if NOT, then the guppies are just... gonna have their own... ((points to computer)) Like before, when... like our other, um... thingie? (You know what I'm talking about?) [Cause the guppies are living alone, and they're gonna die and (inaudible)
- 7 Rogers: [Ok... I mean if you killed, if you get rid of ALL the wolf fish... then the guppies should... recover with no trouble.
- 8 Kera: =Yeh.
- 9 Rogers: =If there's still some wolf fish, [the wolf fish are gonna continue to grow and stuff.
- 10 Kera: [Then they're gonna ((hands trace oscillation))

According to Kera (turn 6), if any wolf fish get through the net wall the graph of the guppy population will “go a little wavy.” This is because “there’s still wolf fish in there eating them,” as she mentions several times. But if the net wall successfully closes the pond to wolf fish, then guppies will grow in isolation “like before” (i.e., referring to their earlier model of guppies alone in the pond).

Another point is important for understanding how students begin to coordinate movement between representing and represented worlds. Kera’s conditional explanation crosses worlds in the sense that shapes in the

representing world (i.e., graph shapes coming out of their “thingie”) depend upon conditions in the represented world (i.e., the passage of fish through a net opening). As the beginning of an activity system that was intended by the curriculum, types of outcomes, as graph shapes, are being associated with types of models, as determined by their assumptions about habitat (i.e., is the pond open or closed to exotic predators) and relations between populations. And critical to a broader understanding of modeling as such an activity system, results are seen to depend upon starting assumptions.

### Implementing and Exploring Predation in an Integrated Model

The “net wall” consulting proposal is an elegant solution to an emergent design problem, and it works at several levels. Guppies will be preserved for rice farmers, since the wolf fish will be blocked from moving down stream. And this can be done without killing any of these predators. As these students have elaborated the fictional world of the task, this will also keep upstream Venezuelans happy (i.e., those who, according to Lisa, must own wolf fish). Up to this point, the group’s work on this proposal is closely tied to a qualitative understanding of the effects of predation. Yet they are far from a functional implementation in computational media that could produce the graphs in question. As Manuel announces at the beginning of their next class period, “Now how do we make it work?”

The three final conversational segments in this paper illustrate the kind of work these students undertook to construct a plausible (if not entirely correct) functional model of predation. In Segment 4, the group has already constructed a user-defined function that links Caribou/Wolf Fish and Guppies population nodes. With this stable network topology in view, they repeatedly adjust node parameters and run the model in an effort to produce a reasonable number of guppies. Just before this segment starts, Lisa complains that they have a “river full of not plants, not insects, but just fishes.”

### Segment 4: Opening Boxes and Adjusting Parameters (5/19/98)

- 1 Lisa: It’s not enough! As long (as you go over) ten thousand ((changes Caribou/Wolf Fish births to 30% every month)) (inaudible) per cent.
- 2 HT: ((huge positive population value for Guppies))
- 3 Lisa: It's still a lot. (inaudible) about guppies. Yeh, that's the problem.
- 4 Manuel: Yeh, see, but the special two is gonna, do (3 sec)
- 5 Lisa: Alright. Could you guys explain this to us? hh
- 6 Manuel: Uh, explain what?
- 7 Lisa: What's a... special two.
- 8 Manuel: Special two is like when you die because of the caribou.

- 9 Lisa: OH! Really?  
 10 Manuel: Yes.  
 11 Lisa: ((mouse circles over Special 2/Predation node)) Oh this is eighteen? And um... how many guppies do they=  
 12 Manuel: =No, let's do three... times thirty is... thirty, ninety. So its caribou times ninety. ((Lisa changes Special 2/Predation)) Every month, and (then) go... That's it, just... Go to build, go to the thing that says build. Then go to the end.  
 13 HT: ((huge negative population value for Guppies))  
 14 Lisa: Negative? [That's a little too (much), yeh.]  
 15 Manuel: [Oh ok ((sighs))] Now we need to reduce the births. Go to births. No don't touch that, do the births. Reduce the births to ten percent every month.  
 16 Lisa: ((changes Caribou/Wolf Fish births to 10% every month))

With the work of implementing predation in these particular computational media well underway, several phenomena are worth noticing. First, Lisa has been adjusting model parameters without understanding how the predation function works. When she asks “you guys” (Manuel and Ned) for an explanation, Manuel describes what the node does from the perspective of Guppies: it is a type of death caused by Caribou/Wolf Fish.

Second, as Lisa looks inside this function and questions how many Guppies are eaten by Caribou/Wolf Fish (turn 11), Manuel proposes and Lisa executes a change in how the predation node is defined. Manuel’s proposal unpacks the monthly value into a daily rate of consumption (i.e., 3 per day, times 30 days in a month, gives 90 guppies per Caribou/Wolf Fish per month).

This exchange is one of many in which students move back and forth between changing model parameters and running their updated model (these are called “Build” and “Play” modes in the interface) to produce a new set of population values. Over the entire series, each adjustment is sensible within the network topology of their model, but none of these changes produce an outcome that the group finds reasonable (e.g., negative assessments after turns 2 and 13). In the face of this stalled progress, Manuel recalls from their earlier research that overcrowding will cause the guppy birth rate to fall. He reduces this parameter and runs the updated model.

### Segment 5: Arriving at a Guppy Crisis (5/19/98)

- 1 HT: ((running Fast))  
 2 Lisa: Too much.  
 3 Manuel: [No:: its not gone into the e's yet. And it hasn't.  
 4 HT: [((Guppies value in population node rises for awhile, but becomes negative and ends with - 2.71826 \* 10^6 Guppies))  
 5 Manuel: ((opens a graph))

- 6 HT: ((graph shows Guppy population rising, then an extinction crisis part way into third year))  
 7 Manuel: Oh my [god:::  
 8 Rogers: [YES::::!  
 9 Lisa: Oh, it's so funny! [What?  
 10 Rogers: [Yes::  
 11 Manuel: Yes what?

Lisa begins to classify this as another unsuccessful run of their model (turn 2), but Manuel, who has been monitoring the value displayed in the Guppy population node, announces that the positive growth of guppies has not yet reached scientific notation. Then as they watch the interface, the value displayed in the Guppy population node goes hugely negative (i.e., the software automatically shifts into scientific notation) and Manuel opens a graph window (see Figure 2).

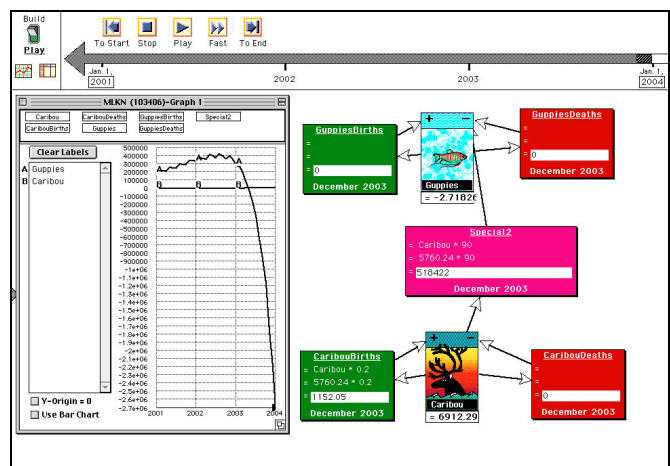


Figure 2. MLKN’s network model of predation and a graph showing an extinction crisis for Guppies during their third year in a Venezuelan pond. The predation node is at the center of the network (right), linking together Guppies (above) and Caribou/Wolf Fish (below).

This graph is striking for members of the MLKN group both because it shows an extinction crisis for Guppies, but also because it catches my eye (lines 8 and 10) as I was working with a group on the other side of the classroom. In a subsequent conversation about this network model and graph, Manuel insists on the influence of overcrowding in lowering Guppies births, while both he and Lisa recount their decision to increase the level of Caribou/Wolf Fish predation. As a final part of their modeling effort, they implemented Kera’s “net wall” as an emigration function, something that was suggested by their teacher as a general strategy for modeling negative influences on population growth.

### Discussion

By the end of the curriculum unit from which these longitudinal selections were drawn, the MLKN group had a sensible and fully implemented model of their consulting proposal, and its behavior was consistent with what they

hoped to achieve in Segment 3 (i.e., Kera's conditional explanation, lines 4 and 6). Since the net wall was implemented as a yearly reduction in the Wolf Fish population, these predators still made it into the pond environment. As a result, some level of predation was ongoing (i.e., this appeared as a scalloped or "wavy" graph of the Guppies population over seasons). But the mechanical "net wall," which they used to remove predators at a regular interval, reversed the outcome of their earlier crisis scenario (i.e., the Guppies population grew steadily over the duration of their scenario).

Predation, as a concept that can be implemented within these particular computational media, was one among several influences in a more complex model of the Venezuelan pond habitat. These influences included (with varying levels of correctness): (a) the starting value established over an earlier period in which Guppies lived alone in the pond, (b) the production of a Guppy crisis after the unregulated arrival of predators, (c) the regulated influence of predation during smaller time cycles within the "net wall" model, and (d) the idea of birth rate suppression during conditions of overcrowding in the pond.

These explicit model components, worked out through repeated cycles of adjusting parameters and holding outcomes accountable to students' qualitative expectations, provided a rich set of resources for their activities on the post test design challenge.

Across these selections from a longitudinal record of group work, more complex forms of coordination appear in the ways that students move between representing and represented worlds. While still far from a technical implementation of their model in computational media (Segment 2), students were able to develop an elegant solution to the problem of stopping or limiting predation. Their work included conversations carried out over a stream gestural stage. Also central in these conversations were processes of animation in which students spoke for (or as) fish in the constructed stream environment, Venezuelan farmers who had diverging interests in these fish, biological consultants concerned with finding a solution for the loss of guppies to predation, and middle school students working on a design project (i.e., as themselves).

As these elaborations of the represented world were carried into computational media, new forms of coordination were required (Segments 3 and 4). These included forms of explanation that linked computational media to aspects of worlds being modeled (e.g., Kera's conditional explanation associates graph shapes with physical events at the net wall in Segment 3). As the structural components of their network model were settled, members of the MLKN group also managed to establish cycles of modeling activity in which they adjusted parameters and compared results with their qualitative expectations.

Through these kinds of activities, students encounter the need to simultaneously look *at* and *through* the interface between representing and represented worlds. As they work through design problems, new conceptual understanding depends upon putting existing concepts and a broader set of representational technologies into coordination. In this sense concepts—as systems of activity—develop in ways that are inseparable from the representational technologies that implement them.

## References

- Cobb, P., Yackel, E., & McClain, K. (1999). *Symbolizing and communicating: Perspectives on mathematical discourse, tools, and instructional design*. Mahwah, NJ: Lawrence Erlbaum and Associates, Inc.
- Goffman, E. (1979). Footing. *Semiotica*, 25, 1-29. Reprinted in Goffman, E. (1981) *Forms of talk*. Philadelphia, PA: University of Pennsylvania Press.
- Goodwin, C. (1994). Professional vision. *American Anthropologist*, 96(3), 606-633.
- Gravemeijer, K. E. P. (1994). *Developing realist mathematics education*. Utrecht, The Netherlands: CD-Beta Press.
- Greeno, J. G. & Hall, R. P. (1997, January). Practicing representation: learning with and about representational forms. *Phi Delta Kappan*, 361-367.
- Greeno, J. G. & MMAP (1998). The situativity of knowing, learning, and research. *American Psychologist* 53(1), 5-26.
- Hall, R. (1996). Representation as shared activity: Situated cognition and Dewey's cartography of experience. *Journal of the Learning Sciences*, 5(3), 209-238.
- Hall, R. (1999). *Case studies of math at work: exploring design-oriented mathematical practices in school and work settings*. Final Report to the National Science Foundation (RED-9553648).
- Holland, D., Lachicotte, W., Skinner, D., & Cain, C. (1998). *Identity and agency in cultural worlds*. Cambridge, MA: Harvard University Press.
- Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Cambridge, MA: Harvard University Press.
- Nathan, M. J., Kintsch, W., & Young, E. (1992). A theory of algebra-word-problem comprehension and its implications for the design of learning environments. *Cognition and Instruction*, 9, 329-389.
- Nemirovsky, R. (in press). How one experience becomes part of another. In K. Beach (Ed.), Special issue of *The Journal of the Learning Sciences*.
- Ochs, E., Jacoby, S., & Gonzales, P. (1994). Interpretive journeys: How physicists talk and travel through graphic space. *Configurations*, 1, 151-171.

# Four letters good, six letters better: Exploring the exterior letters effect with a split architecture.

John Hicks, Jon Oberlander & Richard Shillcock

Division of Informatics

University of Edinburgh

Scotland

United Kingdom

(All correspondence to: [John.Hicks@ed.ac.uk](mailto:John.Hicks@ed.ac.uk))

## Abstract

Recent models employing split neural networks have demonstrated that such architectures are effective for processing visual information. Furthermore, it has been shown that certain emergent strategies of processing are particular to these split architectures. We investigate one such strategy, the exterior letters effect, extending and generalizing it, and go on to discuss the implications that effects which are marked in split architectures bring to bear on lateralization and hemispheric specialization in human cognition.

## Introduction

What might be the advantages for bi-hemispheric processing of visual information? How does real-time high-density information management—such as that employed in the human visual system—cope with the fact that processing of the same thing is done in two halves, in two different places? What is it about the interaction between the hemispheres that allows for the apparently automatic co-operation between them? The answers to these central questions inform almost all other areas of cognition, and discussion of them abounds in the literature. And yet modeling studies on such aspects of gross brain morphology remain relatively under-developed, in spite of the nervous system's clear division centralised in two cerebral hemispheres. The complex relationship that comes into play between particular architectural features and general processing strategies, as well as distinct variations in the nature of the stimuli involved, can play a large role in empirical studies. Although clearly the techniques implicit in learning and execution of a task could be multifarious, models such as the one presented here assist in teasing apart the details of dual processing. Split-architecture connectionist models of cerebral function take as their motivation the well known psychology of the hemispheres, but open out onto a field that is largely uncharted.

## Background

When cognitive science per se was still in its infancy, studies on split brain phenomena were well underway (Gazzaniga, 1970). Work with patients who had undergone commissurotomy made it clear that the two halves of the brain could function autonomously when disconnected. The highlight of this discovery was the apparent inability of the right hemisphere to speak for itself in any real sense (Gazzaniga, 1983). Thus, a century after its initial stipulation, Broca's hypothesis gained even more secure footing. At the same time, the disparate activity resulting from two hemispheres out of touch with each

other, and, in particular, the speechless fumbblings of the right-side, gave a real sense to the distance neuro-anatomically (and thus perhaps experientially) that lay between the hemispheres. This was a distance that was unbridgeable through subcortical structures in the event that the corpus callosum was cut (although see (Sergent, 1987)).

Such severe unlinking is by no means the only evidence of separate identity of the hemispheres. The visual field is split vertically about the fovea in the retina, the right and left halves of the visual field projecting contralaterally into the cortical regions of the left and right hemispheres respectively (Sperry, 1968; Fendrich & Gazzaniga, 1989). Because of this, large scale degradations which are specific to one hemisphere, can lead to marked behavior in tasks reliant on apprehension of the entire visual field, as in cases of unilateral neglect. This deficit, afflicting right-hemisphere stroke victims, manifests itself commonly in the line-bisection task (Halligan & Marshall, 1998; Reuter-Lorenz & Posner, 1990), where the affected portion of the visual field is essentially omitted by subjects asked to designate the midpoint of a line.

The clear contralateral routing of information to opposite hemispheres by the visual system affords a lot of ground for research in normals as well. Key issues about general pattern recognition, symmetry and particularly face recognition can be addressed (Bruce, Cowey, Ellis, Perrett, 1992). Similarly, work in word recognition (e.g. Rumelhart & McClelland, 1981) must at some level be affected by the constraints of the visual processor; assuming the gaze is focussed around the midline of a word, interactionist accounts of processing have to deal at least with the transference of visual information to the locus of letter activation, if not simultaneous activation in different hemispheres.

Jordan's account of letter activation (1990, 1995) bears on the current study. With subjects focusing on a fixation point, stimuli of 200msec or less, containing letter strings of a fixed length but without a full complement of letters (e.g. "d\_k" is a two letter string of length four) were presented and masked with a null string of identical length. Subjects were asked to report the letters that appeared. Significantly, letters coming at the edge points of the string length were more robustly reported than letters that came from interior positions. This "exterior letters effect" (ELE) forms the vehicle for the current discussion on split architectures, and has already been successfully replicated in a connectionist network using a divided "visual field" (Shillcock and Monaghan, in press) each side of which projects separately to one of two hidden layers.

Reggia, Goodall, & Shkuro (1998) describe a word read-



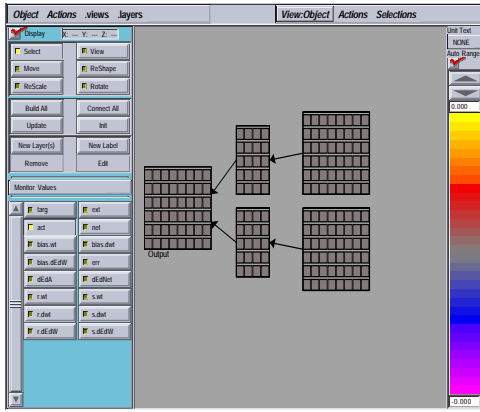


Figure 1: A typical instantiation of a split architecture network, shown here with the aid of the PDP++ graphical interface.

ing task which is learned by a split network. The task is a vehicle for gauging the effects different network parameters have on the degree of lateralization in the fully trained net, lateralization being determined by a “winner take all” competition between two hidden layers given a single input layer. Other modeling work on lateralization deals with the nature of the respective topographies, in terms of cortical organization (Alvarez, & Reggia, 1998; Levitan & Reggia, in press), while elsewhere Shestova & Reggia (1999) do relate a visual identification task to which our models bears an implicit resemblance, insofar as there is a “dual route” strategy for the reception of input.

### Qualitative data using split network

Shillcock and Monaghan (in press) describe a network in which the input field and the hidden units are split in two. With a network similar to that pictured in Figure 1, they present lexical input to the network, but include a positioning technique which allows the four letter words to move across the visual field, being presented in any one of five positions (from occupying only the left hemi-field to occupying only the right hemi-field, passing through the midpoint, where two letters of the four are projected to each side, halfway). It is at root this method of data presentation that ensures that the split net can and will develop a strategy for solution that is not found in the non-split control.

This effect, which relates to Jordan’s work as described above, manifests itself as a diminished reliance in the trained network on the interior letters of words, with a related robustness for recognition for letters in word-final and word-initial positions. Such networks seem to exploit the exterior letters to a greater extent than the nonsplit networks. We claim that the preferential treatment of the exterior letters is provoked by the manner of presentation and the current study is intended to expand upon this idea.

To sum up Shillcock and Monaghan’s findings: there is an ELE, comparable to that found with human subjects, demonstrated by their model. After training the networks with a split architecture showed a significant advantage in recognition of the exterior letters when degraded stimuli consisting of the original words with either the interior or exterior letter pair

“masked” with an ambiguous activation pattern. This finding was true in their study for all positions across the two visual fields. The study was slightly limited however; only four-letter words were used. These are a special case, containing two interior and two exterior letters. Below we explore the affect in the six letter case, also expanding on the criteria used to measure the effect.

## Modeling with a Split Architecture

Rarely are claims made that align connectionist models directly with cellular components of the cortex, upon which the design and operation of simulated neural nets may nevertheless be based.

This caveat is even more salient within the split architecture paradigm, each of the hidden layers ostensibly standing for an entire hemisphere to which input is projected. Other things being equal, it is important to avoid such direct correlations between the neural level and the grain of the model.

## Experiments

Two experiments were performed. For each one, a number of different simulations were run using split and non-split network designs. Each simulation was repeated 10 times and the results all reflect averages for the 10 runs. Subsequent tests using degraded stimuli employed each of the 10 trained nets for that class, the results again being averaged. Details of the nets and the stimuli are given below.

### Materials

A series of simulated neural networks, employing a back-propagation learning algorithm, was trained using the top 60<sup>1</sup> four and six letter words of English respectively. Also used was a list of 60 random strings of the same length<sup>2</sup>. The words were coded following the system of Plaut and Shallice (1994), assigning 8-bit features to each letter, each feature representing an aspect of letter orthography such as “contains closed area” etc. The coded words were then presented to the network through a shift invariant identity mapping (SIIM) task which maintains the integrity of the stimulus organization, while moving it sequentially along the input window. Input nodes that fall outside the location of the word at any time have activation zero, as do the inactive bits within the eight bit feature vector of each letter. The vertical split in the input reflects that of the fovea and thus, as a word is repeatedly presented to the network from all possible positions across the input, it crosses from one “visual hemifield” to the other, activation being redirected to the associated hidden layer accordingly.

Separate networks were used for the four and the six letter tasks, but the number of hidden units, 20, remained the same in each case. Nets not possessing a split hidden layer were used for a control task in which a simple visual field (containing the same number of input units as the non-split network). Networks featured full feed forward connectivity between the layers, save in the case of the non-split models,

<sup>1</sup>Ranking of the words was based on frequency counts from the celex lexical database.

<sup>2</sup>The distribution of letters in these strings was absolutely flat, in opposition to the skewed frequency counts for high frequency words of English.

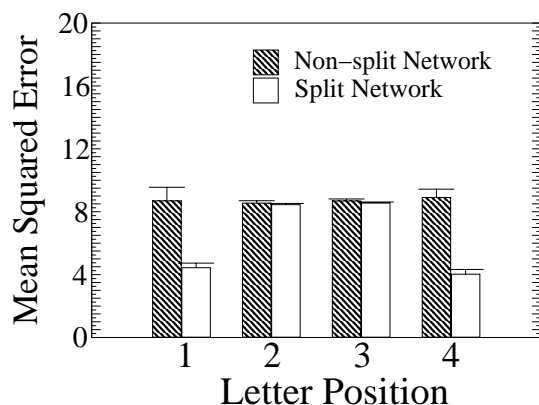


Figure 2: Comparison of nonsplit and split models when networks trained to recognize a set of random strings are fed degraded stimuli, where only the exterior letters are present. The error for the exteriors in the split model is much diminished.

in which the connectivity between the input and hidden layers underwent a random pruning of half of the connection. This was to ensure that the network’s power was consistent with its split counterparts, network power being directly proportional to number of weighted connections.

For all simulations the PDP++ Neural Nets software from CMU was used, running on an Ultra 5 work station.

## Results

### Experiment One: Replication of previous Results

In attempting to replicate the exterior letter effect that Shillcock and Monaghan showed, we trained split and non-split networks on the English and non-word stimuli. As their simulations mirror Jordan’s recognition task for exterior letters, and this involved the presentation of degraded or masked letter strings to trained nets, we used a similar technique. However, it is worth pointing out that we also found a general advantage in *word* recognition for the split networks. This, of course, relates to the size and nature of the lexicon and overall error at the output layer, whereas the letter recognition task is defined in terms of individual letter positions.

On the individual letter scores, for stimuli in which the interior letters were rendered ambiguous, Shillcock and Monaghan found an effect similar to Jordan’s empirical finding, namely that recognition of exterior letters was favorable in such conditions, but significantly more so when the network employed a split architecture. This preference is seen in Figure 2 for non-words and Figure 3 for words. Paired t-tests (two-tailed) checking relative error of exterior letters across networks ( $df = 19$ ) gave  $t = 14.73, p < .0005$  for the study in non-word strings and  $t = 23.32, p < .0005$  for that involving English words, a highly significant effect representing an advantage for the split net in both cases.

Rather than the specific presentation of degraded stimulus that Shillcock and Monaghan demonstrate, to generalize the effects of the split architecture, if they are indeed robust, a more general technique is helpful. The effects of masking letter pairs in strings becomes inordinately complex with

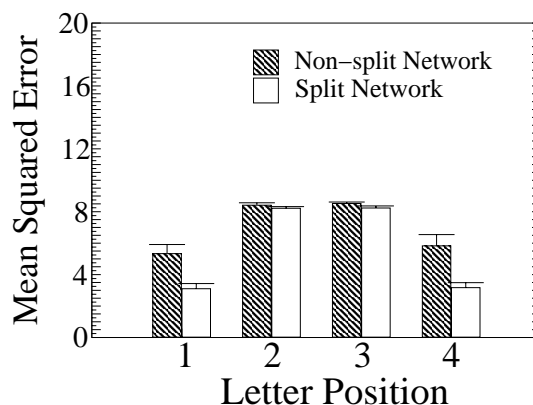


Figure 3: Comparison of nonsplit and split models when networks trained to recognize the English word set are fed degraded stimuli, where only the exterior letters are present, and the interiors masked.

strings even of 6 letters, as we later found (five types of masking means at least a 10 way comparison of masked *on each network*) and generally, we would like to find a more all-encompassing and straightforward view of network behavior, in terms of letter position error after training, for example. To this end we compared the two models, without using masked words.

However, although we were able to replicate and even generalize Shillcock and Monaghan’s findings to a degree, by using degraded stimuli, we found that the effect itself did not significantly cross over into analysis of error levels by letter position as a whole, as Figure 4 shows.

### Experiment Two: Extension of ELE

In the second experiment, our attention was directed to the networks’ performance on the learning task with the six letter stimuli. Again, training consisted of learning over all positions in the visual field, with two different stimulus sets; the top 60 six letter words of English and 60 pseudo words, or random letter strings.

While in the case of the four letter stimuli no significant difference could be demonstrated using error by letter position, for the six letter case there was indeed a notable difference in network performance as seen below. Figure 5 shows the error for each letter position after non-split and split networks had both been trained on the non-word stimuli. In this case a fairly significant drop in error was registered. Taking the difference in error between exterior letters and their adjacent interior letters, we then compared the differences in these (i.e. has the network error dropped significantly for one of the networks on the exterior letters?).  $df = 9$  for each of the following two tailed t-tests: the word initial pair, for each network,  $t = 6.64, p < .0005$ ; the word initial pair in the split network compared with the word final pair in the non-split network,  $t = -3.47, p < .007$ ; the word final pair in the split network compared with the word initial pair in the non-split network  $t = -2.49, p < .034$ ; and the word final pair for both networks  $t = 4.65, p < .001$ . These figures in the main corroborate the story told by the graph: that the split network

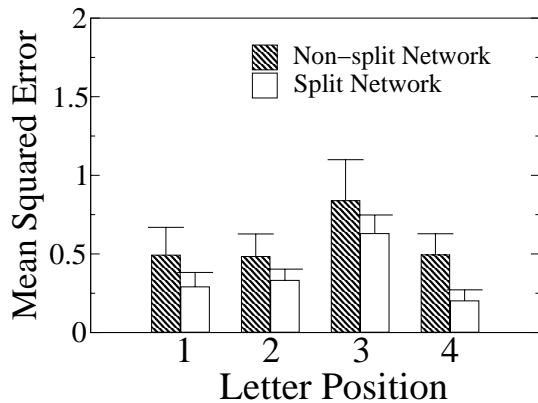


Figure 4: Comparison of nonsplit and split models for the top 60 4 letter words of English. The error is registered after 400 training epochs. Although the error drops across the board for the split model, it does so uniformly, the exterior letters showing no advantage (the best result from 4 separate interior-exterior letter error comparisons between different networks architectures, using a two tailed t-tests,  $df = 9$ , gave  $t = -2.89, p < .018$ )

purchases more success using outside letters than the non-split network. This is statistically clearest for the first and last of the above comparisons, where the only difference was the network architecture (cross word comparisons, e.g. word initial with word final, admit interference from the stimuli). A similar comparison within each network (i.e. seeing if there was a significant drop in performance between interior pairs and exterior pairs not linked to a change in network architecture) yielded,  $t = .97, p < .359$ , for the non-split net,  $t = .54, p < .603$ , for the split, or, no difference.

Figure 6 shows the results for the different nets after training with the English word stimuli. As above,  $df = 9$  for each of the following two tailed t-tests: the word initial pair, for each network,  $t = 6.30, p < .0005$ ; the word initial pair in the split network compared with the word final pair in the non-split network,  $t = -12.07, p < .0005$ ; the word final pair in the split network compared with the word initial pair in the non-split network  $t = -6.81, p < .0005$ ; and the word final pair for both networks  $t = 2.84, p < .019$ . The significant dip in the error of exterior letters reiterates the trend shown in the graph. Of particular interest here is the form of the “arch” in the error by position of the split network, as well as the quasi-sinusoidal effect the non-split net seems to find when presented with the English word strings. These topics are taken up in the general discussion.

## Discussion

In this study we have performed experiments with a series of split and non-split neural networks. The results re-affirm the main finding of Shillcock and Monaghan, that a difference in network performance is based on the architecture, split or non-split, that that network employs. Shillcock and Monaghan’s model produced an ELE, which says exterior letters of strings are favored in conditions of stimulus degradation. This effect was demonstrated by them under very spe-

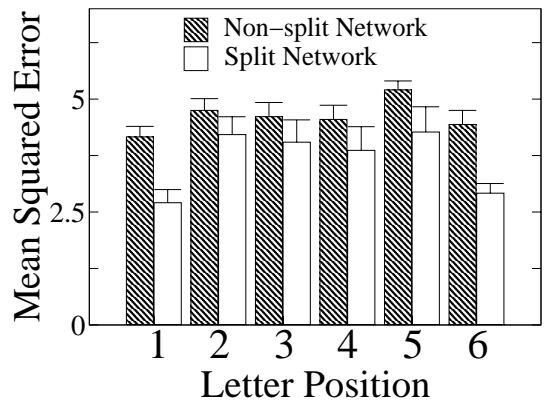


Figure 5: Comparison of nonsplit and split models for 60 random strings of 6 letters each. The error is registered after 400 training epochs. See text for details.

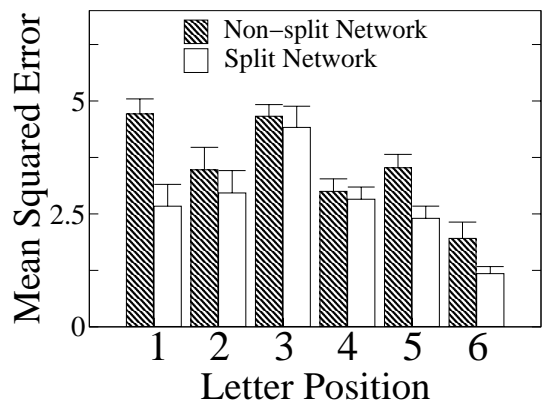


Figure 6: Comparison of nonsplit and split models for the top 60 6 letter words of English. The error is registered after 400 training epochs. See text for details.

cific conditions, which we were able to generalize as holding across the board for degraded stimulus<sup>3</sup>. The effect, a large drop in relative error by the split network for exterior letters only, is clearly seen in the corresponding figures (2 and 3). We tried but were not able to extend Shillcock and Monaghan’s results still further using simple error monitoring criteria, whereas with six letter strings the simple error metric not only revealed the ELE, but did so strongly.

In general, the ELE can be seen as the benefit of having a split hidden layer. With a single hidden layer, the mapping learned by the network for each pattern at each letter position is highly interdependent. Thus instantiations of letters at one position are much more likely to be conflated with their immediate neighbors. What a separate layer for each visual field buys is a foothold for representational independence. The same mapping is learned in either case, but with

<sup>3</sup>It is worth noting that in their actual task, Shillcock and Monaghan read error at single presentation positions of input, as well as the corresponding letter position at output; that is, although they examined every position, we demonstrated the cumulative effect of the error at different positions.

the split network, error back-propagated from the output to hidden layers during learning brings each hidden layer into line with the other through an indirect coordination. Thus a modicum of independence in each layer is retained, and this is used as collateral against an investment, or specialization, that that layer makes in direct proportion to the input it is exposed to. And this input favors, in the case of each hidden layer, the exterior letters of the stimulus, by simple fact of relative exposure (interior letters disappearing across the “fovea” and into the other hemi-field sooner for every pattern presented). This potential “separateness” for marginal phenomena (i.e. exterior letters) licenses, amongst other things, the robust behavior in the face of degraded input the split network demonstrated.

Other questions remain, however. For example, although for the six letter case we were able to show preferential learning for exterior letters just by monitoring error by letter position, the four letter case yielded no such view. A possible reason for this is network competence in terms of the capacity of the hidden layer to find a secure mapping from input to output. The total number of hidden units was the same in both nets; yet the six letter strings required not only a larger input area (two visual hemi-fields of *six*, as opposed to four, letters each), but they also constituted a much larger input set in general, as each word appeared in each possible position (five for the four letter model, but *seven* for the six letter case). Thus at the lower end of the extreme, the smaller net manages its quarry rather elegantly, the residual shape of the error by letter (Figure 4) probably reflecting nothing other than the structural regularities present in English orthography. When this competence envelope is pushed, as in the case of increasing the task load on the hidden units with the introduction of a six letter mapping, the hidden layer is forced to resort to economic measures, visible as the ELE. Indeed, this would provide some explanation of why, at the four letter level, the ELE can only be detected with finer method, the presentation of corrupted input.

If the effect is a conflation of these two trends, then that goes some way to explaining the “arch” of the split bars in figure 6: the pressure on the net to retain as much as it can means a sacrificing of the representations of interior letters in favor of the exterior representations which are easier for each hidden layer to maintain. The contrary shape of the nonsplit network in the same figure suggests that it needs to resort to a different strategy<sup>4</sup> in order to degrade gracefully under the increased weight of the six letter task.

These suggestions form but a part of a larger set of topics to which modeling with a split architectures gives rise. There are many others besides, not the least of which is a retention of the intuitive notion known as “modularity” at some level in the brain. At one time, connectionist models threatened to rule out the idea of “separate parts” altogether. The current study is one which demonstrates the integration of two concepts: the benefits brought by separation—e.g. the independence of the hidden layers as a means of exploiting presentational regularities, like exposure to exterior letters, which are themselves brought about *for free* through the relation-

<sup>4</sup>Perhaps one not unlike taking English C-V-C phonological regularities, or rather the way they manifest in orthography, as a template for resolving input.

ship that obtains between the model and the environment; and the importance of concentration, as that of the units within the hidden layers, without which the error driven learning of such problems would not be possible at all.

## Acknowledgments

The first author acknowledges the generous support of the EPSRC in this work.

## References

- Alvarez, S., & Reggia, J. (1998). Metrics for Cortical Map Organization and Lateralization. *Bulletin of Mathematical Biology*, *60*, 27–47.
- Bruce, Vicki (Ed.), Cowey, A. (Ed.), Ellis, Andrew W. (Ed.), Perrett, D. I. (Ed). (1992). *Processing the facial image*. Clarendon Press/Oxford University Press, Oxford, England.
- Fendrich, R. & Gazzaniga, M.S. (1989). Evidence of foveal splitting in a commissurotomy patient. *Neuropsychologia*, *27:3* 273–281.
- Gazzaniga, M.S.(1970). *The Bisected brain*. Appleton-Century-Crofts; New York.
- Gazzaniga, M.S. (1983). Right hemisphere function following brain bisection: A 20 year perspective. *Am. Psychol.*, *38* 525-549.
- Halligan, Peter W. & Marshall, John C. (1998). Visuo-Spatial neglect: The ultimate deconstruction? *Brain & Cognition*, *37:3* 419-438.
- Jordan, T.R. (1990). Presenting words without interior letters: Superiority over single letters and influence of postmark boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 893–909.
- Jordan, T.R. (1995). Perceiving exterior letters of words: Differential influences of letter-fragment and non-letter-fragment masks. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 512–530
- Levitan, S & Reggia, J. (in press). Interhemispheric Effects on Map Organization Following Simulated Cortical Lesions. *AI in Medicine*.
- Plaut, D. C. & Shallice, T. (1994). *Connectionist modeling in cognitive neuropsychology: A case study*. Lawrence Erlbaum Associates, Inc; Hove, England.
- Reggia, J, Goodall, S, & Shkuro, Y. (1998). Computational studies of lateralization of phoneme sequence generation. *Neural Computation*, *10*, 1277–1297.
- Reuter-Lorenz, Patricia A & Posner, Michael I. (1990). Components of neglect from right-hemisphere damage: An analysis of line bisection. *Neuropsychologia*, *28:4* 327-333.
- Rumelhart, D. E. & McClelland, J. L. (1981). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*, 60–94.
- Sergent, J. (1987). A New Look at the Human Split Brain. *Brain*, *110*, 1375–1392.

- Shestova, N & Reggia, J. (1999). A Neural Network Model of Lateralization during Letter Identification. *Journal of Cognitive Neuroscience*, 11:2, 1277–1297.
- Shillcock, R, and Monaghan, P. (in press). The computational exploration of visual word recognition in a split model. *Neural Computation*.
- Sperry, R. W. (1968). Apposition of visual half-fields after section of neocortical commissures. *Anatomical Record*, 160 498–499

# Eye-Tracking and Conceptual Combination

Dietmar Janetzko

dietmar@cognition.iig.uni-freiburg.de

Institute of Computer Science and Social Research

Center of Cognitive Science

University of Freiburg

D-79098 Freiburg

## Abstract

Processing conceptual combinations has been shown to be based on interactive activation of the concepts involved (Coolen, van Jaarsveld, & Schreuder, 1993). In this paper an approach for investigating conceptual combinations of nouns in an online way by using eye-tracking data is described. Words fixated in a compound-production task form a sequence of symbolic data that can be analyzed by a psychometric method called knowledge tracking (KT) that is based on Markov processes. Empirical evidence has been found that conceptual combinations assessed as medium acceptable attract eye movements more frequently than other ones (especially clearly acceptable or clearly unacceptable conceptual combinations).

## Introduction

Understanding compounds is an ubiquitous cognitive process. Many languages are rich in compounds, and conceptual combination is one of the most important ways of forming new concepts. Currently there are five major theories of concepts: the classical approach, the prototype theory, the exemplar approach, the theory-based model, and the theory of psychological essentialism (Hampton, 1998). It is, however, well known in the literature on concepts that none of these theories provides a convincing explanation to a key issue encountered when processing concepts: conceptual combination. Regarding prototype theory of concepts, for instance, Osherson & Smith (1981, 1984) made evident that there is no generally applicable function that maps the prototypes of PET and FISH to the prototype of the resulting compound PET FISH. However, in many languages – including, e.g., German – conceptual combination is a basic mechanism in both generation and understanding of natural language. In a word: while previous work on concepts had a narrow focus on simple concepts, it is now generally accepted that a theory of concepts can not do without a theory of compounding and methods to carry out empirical investigations accordingly.

There are two general strands of theories that react to the failure of the major theories of concepts. One of these strands has a clear semantic orientation. Theories in this tradition seek to explain conceptual com-

bination by referring to the meaning of compounds. A case in point is the concept-specialization theory (Murphy, 1988), which regards conceptual combination like HOUSE BOAT as a refinement or specialization of the more general concept BOAT. The other strand of theories on compounds has a more syntactic orientation and is rooted in linguistics, in particular in the syntax of words (Selkirk, 1982). Work on compounds in this tradition is based upon the observation that there are striking parallels to fundamental phenomena well known in sentence processing: First, we can generate and understand an unlimited number of compounds on the basis of a small number of simple concepts. Second, we can assess the well-formedness of compounds indicating that there is a “grammar of concepts” with some classes of concepts being more prone to form combinations with others. Recent work on conceptual combination tries to link both approaches, e.g., by analyzing semantic constraints to the compounding process (Keane & Costello, 1997) or by trying to establish a catalogue of semantic relations that link concepts together (Gagné & Shoben, 1997). The methods applied, however, can hardly capture the process of compounding, which has been shown to be highly interactive with the concepts involved in a compound activating each other mutually (Coolen, van Jaarsveld, & Schreuder, 1993).

While rating studies, analysis of thinking aloud protocols and reaction time studies clearly provide valuable insights into conceptual combinations they have difficulties to capture the interactive nature of processing conceptual combinations. We take the view that investigations of conceptual combinations could profit very much from methods that take the interactive nature of processing conceptual combinations into account. Information of this type establishes constraints concerning theories about conceptual combination.

The goal of this paper is to present a method that allows for an approach to online-investigation of conceptual combinations. The paper is organized as follows: First, we briefly report on previous work on Markov processes in cognitive science. Second, an overview of *knowledge tracking* (Janetzko, 1996; 1998; in press) is given. Knowledge tracking is a method that is based

on Markov processes and tailored to analyzing sequential symbolic data so that underlying cognitive structures become explicit. Third is an outline of an empirical validation study that shows more specifically how this approach can be brought to bear in empirical research. In particular, it is described in which way eye-tracking protocols are recorded while subjects build conceptual combinations. Finally, we discuss possible consequences for investigating concepts.

### Eye-Tracking and Markov Processes

The method used to analyze eye-tracking protocols rests on Markov processes, which are usually explained by referring to stochastic processes. A stochastic process is defined by a random variable  $X_n$ , a state space (potential values of the random variable), and transition probabilities between the states. Processes with every state depending on one or many preceding states are called Markov processes. Models based on Markov processes are quite common in fields like pattern recognition – in particular speech recognition – or DNA sequencing. In speech recognition, hidden Markov models (HMMs), viz., a special type of Markov process, are widely used. In HMMs, the states are unknown. Markov processes have also been used and adopted to the analysis of sequential data in cognitive science, in particular eye-tracking data (Suppes, 1990; Salvucci & Anderson, 1998). Here, fixations form a sequence of states, and the outcome of analysis is the identification of a model that accounts best for some observed sequence of states. When using a method based on Markov processes like knowledge tracking for analyzing cognition it is important to remember some of their defining features. In particular, the fact that this technique derives prediction in a strict history-based way has to be considered. For this reason, modelling controlled cognitive processes (e.g., goal-directed cognition like some types of planning or problem solving) via Markov processes raises severe problems. In this case, the phenomenon analyzed clearly conflicts with features of the formal model used. By the same token, modelling cognitive processes that underlie conceptual combination by analyzing eye-tracking protocols appears to be a suitable field for applying this type of models. The reason for this is that processing conceptual combination is (even for novel compounds) often extremely fast (e.g., Zwitserlood, 1994) and thus apparently not a goal-driven process.

### Knowledge Tracking

Knowledge tracking (KT) is a psychometric method that carries out a diagnosis of cognitive representations. Knowledge tracking can be used in confirmative or in a generative mode. The former provides a rationale to decide which of some candidate theories (concept structures) explains a sequence of data

best.<sup>1</sup> The latter may be taken to generate a concept structure on the basis of some start-up structures such that the newly generated structure fits to the data best (Janetzko, in press). We will, however present only the confirmative mode of knowledge tracking. Knowledge tracking rests on Markov processes models (Gardinger, 1990), but it is tailored to analyzing cognition. For instance, knowledge tracking provides more flexibility when calculating goodness of fit scores between empirical data and models. The models may be parametrized such that spreading activation in models is realized (Janetzko, in press). Furthermore, models set up within knowledge tracking are empirically testable, which is not the case in standard HMMs (Dijkstra & de Smedt, 1996).

### The Data: Sequence of Concepts

The input of data required by knowledge tracking is a sequence of symbolic data or concepts (e.g., the sequence of the concepts CAT, DOG, FISH, MOUSE etc.) that refer to the sequence of states in a Markov process. This kind of data may be obtained in eye-tracking studies, thinking aloud studies or studies of HCI (human computer interaction).

### The Theory: Relations and Structures

Knowledge tracking needs a theory to analyze sequences of symbolic data. To specify a theory we have to select one or many relations (e.g.,  $x$  is-a  $y$ ,  $x$  eats  $y$ ). On the basis of a relation we may then add a set of concepts that are taken to instantiate the relations. We end up with concept structures. A very simple concept structure can be described in a Lisp-like notation as (is-a (MOUSE MAMMAL) (HORSE MAMMAL) (SHARK FISH) (HERRING FISH) (FISH VERTEBRATE) (MAMMAL VERTEBRATE)) (cf. Fig. 1). Every network (e.g., hierarchies ontologies, partonomies, semantic networks) of concepts, be it a cyclic or an acyclic graph, can be called a concept structure. Other formalisms of knowledge representation like, e.g., schemas or scripts may also be redescribed as concept structures (Janetzko, 1996).

### Calculating Scores for Goodness of Fit

In KT, the theory, viz., one or many concept structures, is taken to calculate goodness of fit scores on the basis of sequences of symbolic data. The goodness of fit scores describe how well a sequence of symbolic data can be explained by a concept structure. Usually, a number of concept structures is brought to bear, all of which are competing as far as the explanation of the data is concerned. The structure that yields the best goodness of fit score will then be taken as the most suitable model for the cognitive structure explaining the

<sup>1</sup>By explanation we refer to the theory-based prediction of data.

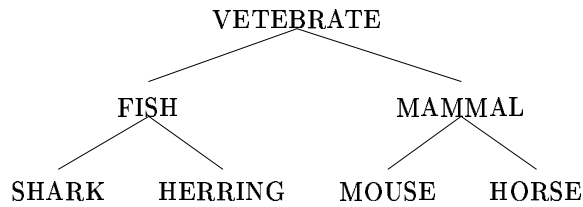


Figure 1: Simple Concept Structure organized by the Relation is-a

empirical data. To compute scores for goodness of fit we have to transform all concept structures into transition probabilities. The technical details behind the calculation of the goodness of fit scores are described in Janetzko (1996; in press).

### Knowledge Tracking in five Steps

In sum, analyzing cognitive structures via knowledge tracking involves five steps:

- eliciting concepts and relations in the domain under study and setting up concept structures,
- recording empirical data (sequences of concepts), e.g., in eye-tracking studies,
- expressing the concept structures by transition probabilities; this is essentially the transformation of knowledge-based models into probabilistic models,
- explaining empirical data by using concept structures and calculation of goodness of fit scores
- selecting the structure that produces the best goodness of fit score.

Empirical validation studies carried out with data collected in human-computer interaction support the claim that the structure that gives the best account of the empirical data is in fact the structure that has dominated cognition while producing the data under study (Janetzko, 1996).

### Eye-Tracking and Conceptual Combination

We used a simple production task to record eye-tracking data while subjects were engaged in conceptual combinations. Subjects were presented with a computer screen where randomly simple German nouns were displayed in circular way (cf. Fig. 2).<sup>2</sup> Presenting the stimuli (nouns) in a circular way does not lead to one big path of overlapping eye-tracks. Nilly-willy, this would have been the consequence, if we had

<sup>2</sup>Translation (in clockwise order beginning with the concept at the 12 o clock position): Way, Fence, Castle, Garden, Guest, King

presented the stimuli in a list-form. Thus, a circular arrangement of the stimuli allows us to analyze the eye tracking more conveniently (cf. Fig. 2).



Figure 2: Arrangement of Concepts in Study 2

The subjects were requested to form noun-noun-compounds by using the concepts presented on the screen. In so doing, subjects had to rely on their eye-movements in order to combine concepts. In preliminary studies, it became obvious that some subjects throw one glance on the screen and rely then heavily on their working memory. Clearly then, in these cases the eye-tracking data are not indicative for the compounding process. To impede this memory-based strategy we introduced a secondary task: Subjects were requested to count backwards from 10 to 0. Whenever they were able to produce a compound they could pause during counting backwards, state the compound and start again counting backwards. Eye-tracking was recorded while the subjects were producing compounds. The fact that our subjects could combine concepts while doing a second task provides supporting evidence to our initial assumption that conceptual combination is not a goal-directed process. Every indication to the contrary would have raised problems concerning the application of a technique based on Markov processes. To balance out sequence effects, we set up a computer program that arranged the items randomly in a circular way for each trial. The presentation of items was never in one line (cf. Fig. 3). In this way, a possible bias towards reading from one item to the next one to the right was minimized. The whole procedure of recording the eye-traces is presented in Zugenmaier and Janetzko (1998). By presenting the stimuli in the way described we are in a position to record the eye-movements while subjects were carrying out the task of conceptual combination (cf. Fig. 4). Note that the sequence of eye-movements can easily be conceived as a sequence of symbolic data. Knowledge tracking allows us to analyze these data by calculating goodness of fit scores with respect to con-



cept structures.

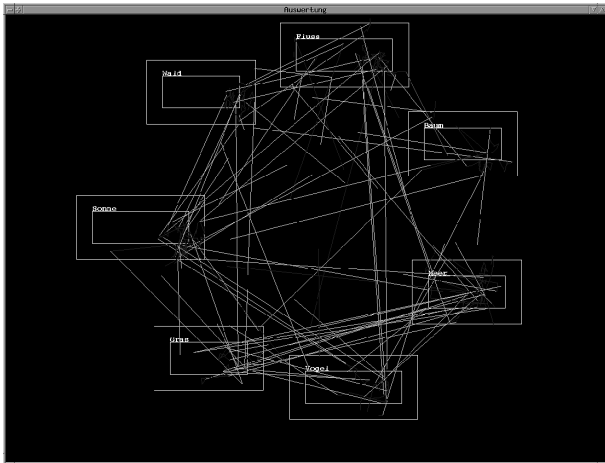


Figure 3: Eye Movements recorded in Conceptual Combination

### Empirical Validation Study

The goal of the empirical validation study was to test the sensitivity of analyzing eye-tracking data with respect to underlying structures by using knowledge tracking. The general steps when applying this approach are as follows:

- Setting up concept structures that slip into the role of hypotheses used to analyze sequential symbolic data via knowledge tracking.
- Administering a compound-production task and recording the eye-tracking protocols.
- Converting the sequence of eye-movements into a sequence of symbolic data.
- Analyzing the sequence of symbolic data via knowledge tracking.

In the following sections we present the steps of this examination in more detail

#### Study 1: Specification of Concept Structures

The purpose of this study was to elicit acceptability scores for compounds that could possibly be of predictive value for eye-tracking behavior as examined in the next study. We used compounds built by the concepts that were also used in the following study.

#### Participants

Participants were 12 subjects (6 male, 6 female). According to a simple questionnaire administered before the investigation all subjects spoke German as their first language.

### Materials

Subjects had to assess the acceptability of 49 nominal compounds that were systematically produced by using the words AUTO, HAUS, PARK, TÜR, SCHIFF, STUHL, ZAUN (Translation: car, house, park, door, ship, chair, fence). All words used for conceptual combinations can be considered as simple German words that are very common according to word frequency indexes like CELEX.

### Procedure

On the basis of the seven concepts stated above all possible noun-noun compounds were produced (AUTOHAUS, AUTOPARK, AUTOSCHIFF etc.<sup>3</sup>) Some of these concepts are true lexicalized compounds that are in everyday usage of German speakers (e.g., HAUSTÜR, engl: housedoor) while other compounds sound rather odd for German speakers (e.g., ZAUNSCHIFF, engl: fenceship). Still other compounds are with respect to their acceptability between these extremes. The subjects assessed the acceptability of each compound on a 5-point rating-scale.

### Results

The results of study 1 was a simple classification into 5 classes of compounds that differed with regard to their level of acceptance. Moreover, each of these classes had an internal structure (Fig. 4), which was employed in the following study. For ease of presentation, we will just give an outline of the summary scores obtained in this rating study (cf. Tab. 1).

### Discussion

The acceptability ratings were transformed into concept structures that could easily be used by knowledge tracking as hypotheses required to analyze symbolic sequential data. If, for instance, we transform the class of highly acceptable compounds into a concept structure, we obtain a structure like (class-1 (AUTO HAUS) (AUTO TÜR) (HAUS TÜR) (PARK HAUS)). The meaning of this structure is simply that by using the nouns listed in pairwise brackets highly acceptable nominal compounds can be built. Similar concept structures can be constructed by the data that lead us to establish the other classes of compounds (Tab. 1).

#### Study 2: Compound-Production Task

In study 2 we collected eye-movement protocols (sequences of symbolic data) that reflect cognitive processes in a compound-production task. Before specifying details of study 2, it is important to see the linkage between both studies. By using the results of study 1 we have established concept structures that express the acceptability of compounds. In study 2 data

<sup>3</sup>In an agglutinative language like German compounds usually form a single compound word.

Table 1: Stimuli and results of study 1

Class	Compound	Rating
1	AUTOHAUS, AUTOTÜR, HAUSTÜR, PARKHAUS	$\bar{x} = 1$
2	AUTOPARK, HAUSSCHIFF, HAUSZAUN, PARKTÜR, PARKZAUN, ZAUNTÜR	$1 < \bar{x} \leq 2$
3	AUTOSCHIFF, HAUSPARK, PARKSTUHL, SCHIFFPARK, STUHLPARK	$2 < \bar{x} \leq 3$
4	AUTOSTUHL, HAUSAUTO, HAUSSTUHL, SCHIFFSHAUS, TÜRZAUN	$3 < \bar{x} \leq 4$
5	PARKSCHIFF, STUHLAUTO, STUHLSCHEIFF, TÜRHAUS, TÜRPAK, TÜRSCHIFF, TÜRSTUHL, ZAUNAUTO, ZAUNSCHEIFF	$4 < \bar{x} \leq 5$

are collected that will be analyzed by using these concept structures. In so doing, we can address the question whether or not acceptability of compounds is important for the eye-movement behavior. We hold the hypothesis that subjects will fixate more often compounds that are at a medium level of acceptability. This should be so since compounds considered very high or very low in acceptability should be analyzed more quickly. Hence they should lead to less pair-wise fixations. In contrast, compounds on a medium level of acceptability should be considered more intensively. Here, we expect a high rate of “jumping” back and forth between the concepts involved.

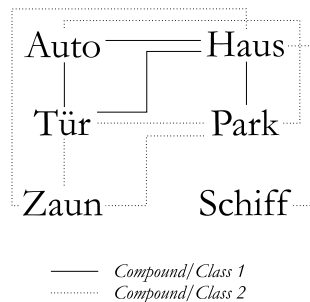


Figure 4: Concept Structures based on Study 1

The types of arrows indicate different levels of acceptance for compounds. Note that only the compounds of classes 1 and 2 are presented (Park=parc, Haus=house, Tür=door, Schiff=ship, Zaun=fence, Auto=car).

## Participants

Participants of study 2 were 5 subjects of whom 2 were female. According to a simple questionnaire administered before the investigation, all subjects spoke German as their first language.

## Materials

The 7 German words, the compounds of which have already been described in study 1, were also used in study 2.

## Procedure

Subjects had to produce compounds according to the procedure described above. Eye-tracking data were recorded by IVIEW, a video-based tool for eye-tracking by Sensomotoric Instruments that uses the corneal reflection technique. The analysis software allowed us to specify rectangular areas laid over the concepts to decide whether or not a word has been fixated.

## Results

The eye-movements were automatically recorded and transformed into a sequence of symbolic concepts (trace). This trace has been analyzed by using the five concept structures that were obtained as a result of study 1 (cf. Table 1). We carried out a descriptive analysis of the data. Figure 5 shows the results of our analysis (goodness of fit scores) on the y-axis. These were obtained by analyzing the eye-tracking data of five subjects across five classes of compounds that are lined up on the x-axis (cf. Fig. 5). The results provide supporting evidence to our hypothesis that highly acceptable (class 1) and also highly unacceptable (class 5) compounds do not lead to intensive processing while compounds that are on a mediate level of acceptability do. Clearly, we need further data to establish a firm empirical ground. However, the tendency of the data testifies to the usefulness of this method.

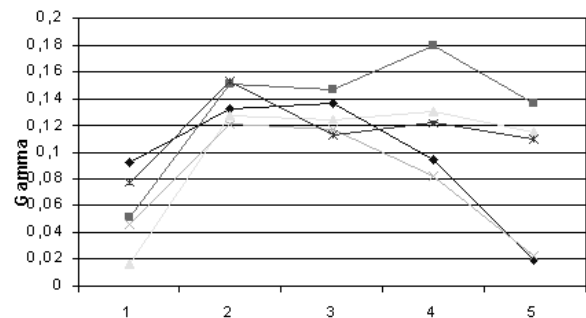


Figure 5: Analysis of Eye-Tracking Protocols via Knowledge Tracking

## General Discussion

The purpose of this study was to show the feasibility of knowledge tracking as a method to analyze eye-tracking in compounding. Knowledge tracking is a general method that can be taken to analyze each type of sequence of symbolic data on the basis of some concept structures (cf. Janetzko, 1998; 1999; in press). In our analysis of eye-movement protocols via knowledge tracking three aspects became apparent: First, the method employed gives a good indication of cognitive processes in conceptual combination. In particular, it is an online-method, and it thus provides insights into conceptual combination by measuring the effort put into this task. However, eye-tracking protocols especially when recorded in exploratory tasks like ours suffer from a bad signal-noise ratio. This is due to the fact, that subjects very often generate and test compounds. Second, we only applied concept structures that essentially express whether or not a compound is or is not acceptable. If knowledge tracking is used to investigate the knowledge used in compounding the concept structures applied have also to represent knowledge. This can be done, if we take compounding relations like *x is\_made\_of y* or *x causes y* (Gagné & Shoben, 1997) to analyze the conceptual combination. Third, we may assume that vast amounts of knowledge are applied in a task like the compound-production task introduced in this paper: Possible relations are tested, and analogues to well-known similar compounds are generated. For a more complete analysis of the knowledge involved in a task like this, a method is required that can tap the theories applied by a problem-solver once he or she forms a compound. To meet this requirement, we have developed a version of knowledge tracking that is no longer restricted to be a confirmative method. This type of knowledge tracking specifies the plausible bridging inferences that may be drawn between pairs of concepts in a symbolic trace. Then, it adds up these inferences to a theory underlying the production of the symbolic trace (Janetzko, in press).

## Acknowledgments

I like to thank Gerhard Strube for valuable comments to this paper. I am grateful to Marc Klinger, Oliver Jäschke, and Dirk Zugenmaier for the programming work put into this project.

## References

- Coolen, R., van Jaarsveld, H. J. Schreuder, R. (1993). Processing novel compounds: Evidence for interactive meaning activation of ambiguous compounds. *Memory & Cognition*, *21*, 235–246.
- Costello, F. & Keane, M. T. (1997). Polysemy in conceptual combination: Testing the constraint theory of combination. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 137–141). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Dijkstra, T. & de Smedt, K. (1996). *Computational Psycholinguistics*. London: Taylor Francis.
- Gagné, C. L. & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 71–87.
- Gardinger, C. W. (1990). *Handbook of stochastic methods*. Berlin: Springer.
- Hampton, G. (1987). Inheritance of attributes in natural concept conjunctions. *Memory and Cognition*, *15*, 55–71.
- Janetzko, D. (1996). Knowledge tracking - A method to analyze cognitive structures. *IIG-Berichte*, *2*. Freiburg, Germany.
- Janetzko, D. (1998). *Knowledge tracking - A tutorial and a device for remote calculations* [WWW document] <http://cogweb.iig.uni-freiburg.de/KT>
- Janetzko, D. (in press). Selecting and generating concept structures. In R. Roy (Ed.), *Industrial knowledge management - A micro-level approach*. London: Springer.
- Murphy, G. (1988). Comprehending complex concepts. *Cognitive Science*, *12*, 529–562.
- Osherson, D. N. & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*, 35–58.
- Salvucci, D. D., & Anderson, J. R. (1998). Tracing eye movement protocols with cognitive process models. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 923–928). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Selkirk, E. (1982). *The syntax of words*. Cambridge, MA: Cambridge University Press.
- Suppes, P. (1990). Eye-movement models for arithmetic and reading performance. In E. Kowler (Ed.), *Eye movements and their role in visual and cognitive processes*. New York: Elsevier Science Publishers.
- Wisniewski, E. & Gentner, D. (1991). On the combinatorial semantics of noun pairs: Minor and major adjustments to meaning. In G. B. Simpson (Ed.), *Understanding word and sentence*. Amsterdam: Elsevier Science.
- Zugenmaier, D. & Janetzko, D. (1998). *EDAD - Eye tracking data analysis device* [WWW document] <http://cogweb.iig.uni-freiburg.de/edad>
- Zwitserlood, P. (1994). The role of transparency in the processing and representation of dutch compounds. *Language and Cognitive Processes*, *9*, 341–368.

# The Role of Mental Imagery in Understanding Unknown Idioms

Armina Janyan (ajanyan@cogs.nbu.acad.bg)

Elena Andonova (elena@cogs.nbu.acad.bg)

Department of Cognitive Science, New Bulgarian University

21 Montevideo St., 1635 Sofia, Bulgaria

## Abstract

In studies of the cognitive processing of idioms, the role of mental imagery in understanding idioms remains a controversial issue. Cacciari and Glucksberg (1995) conducted an experimental study to investigate whether generating mental images of idioms can facilitate their comprehension. Their results appeared to reject both the possible connection between the literal mental image of an idiom and the figurative meaning of the idiom, and the facilitatory effect of mental imagery on comprehension. Our study aims at exploring the facilitatory role of mental images in understanding unknown idioms. We used a paraphrase verification task for transparent and opaque unknown idioms translated from foreign languages into Bulgarian. The results demonstrate that literal mental images of transparent unknown idioms can facilitate their comprehension in terms of error scores in a simple paraphrase verification task. No facilitation effect for opaque unknown idioms was obtained. This points towards a link between the literal mental images of transparent idioms and their figurative meanings.

## Introduction

The bulk of cognitive research on idioms is devoted to comprehension processes. Some have investigated the contribution of the literal and figurative meanings of idioms in the comprehension process, and whether both meanings are computed serially or in parallel (Needham, 1990; Estill&Kemper, 1982; Glass, 1982; Swinney&Cutler, 1979); other studies have shown that at some recognition point literal processing stops and the figurative interpretation becomes available (Cacciari&Tabossi, 1988; Tabossi&Zardon, 1993; Titone&Connine, 1994). Another research area explores the tenet that conceptual metaphors constrain or mediate our understanding of idioms (Gibbs&O'Brien, 1990; Nayak&Gibbs, 1990; Gibbs, 1992). Finally, researchers have also studied the strategies that people use to understand tropes and idioms, for example, using the semantics of the constituent words, analogies, metaphorical extensions, etc. (Cacciari, 1993; Flores d'Arcais, 1993). However, relatively little attention has been paid to the role of mental imagery in the process of understanding figurative language. In some theoretical frameworks, imagery is regarded as an important component

in discovering the figurative meaning of tropes and idioms (Lakoff, 1994; Paivio&Walsh, 1998), although experimental studies have produced contradictory results (Gibbs&O'Brien, 1990; Cacciari&Glucksberg, 1995).

Following Lakoff and Johnson's framework (1980), Gibbs and O'Brien (1990) argue that the meanings of idioms are motivated by conceptual metaphors. For example, the idiom *spill the beans* is motivated by the CONDUIT metaphor which specifies the conceptual mapping that THE MIND IS A CONTAINER and IDEAS ARE ENTITIES. Their claim is that people have conventional images and knowledge for the meanings of idioms. To test this, in Gibbs and O'Brien's experiment, subjects were asked to form a mental image of an idiom and describe it verbally. The results suggest that these images have a dynamic nature and people are able to determine the causes and consequences of the actions in them. The data obtained also confirm the expectation of a high degree of consistency in mental images for idioms with similar figurative meanings. Thus, Gibbs and O'Brien (1990) emphasize that conventional images are "unconscious, automatic, and independent of modularity" (p. 39). They do not propose any algorithm of constructing mental images for idioms but they investigate "the products of speakers' mental images for idioms as a way of discovering the knowledge and information that potentially motivate the figurative meaning of idiomatic phrases in English" (ibid.). Finally, they do not claim that people use mental imagery during 'normal' idiom comprehension given that idioms are processed very rapidly. It is children and non-native speakers of a language but not experienced speakers that may form mental images as a way of understanding idioms.

Contrary to the findings of Gibbs and O'Brien (1990), Cacciari and Glucksberg (1995) claim that the images associated with idioms do not reflect their meanings, moreover, forming mental images does not facilitate the comprehension of idioms. They argue that people cannot bypass the literal meaning when processing idioms and forming a mental image, and that it is much easier to form a literal image of an idiom than a figurative abstract one. In this case the images that reflect the literal meaning of an idiom could not refer to the underlying conceptual

metaphors and should interfere with the comprehension of the figurative meaning of an idiom. Thus, these “wrong” literal images would make comprehension more difficult. Note, however, that Lakoff (1994) does not claim that these conventional images must be figurative; on the contrary, they are rather “literal” and include our general knowledge about the world which maps onto the knowledge of the corresponding conceptual metaphor.

Cacciari and Glucksberg's (1995) study includes three experiments. In the first experiment, subjects were asked to give a paraphrase of the idiom, to form a mental image and describe it. Results showed that of the two -- literal vs. figurative -- predominantly images reflecting the literal meanings of the phrases were generated.

Cacciari and Glucksberg's (1995) second experiment explored the issue whether literal images can facilitate comprehension. They reason that if literal images reflect somehow conceptual analogies, then such images would facilitate comprehension; otherwise, if literal images are unrelated to figurative meanings, they would interfere with comprehension or, at the very least, make it more difficult.

Cacciari and Glucksberg (1995) used a sentence-verification task. Subjects were presented with a sentence which they read and then presented with a second sentence that was either a paraphrase of the first sentence or not. In a between-subject design, subjects either performed the verification task while also generating an image of the first sentence; or without generating images. The results show that verification times were longer when influenced by imagery. In addition, the longer times were not associated with a reduction of errors. It is worth noting here that the rate of errors for idioms did not exceed 3% in any of the conditions, although they used four different types of idioms: familiar transparent, familiar opaque, unfamiliar transparent, and unfamiliar opaque. One possible explanation for the strikingly low rate of comprehension mistakes may be that all idioms were in fact highly familiar since in the familiarity pre-test, subjects were explicitly asked to rate their frequency and not familiarity; alternatively, the idioms were semantically transparent.

Overall, Cacciari and Glucksberg's (1995) results obtained suggest that mental imagery interferes with comprehension and does not facilitate it whether measured by reaction time or by error rate, and literal images of idioms have nothing in common with figurative meanings.

In our view, however, forming literal mental images may facilitate the understanding of completely unknown idioms. If images involve general knowledge of the world, if they can be involved in understanding metaphors alongside linguistic knowledge, then understanding may be an interaction of several processes such as applying knowledge, reasoning, mapping, associations. This combination of processes does not necessarily mean that they entail an equally active participation in comprehension. The degree of involvement may depend on the idiom that is being perceived and its properties, as well as on individual experience and contextual factors.

Let us consider the cognitive processing of an unknown idiom. If the unknown idiom reflects a well-known situation, or if it requires reasoning, applying knowledge, or making associations, then it may be that a literal mental image of the idiom can facilitate the comprehension process. For example, consider the Russian idiom *плавает как монополи* (“swims like an axe”). One possible way to understand it is to imagine an axe in water and “see” the axe sinking immediately. The next step could be realizing that the idiom could be referring to a person who cannot swim. So, the concrete-literal mental image can, in principle, lead to an abstract-figurative meaning.

This path from the literal image of an unknown idiom to its figurative meaning, may certainly depend on what kind of idiom it is. Not all idioms have a literal meaning, hence, a literal image that could be created. Moreover, even if such images are easy to produce, not all imageable unknown idioms could thus be understood. For instance, it is hard to understand the Armenian idiom *if a donkey falls, it will break all its teeth* although it is absolutely cartoon-imageable but not transparent in meaning (the idiom refers to ‘a very rocky area’). So, if foreign/unknown idioms are to be understood, they should be semantically transparent and may include some culturally shared concepts. This study attempts to answer some of these controversial issues.

The main aim of the study is to investigate whether generating a literal image of an unknown foreign idiom can facilitate its comprehension. Facilitation here is measured by the error rate and not differences in processing speed. Cacciari and Glucksberg's (1995) line of reasoning that the two phrases to be compared would need to have the same coded representation is indeed convincing. Here an error is defined as failure to recognize the paraphrase of an idiom.

Another purpose of the study is to examine the differences (if any) in comparing an idiomatic meaning with either a literal paraphrase or with an idiomatic equivalent. Such differences may arise because a literal phrase is much more concrete and clear than an idiom. In the case of idioms, often the exact meaning is known but sometimes difficult to put into words, to explain in a succinct and precise form in a short period of time (an analogy with the recognition and naming of a picture). Moreover, idiomatic meanings are often semantically much richer than literal phrases, and idioms can readily map onto much more diverse situations than literal phrases. Hence, comparing the meaning of a known (or the possible meaning of an unknown) idiom to the overall idiomatic meaning of its equivalent would be different from comparing the meaning of an idiom with a literal paraphrase in terms of reaction times (RTs) and/or in terms of the error rate (in percentages).

This experiment examines the on-line processes of generating images, understanding idioms and comparing meanings with two kinds of paraphrases: literal and idiomatic. The method is similar to that used by Cacciari and Glucksberg (1995). The main difference is that unknown foreign idioms were used as target phrases. These idioms were translated from Armenian and Russian word-

by-word into Bulgarian. Subjects had to verify paraphrases under two main conditions, one with, and the other without forming a literal mental image of the target idiom (in a within-subject design). The following is an example of the experimental material (the set) and its translation.

**target idiom**

*чета проповед в ухото на глух*

(to read a sermon into a deaf person's ear)

**paraphrases:**

**related**

**idiomatic phrase** *преливам от пусто в празно*

(to pour from one empty place to another)

**literal phrase** *правя усилия безполезно*

(to make useless efforts)

**unrelated**

**idiomatic phrase** *хвърлям последния си коз*

(to throw down one's last trump card)

**literal phrase** *не си върша задълженията*

(to not complete one's duties)

## Method

**Subjects** A total of 80 subjects (28 males and 52 females) participated in the experiment. All were native Bulgarian speakers, university students. The age range was from 17 to 28. Subjects were paid for their participation.

**Design and Stimuli** A factorial 2x2x2x2 design was used, with RTs and rate of errors as dependent variables. The within-group factors were Imagery task (Imagery, Non-imagery), Source Language of the target idiom (Foreign, Bulgarian), Type of Paraphrase (Literal, Idiomatic), Relatedness of paraphrase to the target (Related, Unrelated). The stimuli consisted of the word-by-word translations of 30 foreign target idioms (16 Armenian and 14 Russian) and 30 Bulgarian target idioms. All target idioms had the form V(PP)NP, and the verb-form was in the first person, singular, present tense. 30 literal and 30 idiomatic paraphrases for the targets were used in the Related paraphrase condition, and 30 literal and 30 idiomatic phrases were used as "false" paraphrases, i.e., unrelated to the target. The average length (in words) of target phrases was 4.2, of literal paraphrases - 2.9, and of idiomatic paraphrases - 3.7 words. The selected foreign target stimuli did not include similes, and paraphrases did not include words semantically related to the targets.

All 150 Bulgarian idioms were selected after a pre-test with independent familiarity and frequency ratings using a 5-point scale (5 -- most familiar or most frequent, respectively). A total of 28 subjects participated in the pre-test. The age range was from 18 to 30. None of them participated in the main on-line experiment later. The idioms thus selected from the pre-test had the mean value of 4.3 for familiarity and of 3.2 for frequency of use.

**Procedure** 16 randomized lists of 60 sets of stimuli each were constructed so that each subject was presented with all the 60 target idioms and one out of four paraphrases. The

experiment was divided into two main parts, named Non-imagery and Imagery, after the two tasks. Every subject was run on both parts. In the first Non-imagery condition the procedure was the following. The target idiom appeared on a white background, at the center of the screen. Subjects had to read the idiom, decide whether they know it or not (familiarity decision) and press a corresponding button (Yes or No). Immediately after the response (zero inter-stimulus interval), a paraphrase appeared on a light grey background. Subjects then performed a phrase-verification task, i.e., they had to decide whether the meaning of the paraphrase matches the meaning of the target idiom, and press the corresponding button (Yes or No). Each trial began with a central black fixation marker ('+') for 500 ms and the inter-trial interval was 3 sec. The reaction time was measured from the onset of the stimulus (paraphrase) till the subject's response. After the first Non-imagery condition, subjects had a 5-minute break. In the second, Imagery condition, subjects had to read the target idiom, imagine it as a "picture" and press a Yes button immediately afterwards. The remaining procedure was the same as in the first experimental condition (Non-imagery), i.e., subjects performed a phrase verification task. Both experimental conditions started with 8 practice trials.

Every 60-trial list of stimuli was randomly divided into two subsets, consisting of 30 paired stimuli for each of the two experimental conditions (Imagery and Non-Imagery). The assignment of Yes and No values to the buttons was counterbalanced across subjects. The experiment lasted approximately 30-40 min. Subjects were tested individually in a sound-proof room. A Power Macintosh 6400/200 equipped with PsyScope software controlled stimuli presentation, timing, and response collection.

## Results and Discussion

The analysis was carried out by items averaged over subjects. RTs and responses for the phrase verification task were analyzed in a 2 (Imagery vs Non-imagery) x 2 (Foreign vs Bulgarian targets) x 2 (Literal vs Idiomatic paraphrases) x 2 (Related vs Unrelated paraphrases) analysis of variance.

### Analysis of Phrase Verification Task

**Reaction Times (RTs)** Main effects for all four independent variables on RT were obtained. There was a significant main effect of Imagery ( $F_{(1,420)}=37.91$ ;  $p<0.00$ ); the phrase-verification task took longer to complete in the imagery condition (mean response time of 2475 ms,  $SD=589$ ) than in the non-imagery condition (mean  $RT=2179$  ms,  $SD=537$ ). This may mean that the mental image of the target idiom interferes with the linguistic representation of the paraphrase which replicates the results of Cacciari and Glucksberg (1995).

A main effect of Source Language was also obtained ( $F_{(1,420)}=77.46$ ;  $p<0.00$ ). For Bulgarian target idioms, paraphrase verification took less time ( $MRT=2116$  ms,  $SD=506$ ) than for Foreign targets ( $M=2539$  ms,  $SD=578$ ).

Thus, phrase verification was significantly faster for the familiar-familiar pairs than for the unknown-familiar pairs.

The main effect of the Type of Paraphrase also reached significance ( $F_{(1,420)}=8.16$ ;  $p<0.00$ ). The mean RT for verification of literal paraphrases was shorter (2259 ms,  $SD=617$ ) than for idiomatic paraphrases (2396 ms,  $SD=542$ ). This may be due to the different lengths of the paraphrases and/or the less ambiguous meaning of the literal phrase compared with the idiomatic one.

The main effect of Relatedness of Paraphrase was also significant ( $F_{(1,420)}=14.72$ ;  $p<0.00$ ). The mean RT for the Related Paraphrase condition was faster (2235 ms,  $SD=602$ ) than for Unrelated (2420 ms,  $SD=552$ ).

A significant interaction between Source Language and Relatedness ( $F_{(1,420)}=18.60$ ;  $p<0.00$ ) was also found, i.e. the main effect of Relatedness of Paraphrase was not observed in the Foreign condition. Mean response times are presented in Table 1.

Table 1: Mean RTs (ms) for Source Language by Relatedness

	Bulgarian	Foreign
Related	1917 ms	2551 ms
Unrelated	2314 ms	2537 ms

The main effect of Relatedness is visible in these results as well in that the phrase verification task took less time when there was a real paraphrase (the phrase was related to the target). Note, however, that this effect holds only for the familiar Bulgarian targets. Relatedness did not make a difference to the processing of the semantic comparison between paraphrases and unfamiliar targets. Not surprisingly, the task was performed overall faster with familiar than with unfamiliar targets.

**Error Rate (%)** Three main significant effects were found. In the main effect of Source Language the familiarity (Bulgarian target) advantage was obtained ( $F_{(1,420)}=50.20$ ;  $p<0.00$ ); the error rate was lower for Bulgarian targets (12.5%) than for foreign ones (24.6%). The main effect of Type of Paraphrase ( $F_{(1,420)}=16.87$ ;  $p<0.00$ ) showed the advantage of literal paraphrases (15.3% error rate) over idiomatic ones (21.9%). The main effect of Relatedness ( $F_{(1,420)}=37.40$ ;  $p<0.00$ ) revealed fewer comprehension errors for unrelated paraphrases (13.6% of "false alarms") vs. related ones (24.0% of "misses"). These results may be partly explained by the fact that unrelated literal paraphrases were concrete and unusual to serve as possible paraphrases of idioms. Subjects may have chosen a strategy to reject these cases due to their obvious unrelatedness to the target idioms. No significant overall main effect of the Imagery factor was found.

The only significant interaction obtained was that between Relatedness and Source Language ( $F_{(1,420)}=5.06$ ;  $p<0.03$ ). Mean rates of errors are presented in Table 2.

Table 2: Mean Percentage of Errors for Source Language by Relatedness

	Bulgarian	Foreign
Related	16 %	32 %
Unrelated	9 %	18 %

Significant differences across all combinations of the four conditions were found (except for Bulgarian Related and Foreign Unrelated conditions). In both cases unrelated paraphrases were verified with better success than related ones. This again may be partially explained by the way the unrelated literal paraphrases were selected. Overall, paraphrases for Bulgarian target idioms were verified with a lower error rate than the foreign ones.

### Separate analysis by Source Language, Error Rate (%)

In order to reveal the contribution of imagery, a separate analysis over the two levels of Source Language was conducted. For Bulgarian target idioms, no effect was found but for Foreign target idioms, there was a main effect of Imagery ( $F_{(1,215)}=3.94$ ;  $p<0.05$ ). The Imagery condition showed an advantage (only 22% of errors) over the Non-imagery condition (27% error rate). The absence of the imagery effect on Bulgarian targets showed that imagery had no facilitatory effect on the processing of familiar idiom but it did on unfamiliar ones.

To explore the nature of the Imagery effect further, a post-test on the levels of transparency of foreign idioms was carried out.

**Post-test** The 30 foreign idioms were randomly assigned to two separate questionnaire lists, with each idiom placed on a separate sheet of paper. Subjects were 26 native Bulgarian speakers who were asked to guess the meanings / paraphrases of these unfamiliar idioms. There was no time limit in completing the task. The responses were evaluated for accuracy by two independent judges and averaged as the percentage of correct answers for each idiom. On this basis, idioms were categorized as transparent (correct guesses exceeding 60%) and opaque (lower than 60%). As a result, 15 transparent (Mean=74%,  $SD=14$ ) and 15 opaque (mean=20%,  $SD=14$ ) idioms were identified.

### Analysis of Phrase Verification for Foreign targets only

Responses and RTs for items averaged over subjects for the phrase verification task were analyzed in a 2 (Imagery vs Non-imagery) x 2 (Transparent vs Opaque) x 2 (Literal vs Idiomatic paraphrases) x 2 (Related vs Unrelated) analysis of variance.

**Reaction Times (RTs)** The overall main effect of imagery on RT was repeated here as well ( $F_{(1,207)}=19.79$ ;  $p<0.00$ ) with subjects being faster in the Non-imagery condition (2375 ms,  $SD=503$ ) than the Imagery (2706 ms,  $SD=601$ ) one. A significant two-way interaction of Transparency and

Relatedness ( $F_{(1,207)}=15.22$ ;  $p<0.00$ ) was also found. Mean response times are presented in Table 3.

Table 3: Mean RTs (ms) for Transparency by Relatedness

	Transparent	Opaque
Related	2364 ms	2731 ms
Unrelated	2647 ms	2432 ms

There were significant differences across all combinations of the four conditions. It is important that a similar trend is observed here in verifying familiar-familiar pairs and transparent unknown-familiar pairs. In both cases Related paraphrases were verified faster than Unrelated (cf. Tables 1 and 3), with the implication that transparent idioms may be treated as familiar, and similar mechanisms may be involved in their processing in the verification task. For Opaque idioms, the verification time changed in the opposite direction.

**Error Rate (%)** Four significant main effects were found on the rate of errors as a dependent variable: Type of Paraphrase ( $F_{(1,207)}=13.91$ ;  $p<0.00$ ), Relatedness ( $F_{(1,207)}=37.36$ ;  $p<0.00$ ), Imagery ( $F_{(1,207)}=5.38$ ;  $p<0.02$ ), that were replications of the previous discussed, and Transparency ( $F_{(1,207)}=16.65$ ;  $p<0.00$ ) that showed lower rate of errors for transparent idioms than for opaque (20% vs 29%). Two significant two-way interactions were also obtained: Transparency by Relatedness ( $F_{(1,207)}=21.73$ ;  $p<0.00$ ), and Paraphrase by Relatedness ( $F_{(1,207)}=5.97$ ;  $p<0.02$ ). Two significant three-way interactions, Imagery by Transparency by Relatedness ( $F_{(1,207)}=4.38$ ;  $p<0.04$ ) and Imagery by Transparency by Type of Paraphrase ( $F_{(1,207)}=4.03$ ;  $p<0.00$ ), are shown in Figures 1 and 2, respectively. The mean error values for every condition are shown in Tables 4 and 5.

As Figure 1 and Table 4 demonstrate, there was a significant shift in the rate of errors for transparent idioms in the Related paraphrase condition (29% in the Non-imagery vs 14% in the Imagery condition). No imagery effects were found on either semantically transparent or opaque idioms in the Unrelated paraphrase condition, as well as on the opaque idioms in the Related paraphrase condition. The lack of significance and rather low rate of errors in the Unrelated paraphrase condition can be partially attributed to the way the stimuli for the literal unrelated phrases were selected.

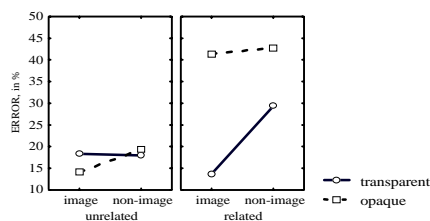


Figure 1: Imagery by Transparency by Relatedness interaction

Table 4: Mean Error Rates for Imagery by Transparency by Relatedness

	Transparent		Opaque	
	Image	N-image	Image	N-image
Related	14 %	29 %	41 %	43 %
Unrelated	18 %	18 %	14 %	19 %

The facilitatory role of imagery (Figure 2, Table 5) was also observed in the verification task results particularly for idiomatic paraphrases of transparent idioms. In the Imagery condition the rate of errors was reduced down to the level of literal paraphrases in both Imagery and Non-imagery conditions. Literal paraphrases of transparent idioms were verified with the same success (low rate of errors) as paraphrases for familiar Bulgarian target idioms (Table 2). There was no imagery effect on opaque idioms although a trend toward improved comprehension in the Imagery condition may be observed in the case of literal paraphrases.

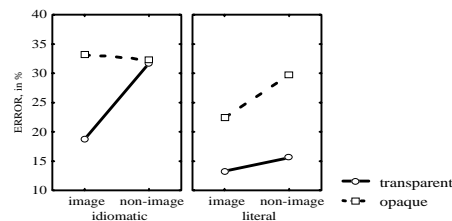


Figure 2: Imagery by Transparency by Type of Paraphrase

Table 5: Mean Error Rates for Imagery by Transparency by Type of Paraphrase

	Transparent		Opaque	
	Image	N-image	Image	N-image
Idiomatic	19 %	32 %	33 %	32 %
Literal	13 %	16 %	22 %	30 %

## Conclusion

The main goal of this study was to explore the facilitatory role of mental images in comprehending unknown idioms. Results have shown that indeed, constructing literal images of unknown idioms can help in understanding the idiom, not in terms of faster processing but in terms of decreasing percentage of mistakes in recognizing a paraphrase of the idiom. This effect is particularly salient in the processing of semantically transparent idioms and is stronger when unknown idioms are compared with an idiomatic paraphrases and not with literal ones. Furthermore, the contribution of mental imagery is such that it produces fewer mistakes of the “miss” type for transparent idioms, i.e., subjects improve their ability to recognize a real paraphrase as equivalent in meaning to the target. Semantically opaque idioms, on the other hand, seem to be indifferent to the imagery task, though a trend toward better



understanding may be observed in the case of literal paraphrases.

Another aim of the study was to test whether different types of paraphrases (literal and idiomatic) could influence the degree of understanding unknown idioms. The hypothesis was that since idioms may be viewed as semantically broader or more vague than literal paraphrases subjects would more readily match idiomatic paraphrases with unknown target idioms than literal paraphrases. As a result they would make fewer mistakes with idiomatic paraphrases than with literal ones. This hypothesis was rejected by the results which revealed the opposite picture - subjects made considerably fewer mistakes with literal paraphrases than with idiomatic ones. One possible explanation derives from the same feature of idioms, i.e., their semantic and 'situational' broadness which may have caused subjects to reach a negative decision on the verification task much more frequently than necessary, hence, these results.

To conclude, the results show that transparency plays only a minor role in comparison with familiarity, and that familiarity itself is only useful as a concept in its own right, not by proxy of frequency. The results also demonstrate that constructing a literal image helps our understanding of unknown transparent idioms whether by unconscious applying general knowledge of the world, unconscious reasoning or some other process involved in understanding. Thus, there exists a close link between figurative meanings of transparent unknown idioms and their literal mental images.

This study, however, has helped to explain further the mechanisms of comprehension of unknown idioms and the role of mental imagery in this process. It remains to be seen whether mental imagery facilitates not only the comprehension but also the process of learning and retrieving from memory of figurative speech.

### Acknowledgements

We are grateful to the people who helped us a lot in different ways: Daniela Angelova, Irina Gerdjikova, Radu Luchianov, Agop Erdeklian, Milena Leneva, Poly Dacheva, and Sonia Tancheva.

### References

- Cacciari, C. (1993). The Place of Idioms in a Literal and Metaphorical World. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, Structure, and Interpretation*. NJ, US: Lawrence Erlbaum Associates, Inc.
- Cacciari, C. and Glucksberg, S. (1995). Understanding Idioms: Do Visual Images Reflect Figurative Meanings? *European Journal of Cognitive Psychology*, 7(3), 283-305.
- Estill, R. B. & Kemper, S. (1982). Interpreting Idioms. *Journal of Psycholinguistic Research*, 11(6), 559-568.
- Flores d'Arcais, G. B. (1993). The Comprehension and Semantic Interpretation of Idioms. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, Structure, and Interpretation*. NJ, US: Lawrence Erlbaum Associates, Inc.
- Gibbs, R.W., JR. & O'Brien, J. (1990). Idioms and Mental Imagery: The Metaphorical Motivation for Idiomatic Meaning. *Cognition*, 36, 35-68.
- Gibbs, Raymond W., JR. (1992) What Do Idioms Really Mean? *Journal of Memory and Language*, 31, 485-506.
- Glass, A. L. (1983). The Comprehension of Idioms. *Journal of Psycholinguistic Research*, 12(4), 429-442.
- Lakoff, G. & Johnson, M. (1980). *Metaphors We Live By*. Chicago: Chicago University Press.
- Lakoff, G. (1994). What is Metaphor? In J. Bardner & K. Holyoak (Eds.) *Advances in Connectionist and Neural Computational Theory. Analogy, Metaphor, and reminding*. Vol.3. Norwood, NJ: Ablex.
- Nayak, N. P. & Gibbs, R. W., Jr. (1990). Conceptual Knowledge in the Interpretation of Idioms. *Journal of Experimental Psychology: General*, 119(3), 315-330.
- Needham, W. P. (1992). Limits on Literal Processing During Idiom Interpretation. *Journal of Psycholinguistic Research*, 21(1), 1-16.
- Paivio, A., & Walsh, M. (1998). Psychological Processes in Metaphor Comprehension. In A. Ortony (Ed.), *Metaphor and Thought*, 2<sup>nd</sup> Ed, Cambridge: University Press.
- Swinney, D. A., & Cutler, A. (1979). The Access and processing of Idiomatic Expressions. *Journal of Verbal Learning and Verbal Behavior*, 18, 523-534.
- Tabossi, P. & Zardon, F. (1995). The Activation of Idiomatic Meaning. In M. Evaraert, E.-J. van der Linden, A. Schenk, R. Shreuder (Eds.). *Idioms: Structural and Psychological Perspectives* NJ, US: Lawrence Erlbaum Associates, Inc.
- Titone, D. A., & Connine, C. M. (1994). Comprehension of Idiomatic Expressions: Effects of Predictability and Literality. *Journal of Experimental Psychology: Learning, memory, and Cognition*, 20(5), 1126-1138.

# Does Collaborative Learning Lead to the Construction of Common Knowledge?

Heisawn Jeong (heis@pitt.edu)

Micheline T. H. Chi (chi@pitt.edu)

Learning Research and Development Center; University of Pittsburgh  
3939 O'Hara Street, Pittsburgh, PA 15260 USA

## Abstract

This study investigated whether collaborative learning leads to the construction of shared knowledge among participants. In this study, college student pairs collaborated to learn a biology text on the human circulatory system. The results showed that pairs shared not only correct knowledge that was presented in the text, but also incorrect knowledge and/or knowledge that had to be inferred from the text. In addition, pairs who interacted more shared significantly more inferred knowledge than those who interacted less did. Taken together, these findings indicate that interaction enables dyads to construct new knowledge and their representations tend to converge after collaboration.

## Introduction

Traditional cognitive psychology has mainly focused on how information is processed within individuals' minds, that is, how individuals represent stimuli, learn new things, solve problems, make a discovery, etc. As a consequence, even when people learn collaboratively, learning has been mainly defined in terms of what individuals learn and not much attention has been paid to the collaborative aspect of knowledge construction that is shared by both partners. This study to be described below was an attempt to examine whether a shared activity such as collaborative learning would lead to the construction of shared knowledge.

Learning often occurs in the context of a group or community, and many researchers propose that learning, a *cognitive* activity, is a joint *social* activity (Lave & Wenger, 1991; Levine, Resnick, & Higgins, 1991; Resnick, Levine, & Teasley, 1991; Rogoff, 1998; Tudge & Rogoff, 1989; Vygotsky, 1978). To say that a cognitive activity is a social activity sounds contradictory at first, but the notion of "socially shared cognition" has been instantiated in two ways: It has been used to refer to shared cognitive activities such as group problem solving (e.g., Larson & Christensen, 1993) or shared representations such as a team mental model (e.g., Klimoski & Mohammed, 1994) or a community memory (e.g., Orr, 1992). Even if we operationalize socially shared cognition to be shared representations that all members of a group have in common, the question remains as to whether this shared representation can come about from a joint social activity such as collaborative learning.

In order to hypothesize whether a shared representation is constructed, we need to reconsider what individual learning

is. Learning requires people to process incoming material (such as an expository text) and to integrate it with their prior representations. Thus, we assume that when individuals can learn by themselves, and they do so by actively constructing new knowledge or skills and/or revising their incorrect understanding (Chi, in press). Although such active construction of knowledge is critical regardless of whether people learn alone or together (Jeong, under review), learning in a collaborative context gives rise to an additional question: Is each member of the dyad constructing and revising her own individual representation, or are they jointly constructing one representation, or is it a hybrid of the two?

If learning is the construction and revision of representation, then there are at least two hypotheses about what might be happening during collaborative learning. In the first case, collaborating partners may each be constructing and revising her own representation, taking the partner's comments and explanations simply as additional input or feedback. In this case, they would each be constructing their own representation, albeit simultaneously. So, partner A and B each would construct their own unique representations, A and B. Representations A and B may be totally distinct or they may overlap, but since partners are constructing and revising their own representations, their representations would not resemble one another. The overlap in their representations, if existed at all prior to collaboration, would not likely to change with collaboration either.

On the other hand, another possibility is that collaborating partners may be constructing and revising jointly a shared representation C. Regardless of whether the shared representation C is constructed *in addition to* or *instead of* their own representations A and B, the resulting representations would reflect the joint learning activity they engaged in during collaboration in that partner A's representation shares portions that are similar to partner B's representation. With collaboration, A and B share more and more parts in common, so that the common representation C gets larger with greater collaboration, whereas the representations unique to A or B would get smaller with collaboration (see Figure 1).<sup>1</sup>

---

<sup>1</sup> Although not examined in this study, a third possibility is that collaboration might encourage the two partners to construct a single representation that is either A or B (that is, they converged upon one of the partner's representation), or neither A nor B, but X.

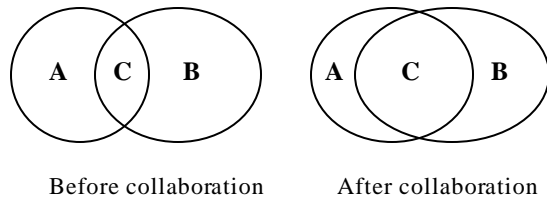


Figure 1: Changes in the shared representation due to collaboration.

In this paper, we define “shared knowledge” as the knowledge that is common to both partners’ representations. The question that we need to address is whether this common knowledge is indeed constructed from collaboration. For example, if people knew that alcohol in moderation reduces the incidence of heart attack, then this is a piece of common knowledge. Such a piece of common knowledge is likely to arise from similar experiences with the environment, but not necessarily from collaborative construction in face-to-face interaction. Members of a group or a culture would possess a set of common knowledge as a result of exposure to the same news media (e.g., there was a fire in New York last month), textbooks (e.g.,  $e=mc^2$ ) or simply being in the same culture (e.g., it is okay to eat in a classroom).

Researchers from anthropology and linguistics have observed that various kinds of groups that have histories of interaction tend to share a set of common knowledge. For example, Orr (1990) reported that people who practice the same job (e.g., photocopier repair technicians) hold a community memory about machines and customers. Similarly, teams are reported to have a shared mental model about their task requirements, procedures, and their responsibilities, which in turn helps them to work more efficiently especially in emergencies (Cannon-Bowers, Salas, & Converse, 1993; Klimoski & Mohammed, 1994; see also Hazlehurst, 1994 and Sherif, 1936). Thus, it seems that there exists common knowledge, whether it is memory, team mental model, or a way to interpret an image, common to interacting group members in the real world.

However, these observations of common knowledge may arise because the group all experienced the same input, rather than because they co-constructed it. It is difficult to tease these two interpretations apart, because people often share input as well as co-construct during collaboration. Thus, to clearly differentiate the two, we need to examine and demonstrate that some new or incorrect knowledge, *knowledge that cannot be experienced directly from the environmental input* (such as incorrect knowledge or inferred knowledge) has been constructed and shared after collaboration and that construction of such knowledge is clearly linked to the extent of interaction.

Very few studies have even attempted to capture the existence of shared knowledge that resulted from collaboration rather than direct experience. Roschelle’s (1992) study did attempt to show a convergence of representation while one high school student pair was learning physics concepts of velocity and acceleration. The students’ representation became more similar to each other’s after collaboration. How-

ever, it was difficult to assess whether the resulting converged representation reflects knowledge that both experienced or was co-constructed, especially based on a single case.

In general, the evidence for socially shared cognition has relied on qualitative evidence based on a select few cases. One of the goals of this study was thus to provide quantitative evidence using clear operationalization of shared knowledge. In this study, college student pairs were asked to collaborate to learn a biology text about the human circulatory system. We were interested in exploring and identifying the role of several variables in collaborative learning, their collaboration was unstructured other than the instruction to collaborate and to talk. Students were individually pre-tested and only those who had inaccurate mental models about the circulatory system (see method section for what constitutes inaccurate models) were allowed to participate. After the pre-tests, students were paired with another student who was equally naive about the topic and collaborated to learn the text and then came back for an individual post-test. Students were given two tests before and after collaborative learning: Terms Task and Blood Path Drawing Task. The Terms Task was to assess what students knew about the topic, specifically about various terms important in understanding the circulatory system. The Blood Path Drawing Task was to assess what students knew about the blood flow in the human body.<sup>2</sup>

## Method

### Participants

Twenty (nine male and eleven female) pairs of undergraduate students at the University of Pittsburgh participated in the study for course credit. Students were asked to participate if they had not taken any college-level biology classes. Students were asked to stay in the study only if they had inaccurate models (see later coding section) at the pre-test, did not have relevant personal experiences (e.g., open heart surgery), and could be paired with another student of the same gender who could come in around the same time for collaborative learning session. The pairs did not interact with their partner prior to the study except in one pair.

### Materials

**Text** The text used in Chi et al. (1994) was used with a slight revision (the text was originally taken from the chapter on the human circulatory system in a high school biology text by Towle, 1989). The resulting text contained 73 sentences. They were presented in a binder with each sentence printed on a separate page.

<sup>2</sup> A set of Knowledge Questions was also given to students, but are not included in this paper. It was administered at the post-test (after Terms and Blood Path Drawing Tasks) and did not allow comparison between pre-test and test as to how much new common knowledge was constructed after collaboration.

**Terms Task** Students were given 19 terms about the human circulatory system (e.g., atrium), and asked to talk about everything they knew about each term, even if it seemed unimportant to them.

**Blood Path Drawing Task** Students were provided with an outline of a human body (with a heart in it) and asked to draw the blood path of the circulatory system. They were asked to talk about everything that came to their mind as they drew.

## Procedures

**Pre-test** Participants were tested individually on the Terms Task and the Blood Path Drawing Task. This session was audio-taped. The pre-test session took about 15 to 30 minutes. At the end of the session, they were asked not to do outside reading on this topic while the study was in progress.

**Collaborative Learning** About a week after the pre-test, students were paired with a partner to learn the text. Since most of them had never met each other, they were given some time to get to know each other before the session started: after the experimenter introduced them and initiated conversation (e.g., who were their psychology instructors), she left the room supposedly to check the equipment. The experimenter watched their interaction from a monitor in another room until they seemed to be comfortable with each other. Most students quickly established a rapport with each other (in about ten minutes), discovering a common friend or exchanging information about classes.

Students were asked to help and encourage each other to learn and understand the materials during the collaborative learning session. They were asked to read the text out loud at least once. Participants were informed that they would be tested after the learning was over (a few sample test questions were provided). The pairs shared the text binder, and were provided with paper and pens in case they wanted to take notes or draw. This session was audio- and videotaped. The experimenter was not present in the room during this session, but could hear and watch them from the control room. Participants knew that the experimenter could hear them, but not necessarily that she could watch them. They were allowed to take as much time as they needed to study the text. The actual learning session took about an hour on average, ranging from 40 minutes to one hour and 45 minutes.

**Post-test** Participants were tested individually on the Terms task and Blood Path Drawing Task about a week after the collaborative learning session. This session was audio-taped. Post-test sessions ranged from 45 minutes to 2 hours.

## Coding

All the sessions were transcribed. From the protocol, three measures were collected. First, individual knowledge pieces (KPs) students knew were coded using a template from the students' answers in the pre-test and post-test. Second, students' mental models about the human circulatory system

were analyzed. Third, turns the two students took during the collaborative learning session were coded in terms of whether the turn was relevant to their partner's previous contributions. A more detailed coding scheme is reported below along with the reliability measures. A second coder coded 20% of the data independently from the first coder. The coding of the first coder was used throughout.

**Template Scoring** A template was created to assess how much students knew about the topic presented in the learning text. The template was created based on the information presented in the text. The 73 sentences in the learning text were segmented and collapsed into individual knowledge pieces (KPs) that roughly corresponded to a proposition (e.g., "aorta is an artery"). The template contained a total of 173 KPs. There were two types of KPs: KPs that were explicitly stated in the text (Stated KPs) and KPs that could be inferred from the text (Inferred KPs). An example of a Stated KP is "atrium is the upper part of the heart" which is directly stated in sentence 20 "Each upper chamber is called an atrium." An example of an Inferred KP is "the heart has four chambers." This KP is not explicitly stated in the text but can be inferred by integrating sentence 17 "The septum divides the heart lengthwise into two sides" and sentence 19 "Each side of the heart is divided into an upper and a lower chamber." The template contained 115 Stated KPs and 58 Inferred KPs. The KPs were coded from the students' protocol during the Terms and the Blood Path Drawing Task. The agreement between the two coders was 87%.

**Mental Model Analysis** Students' initial and final mental models about the human circulatory system were coded to assess changes in how individual knowledge is integrated to form a coherent model of the circulatory system as a whole. Based on students' protocols during the Terms and the Blood Path Drawing task, each student's initial and final mental models were coded into one of the following models: (1) No Loop (NL) model, (2) Ebb and Flow (EF) model, (3) Single Loop (SL) model, (4) Multiple Loop (ML) model, (5) Single Loop with Lungs (SLL) model, (6) Double Loop-1 (DL1) model, and (7) Double Loop-2 (DL2) model. The seven models differ from each other in terms of the presence and the kind of incorrect conceptions (e.g., blood returns to the heart by way of the same blood vessels) and/or the correct conceptions (e.g., heart pumps blood to the lungs versus left ventricle pumps blood to the lungs). Both the Double Loop-1 and Double Loop-2 models represent the accurate flow of blood through the circulatory system with Double Loop-2 being the most complete model (see Chi et al., 1994 for more details on this analysis). The inter-rater agreement on mental model coding was 94%.

**Turn-taking** Each turn that a student took during collaboration was coded whether or not it was relevant to their partner's previous turn. A turn can be relevant in several different ways. A turn was coded as relevant, for example, if students answered questions that their partner asked, repeated and/or continued the statement and topic that their partner initiated, or acknowledged what their partner said. A turn

was defined in this study as a change of speaker in their learning dialogue. The transcript occasionally contained non-verbal (e.g., laughs, gestures) turns, but it was coded as a turn if it had information potentially relevant to the partner. Thus, turns that contain only “ok” or “umm” were coded as a separate turn when it could be answers or acknowledgements. Similarly, turns that contained only gestures were coded as a separate turns if it was communicative (e.g., nodding indicating “yes”). Based on this identification of turns, a second pass over the transcript was done to determine whether each turn was “relevant” to their partner’s previous turns. A turn was coded as relevant as long as the turn contained information relevant to their partner’s previous contribution in some way (see Jeong, under review, for more details). The reliability for this coding was 85%.

## Results

The process and outcome of knowledge construction were considered to be interdependent between the two members of the pair in this study. Thus, the unit of analysis in this study was pairs rather than individuals. Although students’ pre-test scores were mostly independent from each other’s (unless we start considering cultures), their post-test scores, although tests were individually administered, were partly dependent on their partner’s score due to their collaboration. Thus, we calculated common KPs as well as unique total KPs to deal with this dependency. In this section, we first describe how much learning occurred and how much common knowledge was constructed after collaborative learning. We then examine in more detail whether the increase in common knowledge was indeed co-constructed from interaction.

### Learning and Common Knowledge

Learning was assessed by addressing (1) the number of Knowledge Pieces (KPs) that were learned after collaborative learning and (2) improvement in the pairs’ mental model.

**Template Scoring: Knowledge Pieces (KPs)** Since template scoring gives scores for each partner, the amount of knowledge that the *pairs* knew as a whole was calculated by: (a) an average score of the two students in the pair and (b) a *unique total* score. These scores can be best understood by looking at Figure 1. Circle A represents what Partner A knows, Circle B represents what Partner B knows. *Common knowledge* is defined as the knowledge that both partners possess, represented by the area C, the overlap of the two circles. For example, if both partners know the KP that the heart has four chambers, then they are said to share that piece of common knowledge. On the other hand, *unique knowledge* is defined as the knowledge that only one member of the pair possesses. Learning for each individual is represented by an increase in the size of each circle (A, B) from the pre-test to the post-test. On the other hand, learning for the pair as a whole can be best represented by examining a *unique total* score that represents the number of distinctive KPs that the pair knew as a whole (see Table 1).

The average score of pairs increased significantly from 19.70 KPs at the pre-test to 47.70 KPs at the post-test,  $t(19)=10.08, p<.001$ , and the unique total score also increased significantly from 32.15 KPs at the pre-test to 72.85

Table 1: The relationship between various scores.

Scores	Areas in Figure 1
Common KPs	C
Unique KPs	(A-C) or (A-B)
Average KPs	(A+B)/2
Unique total KPs	(A+B-C)

KP at the post-test,  $t(19)=11.92, p<.001$ , indicating that students’ understanding of the human circulatory system increased significantly after collaborative learning. The amount of common knowledge also increased significantly from 7.25 KPs at the pre-test to 22.55 KPs in at post-test,  $t(19)=6.13, p<.001$ .

**Mental Models** Consistent with the overall gain in the KPs, there was an overall improvement in students’ individual mental models about the circulatory system after learning. Recall that none of the students had the correct Double Loop models at the pre-test, since students were selected that way for the study. The majority of them started with the Single Loop model (55%), followed by the Single Loop with Lungs model (25%). After learning, the majority of the students possessed the most accurate and complete Double Loop-2 model (52.25%), followed by the next most accurate Double Loop-1 model (37.5%). Thus, learning the text with a partner improved the accuracy of the students’ individual mental model as well as increasing the number of individual knowledge pieces that they knew, as in Chi et al. (1994).

To determine whether partners’ mental models converged onto the same model, each student’s mental model was compared to their partner’s. At pre-test, 10 pairs (50%) had different initial incorrect models. As stated earlier, none of the models was the correct Double Loop models. Six of these 10 pairs converged onto the same final mental models. However, five of the six pairs’ final models were the correct Double Loop-2 model, so we cannot rule out the interpretation that each partner’s model converged on the correct model, independent of interaction.

### Collaboration and Common Knowledge

Students were not coming up with arbitrary knowledge (e.g., how to name an ambiguous geometric figure) in this study. They were learning a science text that strongly constrains their interpretation and knowledge construction. Since they all learned the same text, the increase in common knowledge and the convergence toward the correct final model, could be the result of individuals learning the same materials from the text rather than their collaboration. In this section, we further examined whether there was any evidence that collaborative dyads co-constructed knowledge from interaction, rather than merely self-constructing their own knowledge, in the presence of an enabling partner.

**Common and Unique Knowledge** Although pairs had more common knowledge after collaborative learning, they also knew more after learning. Thus, just looking at the number of common KPs could give a false picture without taking into account the increase in total amount of knowledge due to learning. To address this problem, the percentage of common knowledge (over the unique total KPs) was calculated. The percentage of common knowledge increased after collaboration (from 23% to 31%), whereas the percentage of unique KPs decreased after collaboration (from 77% to 69%),  $F(1, 19)=11.05, p<.005$ . This significant interaction indicates that the increase in common knowledge was not a mere reflection of knowing more. In sum, after collaborative learning, pairs gained more KPs overall, but they learned proportionately more common knowledge than unique knowledge.

**Nominal Pair Analysis** If some parts of common knowledge is co-constructed (rather than learned individually by each partner), then collaborative pairs ought to learn more common knowledge than nominal pairs who did not collaborate. A hypothetical nominal pair was constructed by randomly pairing each member of the pair with a member of another pair. The results showed that there was an increase in common KPs in nominal pairs as in real pairs, but the increase was greater in real pairs (8% versus 4%). Although ANCOVA (controlling for their pre-test scores), did not reveal significant difference between the two conditions,  $F(1,36)=2.36, p<.14$ , the increase in the proportion of common knowledge from pre-test to post-test was significant in real pairs,  $t(19)=2.8, p<.01$ , but not in nominal pairs,  $t(19)=1.20, p>.10$ . Thus, although part of the common knowledge constructed during collaboration was due to learning from the same text (as can be seen in the small increase of shared knowledge in nominal pairs), it seems that part of the increase in common knowledge can be undoubtedly attributed to collaboration.

**Incorrect Knowledge Pieces** We also examined incorrect knowledge at the knowledge piece level from the pre-test and post-test answers. In total, pairs had 69.25 incorrect KPs at the pre-test and 91 KPs at the post-test. Out of these, the real pairs did not share any incorrect KPs at the pre-test, but shared a total of 4 at the post-test after collaboration. On the other hand, nominal pairs had a total of 3 common incorrect KPs at the pre-test, but 0 KP at the post-test. Although the numbers are small, the fact that pairs shared 4 incorrect KPs after collaboration suggest that these incorrect KPs must have been co-constructed with their partners during collaboration, rather than encoded and inferred from the text alone independently from their partner.

**Common Incorrect Mental Model** As mentioned earlier, six of the ten pairs of students who had different initial mental models converged onto the same final mental model. Of these six pairs, five pairs converged on the correct Double Loop-2 model, which could be attributed to having read a text that described such a correct model. One pair, however, converged on an incorrect model. Both of their models had

the same “error”: They both thought that blood from the lungs goes back to the heart through the ventricle, rather than through the atrium as in the correct model (see Figure 2). Thus, an incorrect model that both partners share strongly indicates that they somehow co-constructed it.

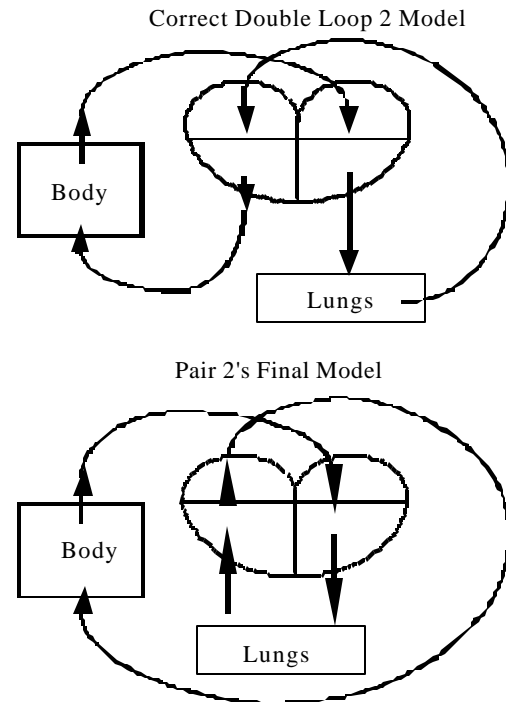


Figure 2: Pair 2's final mental model in comparison to Double Loop-2 model.

**Interaction and Common Knowledge** If dyads co-constructed common knowledge from interaction, rather than merely self-constructed their own knowledge, it would suggest that the more interaction they engaged in, the more common knowledge they would construct, especially the knowledge that cannot be obtained directly from the text, that is, knowledge that need to be inferred. To test this hypothesis, the pairs were grouped into high-interaction pairs (N=10) and low-interaction pairs (N=10) based on the amount (percentage) of relevant turns they took during collaborative learning.

As can be seen in Figure 3, high-interaction pairs shared more inferred knowledge than low-interaction pairs even after the pre-test difference was controlled,  $F(1, 17)=6.10, p<.05$ . On the other hand, high-interaction pairs did not necessarily shared more stated knowledge than low-interaction pairs,  $F(1, 17)=.107, p>.10$ . Thus, the more interaction pairs engaged in, the more likely they were to construct knowledge that was inferred (i.e., knowledge that was not given in the text). Since the knowledge was never presented in the text, it was more likely that dyads constructed them together through collaborative interaction.

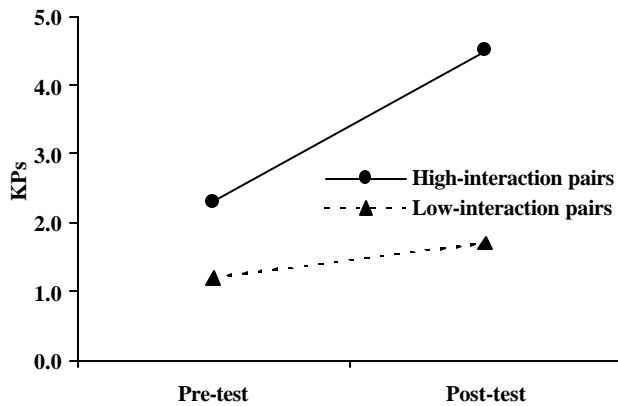


Figure 3: The increase in shared Stated KPs and shared Inferred KPs after collaborative learning in high-learning and low-learning pairs.

### Conclusions

In this study, we examined whether collaborative learning, a shared learning activity, leads to the construction of shared knowledge. Among the several potential representational outcomes of collaborative learning, one distinct possibility was that collaborating members of dyads (or groups) would construct common knowledge. To examine whether the common knowledge would really come from interaction rather than sharing the same environmental input, we examined whether students common knowledge when the knowledge cannot be obtained directly from the input. The results of this study showed that collaborating pairs shared more knowledge (correct and incorrect, stated and inferred) after collaboration. Since the incorrect knowledge and the correct but inferred knowledge was never presented in the text, it is more likely that they constructed it during collaboration. Above all, those who interacted more shared significantly more inferred knowledge than those who interacted less did. Even though each of these analyses produced a small effect and/or small amount of data, taken together, these findings indicate that participation in joint activity allows participants to construct a common knowledge.

There are several ways that the pairs went about constructing common knowledge in this study. In one scenario, the two pairs might have contributed to the construction of knowledge more or less equally, each generating part of inferences to complete the knowledge construction. In another scenario, one student might have made an inference, regardless of whether it is correct or incorrect, from the text by herself and tells her partner about it. At this point, the other partner had two choices: he or she could either accept it or reject it (Clark & Schaefer, 1989). It is only when the partner accepted the other's contribution that both of them get to possess the common knowledge. The partner who just heard the inference was more passive than the other person, but nonetheless participated in the construction process.

### Acknowledgments

This study was partially supported by the Spencer Foundation (Grant 199400132) to the second author.

### References

- Brown, A. L., & Palinscar, A. S. (1989). Guided, cooperative learning and individual knowledge acquisition. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser*. Hillsdale, NJ: Erlbaum.
- Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In J. J. Castellan (Eds.), *Current issues in individual and group decision making*. Hillsdale, NJ: Erlbaum.
- Chi, M. T. H. (In press). The dual processes of self-explaining: Generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology*. Mahwah, NJ: Erlbaum.
- Chi, M. T. H. de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439-477.
- Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, 13, 259-294.
- Jeong, H. (Under review). Being constructive and interactive during collaborative learning.
- Klimoski, R., & Mohammed, S. (1994). Team mental model: construct or metaphor. *Journal of Management*, 20(2), 403-437.
- Larson, J. R. J., & Christensen, C. (1993). Groups as problem-solving units: Toward a new meaning of social cognition. *British Journal of Social Psychology*, 32, 3-30.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.
- Levine, J. M., Resnick, L. B., & Higgins, E. T. (1991). Social foundations of cognition. *Annual Review of Psychology*, 44, 585-612.
- Orr, J. E. (1990). Sharing knowledge, celebrating identity: War stories and community memory among service technicians. In D. S. Middleton & D. Edwards (Eds.), *Collective remembering: Memory in society*. London: Sage.
- Resnick, L. B., Levine, J. M., & Teasley, S. D. (1991). *Perspectives on socially shared cognition*. Washington, DC: American Psychological Association.
- Rogoff, B. (1998). Cognition as a collaborative process. In W. Damon (Ed.). *Handbook of Child Psychology*. New York: Wiley.
- Roschelle, J. (1992). Learning by collaborating: Convergent conceptual change. *The Journal of the Learning Sciences*, 2(3), 235-276.
- Sherif, M. (1936). *The psychology of social norms*. New York: Harper & Row Publishers.
- Tudge, J., & Rogoff, B. (1989). Peer influences on cognitive development: Piagetian and Vygotskian perspectives. In M. Bornstein & J. Bruner (Eds.), *Interaction in human development*. New Jersey: Lawrence Erlbaum.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

# Learning from a Computer Workplace Simulation

Heisawn Jeong (heis@pitt.edu)  
Roger Taylor (rtaylor@pitt.edu)  
Micheline T. H. Chi (chi@pitt.edu)

Learning Research and Development Center; University of Pittsburgh  
3939 O'Hara Street, Pittsburgh, PA 15260 USA

## Abstract

Workplaces are rapidly changing, placing increased cognitive demands upon workers. The use of computer workplace simulations has been proposed to help students successfully make the transition from school to work. In this study, we examined what kinds of learning occurred when students used a computer workplace simulation called Court Square Community Bank. We hypothesized that three types of learning would occur: (1) students would gain knowledge about the banking business in which the simulation is situated and (2) students would also learn general business knowledge and problem solving/decision making skills that they could apply in other work contexts. Thirteen pairs of high school students used a workplace simulation. The results showed that students knew significantly more knowledge about the banking business. Students also adopted a new perspective to organize their knowledge and their problem solving activities became more coordinate. Taken together, the results of this study showed that computer workplace simulation can serve as a useful tool to prepare students to make a better school-to-work transition.

## Introduction

There is a growing concern that many of today's high-school graduates are ill-prepared for succeeding in today's demanding and rapidly changing workplaces. The world of work is experiencing a dramatic transition: jobs increasingly require complex thinking skills and adaptive performance. As a consequence, there have been many calls for school-to-work transition programs such as youth apprenticeship or technical preparation. Recently, Ferrari, Taylor, and VanLehn (1999) advocated the use of computer simulations as a way to facilitate the school to work transition. They argued that computer simulations of workplace environments can help familiarize students with a particular workplace and assist them in developing the analytical/problem-solving skills needed to successfully participate in the workplace, while allowing them to remain safely situated in the classroom. The goal of this paper is to assess what kinds of learning opportunities are afforded and how much learning actually occurs when students use such a computer simulation.

For our study, we selected a workplace simulation called Court Square Community Bank (CSCB), one of the simulations recommended by Ferrari et al. (1999). CSCB is an episode-based simulation. Students play the role of vice president, engaging in activities such as interacting with

bank customers, consulting the opinions of other bank employees, and making business decisions. Each of the 14 episodes poses different kinds of problems that the vice president of a small community bank must deal with such as approving mortgages or selecting the best candidate for a position (see Ferrari et al., 1999 and McQuaide, Leinhardt, & Stainton, 1999 for more details on the program).

In assessing learning from CSCB,<sup>1</sup> we were less concerned with evaluating the specific workplace simulation and were more interested in understanding the learning issues involved in workplace simulations in general. We hypothesized that two types of learning can occur when students interact with a computer workplace simulation. Students could acquire (1) knowledge about the banking business in which the simulation is situated, and (2) general business knowledge and problem solving skills that are applicable to a wide variety of workplaces. Below, we describe these in more detail and speculate on how learning such knowledge/skills might occur.

## 1. Knowledge about the Banking Business

One of the most notable features of the computer workplace simulation in comparison to other medium of instructions (e.g., reading an expository text or listening to a lecture) is its contextualized nature. In the case of CSCB, the specific business context was a small town community bank. The contextualization was done by using the problems that arises from the banking business (e.g., when to approve a mortgage) and implementing interactions with simulated characters who are primarily bank personnel or customers. This means that much of the information about banking business is embedded in the problem descriptions and the students' interactions with the characters of the simulation. In addition to this contextualization, declarative banking knowledge is presented in the form of on-line dictionary and procedural manual. In sum, the simulation provides extensive amount of information about banking business, either implicitly or explicitly. Although this banking knowledge is never the focus of the simulation (e.g., the program never asks stu-

---

<sup>1</sup> The effect of the simulation is likely to be different when used in schools compared to when it was used in the laboratory as in this study. For example, in one of the schools that used the CSCB as part of their curriculum, it was augmented with instructional and teacher supports (see McQuaide et al., 1999 for more details).



dents to supply the definition of various financial terms), it seems reasonable to expect that students would at least learn some amount of banking knowledge as a result of using the simulation.

## 2. General Business Knowledge and Skills

Although learning about banking is meaningful and could be helpful in other contexts, we hope that students would also learn general knowledge or skills that can be used in contexts other than banking. When students go through the simulation episodes, they need to process episode-specific information (e.g., salary of the mortgage applicants) as well as banking specific knowledge (e.g., interview loan applicant before approving the loan). Such knowledge, although useful in making banking-related decisions required in the episode, is largely irrelevant in other contexts.

Concrete contexts can help initial learning because they can be elaborated and help students appreciate the relevancy of new information in problem solving. Context can also be helpful to learning by facilitating the construction of a more accurate representation in a manner similar to how context helps to disambiguate word meaning. On the other hand, context can also present a problem for abstracting general principles or features (e.g., category structure). Overly contextualized learning tasks could potentially impede the abstraction of general principles (Bransford & Schwartz, 1999). Despite its benefits for learning, learning in context could also present a challenge to students in that the contextualized nature of the simulation may impede learning general knowledge or skills that can be useful in other work contexts.

It should be noted that this problem is not unique to computer workplace simulations. The same issue is present in on-the-job or apprenticeship training. There is an abundance of context/job-specific information in the actual work context. Although people need to pay attention to this information, often the ideal learning goal can be met only when people go beyond this information and understand more general business issues (Pearlman, 1997). For example, students need to learn that one needs to consider all the available options before making the final decision or that the role of the vice president may be complicated due to potential conflicts of interest or concerns about nepotism.

According to the analysis of Ferrari et al. (1999), approximately two thirds of the information presented in the two simulations that they examined in detail were specific to the particular type of the industry simulated (i.e., banking and software development). Only one-third of the information was general work knowledge that dealt with issues such as decision making, information management or interpersonal relations that occur across many jobs. In other words, unlike textbooks that often present this decontextualized knowledge in the form of abstract principles or concepts, simulations like CSCB embed this general knowledge in a specific context. The question then is what the kinds of general knowledge or skills students could learn from their experience with the simulation that could be used in other work or business contexts. In this study, we examined the following two candidates.

**Perspective Taking** In the simulation, students are asked to play the role of bank's vice president and interact with various characters related to the business (e.g., customers, managers, etc.). Given that prior to using the simulation, almost all of the students' interactions with business would have occurred while they were in the role of a consumer (i.e. purchasing items from stores), one would expect students to answer banking terms to be answered from a similar perspective. For example, a banking term such as *interest* can be defined from the perspective of a consumer or a business. When students approach the term from a customer's perspective, *interest* is more likely to be defined as an expense paid to banks. On the other hand, when students approach the same term from a business perspective, *interest* is more likely to be defined as a means for earning profit. The ability of students to take multiple perspectives would reflect a richer understanding of the terms and principles under discussion. Thus, being given the opportunity to take on new roles might affect how students frame their experiences, allowing them to think about the same issue in multiple ways.

**Decision Making/Problem Solving Skills** Most of the episodes in CSCB require some kind of decision to be made. In each episode, students are first presented with basic descriptions of the problem (e.g., the treasurer announces that the downtown branch has a shrinking profit margin). Students attend a meeting to hear what managers think may be the cause of the problem, read newspaper article about how the community is reacting to the possible closing of the branch, and evaluate various options to address the problem of the branch. The simulation provides a set of alternatives regarding the closing of the branch and asks the students to choose and justify their choice.

Although the decisions in the simulation are simplified by using multiple choice format, they are still complex in nature, resembling the kinds of decision making that might occur in real workplaces. In general, the problems given in the simulation episodes are different from the problems commonly dealt in the classroom (e.g., algebra problems). First, they require an understanding of a both specific context (banking business in this case) and general problem solving skills (e.g., goal state). Second, there is no single right answer as is the case in typical problems taught at schools. There often exist multiple equally viable options that can solve the problem, and even the best solution can be flawed in some way. Third, the problems are complex in that the goals and solution options of the problem are often unclear. The problems also have multiple, interacting causes.

Engaging in such decision making is a complex task. Successful problem solving requires students to understand several factors and their relationships (e.g., governmental regulations or why the branch is losing money). Like in real workplaces, problems in the simulation often do not have a single right answer. Additionally, students often need to evaluate each option based on their own criteria. For example, students need to evaluate the relative merits of advanced technology (e.g., ATM) versus personal attention to cus-

tomers (e.g., human tellers). Due to such characteristics of the problem presented in the simulation, we postulated that students would learn how to solve such ill-defined problems better after using the simulation. According to the notion of "Preparation for Future Learning" proposed by Bransford and Schwartz (1999), the benefits of previous experience often do not reveal themselves immediately. Instead, the benefit takes the form of helping to prepare students to learn new information. Thus, we examined not only how the overall quality of their decision improved, but also whether they had a better understanding of the problem solving process after using the simulation.

In this study, to get a detailed picture of students' learning, we asked high school students to do eight (out of fourteen) CSCB episodes in the lab. We constructed three assessment tools to test and elaborate our hypotheses about potential learning outcomes from the workplace simulation. They were: (1) Definition Task, (2) Question Answering Task, and (3) two transfer problems.

## Method

### Participants

Twenty-six high school students (23 from public and 3 from parochial schools) participated in this study. The students were either juniors or seniors from local urban high schools. During recruitment, students were asked to bring a friend of the same gender to participate in the study, which resulted in four male and nine female pairs. On average, they had known their partner for about 4 years, having engaged in academic and after-school activities together. As compensation for participation in the study, students were provided with a base pay of \$75 and up to \$25 as a bonus if they kept their appointments. With regard to their familiarity with computers, about half of the 26 students (54%) reported that they had computers in their home. Almost all students (96%) reported that they used computers 1-15 hours per week and had experience with word processing or e-mail. A majority of the students (58%) reported that they used other computer simulations or games.

### Materials

**CSCB Episodes** A representative subset of CSCB episodes (8 out of 14) was selected so that diverse topics would be covered (e.g., ethical as well as financial issues) with minimum overlap.

**Definition Task** The aim of this task was to assess the context-specific knowledge that students might learn about the banking business. This task consisted of 13 terms relevant to banking (e.g., collateral) that were covered in the eight selected episodes. Students were asked to talk about everything they knew about these terms.

**Question Answering Task** This task was to assess general business knowledge that students might have abstracted from their experience with the simulation. This task

consisted of 12 questions that were constructed based on the propositional content of the simulation episodes. These were general questions about how business operates (e.g., name 3 ways that a business can stay competitive) or about a vice president's role (e.g., name three kinds of activities/jobs that a CEO or a vice president has to do in a company). We expected that students would find this task difficult due to the lack of specific contexts. We thus included in the instructions that they could use examples of specific businesses to help to answer these questions.

**Problem Solving Task** This task was to assess broad changes in students' problem solving. Two transfer problems called Fresh Food and Giant Gallery were constructed based on two episodes of the simulation (Episode 9 and 10, respectively). They were identical to the problems presented in the simulation episodes in terms of the underlying problem/goal, constraints, and options, but differed from the simulation problems in two respects.

First, instead of banking, the transfer problems used the supermarket business (also familiar to high school students) as the context. Thus, although the surface features were different, the underlying structures of the transfer problems were identical to the problems presented in the simulation.

Second, transfer problems were less structured than the problems presented in the simulation episodes. Unlike the simulation episodes that provide a set of alternative choices, in the transfer problems students were asked to generate their own solutions. A set of seven open-ended probes about various aspects of students' reasoning were included. Each transfer problem consisted of two general phases: information interpretation and probe answering. Students were first presented with a set of documents about the problem and then responded to a set of five probes. They were then provided with an additional set of documents and they responded to two more probes. The seven probes were the following:

- Probe 1: Could you please state the store's problems in your own words?
- Probe 2: If you could request more information about this supermarket's problem, what information would you request and how would you get this information?
- Probe 3: As vice president of this supermarket, how (or where) would you get this information?
- Probe 4: What factors do you have to take into account to solve the supermarket's problems?
- Probe 5: As vice president, how would you go about implementing your options?
- Probe 6: What do you think is the best way to solve the problem and why?
- Probe 7: What kind of information did you use to make your decision?

### Procedures

The study was carried out in the laboratories located in the Learning Research and Development Center at the University of Pittsburgh. Students visited the laboratory over four sessions: (1) pre-test, (2) simulation session I, (3) simulation session II, and (4) post-test. On average, each session

was separated by approximately four days. All sessions were audio-taped. In addition, the two simulation sessions were video-taped to provide a context for the interaction between the students.

**Pre- and Post-test** During the pre-test and post-test, students were given the three tasks described earlier. Students first generated answers to the terms in the Definition Task. They then solved the two transfer problems, Giant Gallery and Fresh Food. Lastly, they completed the Question Answering Task. The order of the two transfer problems was counterbalanced across pairs and pre/post-test sessions. Throughout the sessions, students were asked to think aloud and to talk about everything that came to their mind. To familiarize them with think-aloud procedure, students were given a short think-aloud practice at the beginning of the problem-solving during the pre-test. The pre-test and post-test sessions were individually administered. The pre-test took about two hours, and the post-test took about one and a half hours on average.

**Simulation Sessions** The first three episodes in each simulation session (Episode 1, 2, & 5 in the simulation session I, and Episode 9, 10, & 11 in the simulation session II) were done collaboratively by the pair, and the last episode in each session (Episode 7 in the simulation session I, and Episode 12 in the simulation session II) was done individually. In the collaborative simulation session, the pairs were instructed to work as a team, in discussing how to handle the problems and to reach a consensus before making a decision. After the collaborative learning session, students were led to separate rooms and completed an episode alone while thinking aloud. Students took approximately 40 minutes per episode, a total of five hours on the simulation over both sessions (excluding time spent on breaks).

## Results

Please note that only a subset of results was reported in this paper for the two types of learning examined: (1) learning about the banking business and (2) learning about general knowledge or problem solving skills. We presented two sets of results for the first type of learning and four sets of results for the second type of learning.<sup>2</sup>

### 1. Knowledge about the Banking Business

**Increase in Correct Knowledge and Decrease in Incorrect Knowledge** A Knowledge Piece (KP) roughly corresponds to an idea (e.g., ATM costs banks less than tellers). A template was constructed by identifying individual Knowledge Pieces (KPs) relevant to the 13 terms in the Definition Task. The template consisted of 113 KPs captured from the information presented in the simulation in various formats (e.g., on-line dictionaries or reports given to

the vice president). The template represented the maximum possible knowledge that students *could* learn from the eight episodes of the simulation. Based on their answers to the Definition Task, students received one point for every unique KP that they stated (partial credit was given if their answer was vague or they expressed uncertainty). Students knew on average 24 KPs at the pre-test and 27.98 KPs at the post-test,  $t(24)=3.79$ ,  $p<.01$ . In addition, we examined the terms in which students either provide no answer or provided incorrect (or irrelevant) answers (e.g. *principal* was a “person in school”) and found that such answers decreased significantly from 2.28 (18%) at the pre-test to .88 (7%) at the post-test,  $t(24)=4.09$ ,  $p<.001$ .

**Schema about Banking Business** A subset of questions in the Question Answering Task asked students to explain general business operations. The answers to these questions were analyzed in terms of the type of business schema used (e.g. manufacturing, retail, banking, etc.). For example, to the question “Please name three kinds of expenses that a business has,” one student answered: (1) rent, (2) expense to make the product, and (3) expense to get the products out to the public (e.g., shipping or mailing expenses). In this case, the first answer was an expense that was applicable to the banking business, whereas the second and the third answers were not. On average, students’ answers applicable to the banking business significantly increased after using the simulation (from 62% to 77%),  $t(25)=5.02$ ,  $p<.01$ , suggesting that students learned some rudimentary schema about the banking business.

### 2. General Business Knowledge and Problem Solving Skills

**From Customer to Business Perspectives** To assess the change in the perspectives, we examined the perspectives that students used in defining the terms in the Definition Task. We coded their answers to each term in the following two perspectives: (1) customer perspective, (2) business perspective. The results showed that there was a significant decrease in answers with customer perspective (3.44 to 2.44),  $t(24)=2.45$ ,  $p<.05$ , and a significant increase in answers with a business perspective (1.04 to 2.44),  $t(24)=3.03$ ,  $p<.005$ . Thus many students now demonstrated the ability to think about banking terms from additional framework (i.e. a business perspective in addition to their initial consumer perspective).

**Improvement in the quality of the students’ final decision** As mentioned before, due to the nature of the problem, it was often difficult to determine optimal solutions. What constituted the “best” solution was highly dependent on one’s beliefs and priorities (e.g., the importance of technology versus that of renovation in business). Nonetheless, we examined whether there were any improvements in the quality of students’ final decisions that they chose in the two transfer problems. This was done based on their response to Probe 6 (“What do you think is the best way to solve the problem and why?”). Based on the analysis of the

<sup>2</sup> Due to recording errors, the following data was lost: student 9B’s pre-test answers to the Definition Task and in the Giant Gallery problem and student 8B’s post-test answers to Question 3 in the Giant Gallery problem.

simulation episodes, we selected five constraints important to the decision-making and examined how many constraints each of their final decision satisfy. In the Giant Gallery problem, students' final decision met 3.20 constraints (out of 5) at the pre-test and 3.96 constraints at the post-test,  $t(24)=1.93$ ,  $p<.05$ . In the Fresh Food problem, there were no significant changes in the numbers of constraints (3.62 to 3.54).<sup>3</sup>

**Improved Understanding of the Problem Solving Component** We examined students' understanding of problem solving components, specifically, their understanding of solution option. At pre-tests, when asked to generate options to solve the store's problem in Probe 3, students were more likely to generate options that did not really *solve* the problems that the supermarket was facing. For example, students answered "look [at] every aspect of the store" or "get ideas from other stores." These answers may be steps to arrive at the final solution, but were not options that could *solve* the specific problems that the supermarkets had. Generation of such *non-options* decreased significantly both in the Fresh Food problem (1.31 to .52),  $t(25)=2.33$ ,  $p<.05$ , and in the Giant Gallery problem (.78 to .32),  $t(24)=1.83$ ,  $p<.05$ , suggesting that students understood better what a solution option was after using the simulation.

**Integrated Problem Solving** In addition, students' problem solving activities became more integrated and coherent. First, students became better at gathering the information. In Probe 2, students were asked to request information they need to understand and solve the problem and then specify how to get it. At pre-tests, they tended to specify the steps to gather information in general rather than the information they requested. For example, in the Fresh Food problem, one student requested three pieces of information: (a) how old the current machines at the checkout counter were, (b) whether customers use the machines properly, and (c) why the machines at the checkout counter was not connected to the main computer system. Then, she specified that she would get that information by (a) asking the company that made the checkout machines in the store about how to fix it, (b) talking to the competitor whether they had similar problems, and (c) talking to the bank about the compatibility of the card and the machine. In this example, none of her information gathering methods were about the information she requested, although they were valid ways of getting information. At post-tests, students were more likely to specify information gathering methods to get the information they requested. Although the increase was only significant in the Fresh Food problem (53% to 71%),  $t(25)=2.45$ ,  $p<.05$ , the trend was also present in the Giant Gallery problem (58% to 69%).

A similar finding was obtained when students were justifying their final decisions in Probe 6. We coded whether students actually considered the constraints that they listed in

their response to Probe 4 ("What factors do you have to take into account to solve the supermarket's problems?"). After students did the simulation, they were more likely to consider the constraints that they listed in Probe 4. In the Fresh Food problem, students used .81 reasons at the pre-test (out of 1.58 reasons they used to justify their final decision) that they had named and 1.17 reasons (out of 1.96) at the post-test,  $t(25)=1.74$ ,  $p<.05$ . In the Giant Gallery problem, students used .84 reasons at the pre-test (out of 2.04) that they had named and 1.24 reasons (out of 2.60) at the post-test,  $t(24)=1.79$ ,  $p<.05$ . Thus, it seemed that students' problem solving activities became more connected or coherent in that they generated ways to get the information they requested and used more of the constraints that they initially thought were important in solving the problems.

## Discussion

In this study, we attempted to identify and assess the learning outcomes of computer workplace simulations. We first speculated that students would learn about the work context of the simulation. The CSCB simulation uses banking as a context, but other simulations have used other businesses (e.g., hotel management or software development). We also speculate that students would learn general knowledge about business and how to solve complex real-world like problems.

First, the results showed that the simulation helped students to learn about the banking business. They knew more knowledge about banking, which was accompanied by a corresponding decrease of incorrect knowledge. Students also acquired a schema about banking business. Such results are interesting, considering the fact that (a) students were never asked to learn about banking explicitly and (b) they were not likely to have accessed all the relevant information about the banking business even though they were provided in the simulation.

Second, the results also showed that the simulation helped students to learn general business knowledge and problem solving skills. Students learned to take a business perspective, one that they would not have likely gained from their everyday experiences. As a result, students' answers to the banking terms became more organized from a business perspective rather than from a customer perspective. In addition, students understanding of the problem improved and became more coherent, which seems to be one of the reason why the quality of their final answer improved after the simulation. Considering the fact that the simulation never teaches these knowledge and skills didactically and embeds them in contexts, it is encouraging to discover that students can not only learn about the banking business in which the simulation is contextualized, but also some general knowledge and problem solving skills that they can use in other business and work contexts as well. These results of this study are consistent with the findings obtained with other instructional mediums that use contexts or cases, although they did not deal with workplace issues (e.g., mathematical problem solving as in Jasper Series; see Barron, Zech, Schwartz, Bransford, Goldman, Pelligrino, Morris, Garrison, Kantor, 1995). Taken together, it seems that workplace computer

---

<sup>3</sup> This seems to be due to the fact that students pre-test decisions were highly similar to the decisions that were reinforced in the simulation in the Fresh Food problem.

simulation could be useful in preparing students for the future workplaces.

### **Acknowledgments**

This study was funded by the grant to Michelene T. H. Chi, Gaea Leinhardt, and Kurt VanLehn from the A. W. Mellon and Russell Sage Foundation (495006544). The authors acknowledge the help of Cindy Hmelo and Alex Vincent who participated in the initial phase of the study.

### **References**

- Barron, B., Vye, N., Zech, L., Schwartz, D., Bransford, J., Goldman, S., Pellegrino, J., Morris, J., Garrison, S., & Kantor, R. (1995). Creating contexts for community-based problem solving: The Jasper Challenge Series. In C. N. Hedley, P. Antonacci, & M. Rabionwitz (Eds.), *Thinking and literacy: The mind at work* (pp. 47-71). Hillsdale, NJ: Erlbaum.
- Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* (Vol. 24). Washington, DC: American Educational Research Association.
- Ferrari, M., Taylor, R., & VanLehn, K. (1999). Adapting work simulations for schools. *Journal of Educational Computing Research*, 21(1), 25-53.
- McQuaide, J., Leinhardt, G., & Stainton, C. (1999). Ethical reasoning: Real and simulated. *Journal of Educational Computing Research*, 21(4), 425-466.
- Pearlman, K. (1997). Twenty-first century measurer for twenty-first century work. In A. Lesgold, M. Feuer, J., & A. Black, M. (Eds.), *Transitions in work and learning implications for assessment*. Washington, D. C.: National Academy Press.

# A Dynamical Model of Insightful Memory Retrieval

Koji JIMURA, Hisaaki KOMAZAKI

Takashi MATSUOKA, Masanori NAKAGAWA, Takashi KUSUMI

{ jimura | komaspee | matsuoka | nakagawa | kusumi }@tp.titech.ac.jp

Department of Human System Science, Tokyo Institute of Technology

2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, JAPAN

## Abstract

The authors propose a dynamical model of memory retrieval that explains how people break an impasse or a memory block spontaneously without an external stimulus. We describe the process as “insightful memory retrieval”. First, an experiment was conducted in which 15 participants retrieved eight Chinese characters from memory space using figural pattern cues. The results indicated that the retrieval process was divided into three phases: (1) direct retrieval, (2) indirect retrieval, and (3) an impasse and insightful retrieval. Second, a dynamical model named DIMeC was developed from the results. The direct and indirect phases depend on constraint relaxation, and the insightful retrieval phase is simulated using a chaotic neural network. Third, the DIMeC model was implemented on a computer. The results of the simulation indicate that the model reflects the typical dynamic retrieval process of the participants.

## Introduction

Insight and memory block resolution are connected (e.g., Weisberg & Alba 1981; Bowers, Balthazard, & Parke 1990; Yaniv, Mayer, & Davidson 1995). Insight is characterized by spontaneity, suddenness, unexpectedness, and satisfaction (Seifelt et al. 1995), however, research into memory blocks, including the tip-of-the-tongue phenomenon, output interference, fixation, and priming effects has not shown a “spontaneous” mechanism (e.g., Smith 1995; Smith & Tindell 1997; Yaniv et al. 1995). These researchers have explained a memory block and its resolution mechanism by setting an external stimulus, such as priming. We assume that “insightful memory retrieval” expresses a spontaneous mechanism, which breaks a memory block or an impasse without an external stimulus.

An artificial neural network is one of the cognitive models that can explain the human memory system. Many models use a minimized energy function to retrieve a memory from an initial pattern, and cannot search for additional memories. Inevitably, they cannot express dynamical retrieval processes, such as retrieving one memory from another. However, Nara, Davis, & Totsuji (1993) developed a model that could travel among memories, using chaotic dynamics by controlling the number of connections among units. Tani (1996) developed the “chaotic steepest descent” model, which

could also travel among memories. Tani employed a non-linear resistance to control the chaotic transition. It is interesting to apply such dynamics to the memory retrieval process of actual psychological phenomena.

Therefore, our purposes were (1) to explore how people reach an impasse and break it by insight in the memory retrieval process, (2) to develop a dynamical model using a chaotic neural network from the results, and (3) to examine the model using computer simulations.

## Experiment

### Method

**Participants** The participants were 15 Japanese graduate students at the Tokyo Institute of Technology.

**Material** The problem they were given was to find all the Chinese characters (Kanji) that can be made by adding one straight line to “*I*” without rotation, where “*I*” denotes the initial character in Figure 1. Eight Chinese characters can be constructed from “*I*”. Figure 1 shows the initial character (*I*) and the target characters ( $C_1 \dots C_8$ ) and Table 1 shows the operations that must be made on the initial character to retrieve the targets. All Japanese people should have learned all these characters and the initial character in school between ages seven and fifteen years. We call the problem the ALIC (Add a Straight Line to the Initial Character) task.

**Procedure** Each participant was tested individually, and participants’ actions and speech were recorded using a VCR. The participants were told to write down their answers on a sheet. During the session, they were urged to speak their thoughts aloud and to write, regardless of incorrect answers. When the participants indicated that they could not think of any more targets, they were told the number of remaining targets. Each session lasted until all eight targets were retrieved.

### Results

**Retrieval Times and Targets** All the participants retrieved all the targets within 30 minutes. Figure 2 shows the retrieval patterns of the 15 participants. Each line represents the retrieval pattern of one participant. The horizontal axis indicates the cumulative number of retrieved targets and the vertical axis indicates the cumulative time. These patterns indicate that the retrieval processes were “insightful”. The participants routinely retrieved five to seven targets regularly in 0 to 120 seconds, however, they could not retrieve the remaining targets for a long time; they reached an impasse (a memory

Table 2: Retrieval times for targets

Target	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$C_7$	$C_8$
Mean*	44.1	26.5	66.0	71.0	283.0	294.5	291.9	141.5
S.D.	61.5	29.3	64.3	55.1	483.4	487.0	455.1	189.2

Note:  $n = 15$ , \* seconds

日 田 目 旧 且 甲 由 申 白  
 $I$   $C_1$   $C_2$   $C_3$   $C_4$   $C_5$   $C_6$   $C_7$   $C_8$

Figure 1: Initial ( $I$ ) and target ( $C_1...C_8$ ) characters

Table 1: Targets and operations

Target	Operation on the initial character
$C_1$	Add a vertical line inside $I$
$C_2$	Add a horizontal line inside $I$
$C_3$	Add a vertical line on the left outside $I$
$C_4$	Add a horizontal line under $I$
$C_5$	Add a vertical line that goes through $I$ and protrudes from the bottom
$C_6$	Add a vertical line that goes through $I$ and protrudes from the top
$C_7$	Add a vertical line that goes through $I$ and protrudes from both the top and the bottom
$C_8$	Add a slanted line on the top of $I$

block developed). Then, they broke the impasse, and subsequently retrieved the remaining targets relatively quickly.

Table 2 shows the mean retrieval time and the standard deviation for the eight targets by the 15 participants. There was a relationship between the figural features of the targets and retrieval time. Targets  $C_1$  and  $C_2$ , which are made by adding a line inside the initial character, were retrieved quickly. Targets  $C_3$  and  $C_4$ , which are made by adding a line outside the initial character, were retrieved next. It took longer to retrieve targets  $C_5$ ,  $C_6$ , and  $C_7$ , which have same added feature, a protruding line. Figure 3 shows the cluster tree for the eight targets obtained by cluster analysis using the flexible  $\beta$  method, where  $\beta = 0.25$ . Each target was clustered by retrieval time, so that the distance between targets indicates the retrieval time interval. It is clear that  $C_3$  and  $C_4$  were retrieved within a short interval, and the same applies to  $C_1$  and  $C_2$ , and  $C_5$  and  $C_6$ . These results show that the participants retrieved the targets using figural cues.

**Protocol analysis** The protocol data revealed that the participants repeated the following processes: (1) drawing a straight line, (2) confirming whether the character was a target. Repeating the processes, they reached the following three phases. *Phase1* (direct retrieval): 14 out of 15 participants reached this phase, and retrieved some targets without failure. *Phase2* (indirect retrieval): All the participants reached this phase, and retrieved some targets with retrieval failures, repeated retrievals, or by making writing motions with the hand without producing a visible trace. *Phase3* (impasses and insightful retrieval): 13 out of 15 participants reached this phase, and were unable to retrieve the remaining targets for a long period (over 50 seconds). In this period, they drew curved lines, added two strokes and found a Chinese char-

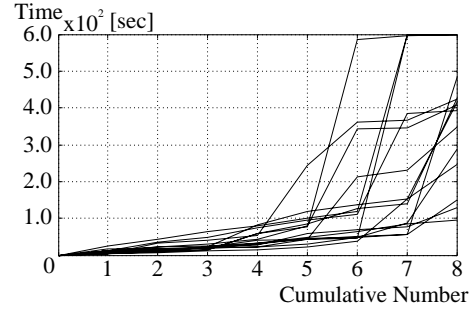
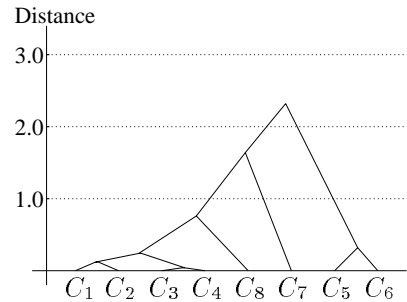
Figure 2: Retrieval patterns ( $n = 15$ )

Figure 3: Cluster tree of the targets

acter formed by adding two lines to the initial character, or did nothing. They failed repeatedly and their mental state fluctuated. However, they then suddenly retrieved one of the remaining targets. Once they found the figural pattern of the retrieved target, the rest were retrieved.

Table 3 shows the frequency of retrieved targets and time spent in each phase for the 15 participants.  $C_1$  and  $C_2$  (add inside) were retrieved mainly in *Phase1*.  $C_3$  and  $C_4$  (add outside) were retrieved mainly in *Phase2*.  $C_4$ ,  $C_5$ , and  $C_6$  (add protruding line) were retrieved mainly in *Phase3*.  $C_8$  (add a slanted line) was retrieved in all phases. The duration time indicates that it took much longer to complete *Phase3* than *Phase1* or *Phase2*, so that the subjects smoothly retrieved some targets in *Phase1* and *Phase2*, with some failures, but were deadlocked for a long time in *Phase3*. This indicates that retrieving targets by adding a line inside “ $I$ ” was easy, adding a line outside was relatively easy, and adding a protruding line was difficult.

## Discussion

**Chinese Characters for Japanese People** Japanese people store Chinese characters as not only letters but also words, and they can understand their meanings from their shapes, and it is natural and routine for them. We regard the ALIC (Add a Straight Line to the Initial Character) task as similar for Japanese people to the Word Fragment Completion task (e.g., Smith & Tindell 1997) for English-speaking people in terms of making up deficiency to retrieve a word.

Table 3: Frequency of retrieved target and duration time in each phase

	Drawing Operation										Duration time*		
	Inside line			Outside line			Protruding line			Slanted line	Mean	S.D.	
	$C_1$	$C_2$	Subtotal	$C_3$	$C_4$	Subtotal	$C_5$	$C_6$	$C_7$	Subtotal			$C_8$
<i>Phase1</i>	9	12	11	2	4	6	4	4	5	13	5	23.2	15.2
<i>Phase2</i>	5	3	8	12	10	22	5	4	4	13	5	57.4	38.1
<i>Phase3</i>	1	0	1	1	1	2	6	7	6	19	5	540.3	547.6

Note: n = 15, \*seconds

**Retrieval with Constraint Relaxation** It has been shown that there are constraints on insight problems and that insight arises when the constraints are relaxed (e.g., Hiraki & Suzuki 1998; Knoblich et al. 1999). In the ALIC task, there seemed to be stronger constraints to add lines inside and outside the initial character, and a weak constraint to add a protruding line. Their strength seemed to change with repeated failure as the retrieval process progressed. When the constraints were relaxed, the participants could retrieve the remaining targets by insight.

## Model

### Hypothesis of the Model

**Retrieval Process and Constraint Relaxation** Hiraki & Suzuki (1998) maintained that the problem can be expressed by three components (object, relation, and goal), where each component has a constraint, and that insight problems are solved by cooperation among the components and with relaxing their constraints, which triggers a representation change. We assume that memory retrieval processes in the direct and indirect phases depend on the relaxation of constraints (object and relation), evaluated by the goal constraint.

**Insightful Retrieval with Chaotic Process** Finke & Bettle (1996) maintained that insight is spontaneous, and occurs at levels of processing that lie below conscious awareness and control of the underlying process, which characterizes chaotic thinking, and that a chaotic process can often be employed when normal pathways are blocked. Then, we assume that chaotic dynamics explain the insight process.

**Conscious and Unconscious Layer** Since Finke & Bettle (1996) maintained that a chaotic process is employed without awareness when logical strategies fail, we developed a model including conscious and unconscious layers. The former characterizes direct and indirect retrieval and the latter characterizes an impasse and insightful retrieval. In the conscious layer, a strategic process generates an image by adding one straight line to the initial character “I”, and then the image is sent to the unconscious layer. In the unconscious layer, the memory space associates the image and the retrieved result is sent to the conscious layer, where it is evaluated. Strategic procedures fail with repeated retrieval failure, so a chaotic retrieval process is then employed in the unconscious layer. Consequently, the state of the memory space repeats the chaotic transition and retrieves the remaining targets by insight.

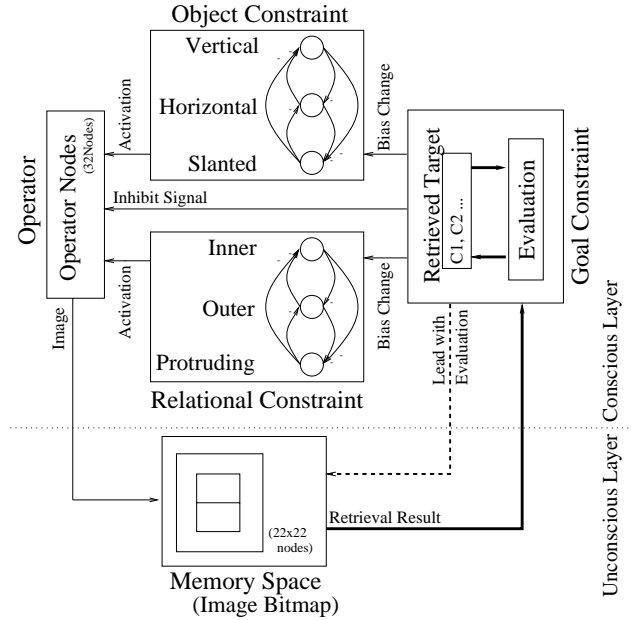


Figure 4: The DIMeC model

### Retrieval Process

Figure 4 illustrates the model. We named the model “DIMeC” (Dynamical model of Insightful Memory retrieval with Constraint relaxation). Narrow solid lines are enabled only when a strategic process is employed, the dashed line is enabled only when a chaotic process is employed, and bold solid lines are enabled all the time.

DIMeC repeats the following four steps: it generates an operator to draw a straight line; it generates an image by adding the line to the initial character; it associates the image with memory space and retrieves the result; and it evaluates the result. This process retrieves some targets and this period includes *Phase1* and *Phase2*.

After retrieving some targets, an impasse arises, with constraint relaxation because of repeated retrieval failure. Then, the chaotic process is employed and the remaining target is retrieved by insight. This period is *Phase3*.

### Definition of Memory Space

The image generated by the operator is associated with the memory space to be evaluated. The memory space is defined by learning Chinese characters with the Hebb rule, and representing their shape on a  $22 \times 22$  pixel image bitmap, where one unit represents one pixel. We employ a Hopfield network (Hopfield & Tank 1985) where the internal state of the  $i^{\text{th}}$  unit is defined as  $x_i$ ,



$$\frac{dx_i}{dt} = -\frac{x_i}{\tau} + \epsilon \left( \sum_k w_k u_i + \theta_i \right) \quad (1)$$

where  $\tau > 0, \epsilon > 0$  and  $\theta_i$  is the threshold of  $i^{\text{th}}$  unit. The output of the  $i^{\text{th}}$  unit is defined as  $u_i$ ,

$$u_i = \frac{2}{1 + e^{-\alpha x_i}} - 1 \quad (2)$$

where  $\alpha > 0$ . The energy function of the memory space is defined as  $E$ ,

$$E = - \sum w_{ij} u_i u_j \quad (3)$$

where  $w_{ij}$  is the connective weight between the  $i^{\text{th}}$  and  $j^{\text{th}}$  units. The weight is based on the one-year frequency data for Chinese characters in the daily *Asahi Shimbun* (Nozaki et al. 1996). Only the initial character and the eight targets are trained.

### Retrieval with Constraint Relaxation

As mentioned above, the retrieval process in the direct and indirect phases depends on constraint relaxation (object and relation). In the ALIC task, the object constraint consists of three elements (horizontal, vertical, and slanted) and the relational constraint consists of three elements (inside, outside, and protruding). Table 4 shows the relationship between constraints and targets. The goal constraint is not relaxed, but evaluates the result of retrieval and acts on the object and relational constraints.

**Relational and Object Constraints** A Hopfield network (Hopfield & Tank 1985) is employed to represent object and relational constraints, which are called object and relational constraint networks, respectively. One unit is introduced for each constraint element, and each inhibits the others within each network. The strength of each constraint is represented by its threshold. The initial state of the unit is set to zero, and the network repeats the transition, minimizing the energy function. After it stabilizes, an operator corresponding to the output of the constraint networks generates an image by adding a straight line to the initial character. The operator consists of 32 units, previously defined with the protocol data of the experiment. The output of operator  $\mathbf{q}$  is calculated by the transformation matrix  $\mathbf{W}$ ,

$$\mathbf{q} = \mathbf{W} \begin{bmatrix} \mathbf{Z} \\ \mathbf{R} \end{bmatrix} \quad (4)$$

where  $\mathbf{Z}$  is the output of the object constraint network and  $\mathbf{R}$  is the output of the relational constraint network. Then, the memory space associates the image and the result is retrieved.

**Goal Constraint** As mentioned above, the goal constraint evaluates the retrieval result and acts on object and relational constraints. There are five functions

(a) When a target is retrieved, the goal constraint sends an inhibitory signal  $\mathbf{G}$  to the units of the operator corresponding to the previous operation,

Table 4: Targets and constraints

Target	Constraint					
	Object			Relation		
	ver.	hor.	sla.	inn.	out.	pro.
$C_1$	+	-	-	+	-	-
$C_2$	-	+	-	+	-	-
$C_3$	+	-	-	-	+	-
$C_4$	-	+	-	-	+	-
$C_5$	+	-	-	+	-	+
$C_6$	+	-	-	+	-	+
$C_7$	+	-	-	+	-	+
$C_8$	-	-	+	-	-	+

Note: + indicates positive and - indicates negative

$$\mathbf{G} = -\mathbf{TA} \quad (5)$$

where  $\mathbf{A}$  is the retrieved target vector and  $\mathbf{T}$  is the transformation matrix from the target to the inhibitory signal. Consequently, in the next operation, a different operator with the same constraints generates an image.

(b) In the case of repeated retrieval of a target, the constraints corresponding to the previous operation are relaxed by,

$$\frac{\partial \Theta}{\partial t} = -\zeta_r \mathbf{s} \quad (6)$$

where  $\zeta_r > 0$ ,  $\Theta$  is the strength of the constraints (i.e., threshold vector of the object and relational constraint networks), and  $\mathbf{s}$  represents the relaxed unit.

(c) In the case of retrieval failure (i.e., finding no Chinese characters), the constraints corresponding to the previous operation are relaxed by using equation (6) and replacing  $\zeta_r$  with  $\zeta_f$  where  $\zeta_f > 0$ .

(d) In the case of an impasse, the state transition in the memory space is led by an evaluation function. Details are presented in the following section.

(e) In the case of an insightful retrieval, constraints corresponding to the figural pattern of the target are strengthened,

$$\frac{\partial \Theta}{\partial t} = \zeta_s \mathbf{VA} \quad (7)$$

where  $\zeta_s > 0$  and  $\mathbf{V}$  is the transformation matrix from the target vector to the figural pattern vector.

### Impasses and Insightful Retrieval

Impasses and insightful retrieval processes arising in the memory space are simulated using the ‘‘chaotic steepest descent’’ (CSD) model (Tani 1996) led by an evaluation function.

**Chaotic Transition** Tani (1996) developed the CSD model with a neural network employing a nonlinear resistant  $f_i$  for the  $i^{\text{th}}$  unit.

$$m\ddot{x}_i + f_i(\dot{x}_i, t) = -\epsilon \frac{\partial E}{\partial u_i} \quad (8)$$

$$f_i(\dot{x}_i, t) = (d_0 \sin \omega t + d_1)\dot{x}_i + d_2 \dot{x}_i^2 \text{sgn}(\dot{x}_i)$$

where  $m > 0, \epsilon > 0, d_0 > 0, d_1 > 0, d_2 > 0, \omega > 0$ ,  $u_i$  is the output of the  $i^{\text{th}}$  unit,  $x_i$  is the internal state of the  $i^{\text{th}}$  unit,  $\ddot{x}$  is the acceleration of  $x$ , and  $\dot{x}$  is the velocity of  $x$ . With this model, Tani showed that the state of the network travels from one energy basin to another

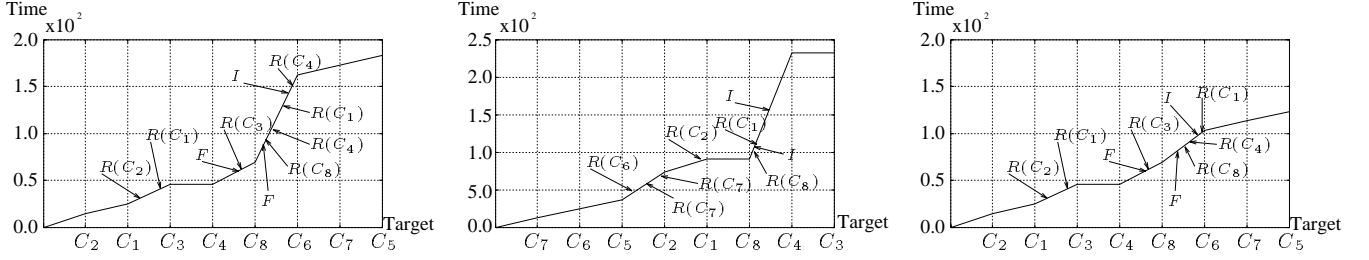


Figure 5: Retrieval patterns simulation 1 (left); simulation 2 (center); simulation 3 (right)

with chaotic dynamics as the resistance characteristics change from positive to negative, and the transition pattern corresponds to the cluster tree of the memory pattern. We employ the CSD model to simulate traveling among memories in an insightful retrieval process.

**Transition by Evaluation Function** Smith (1995) maintained that metacognitive monitoring towards a goal was often predictive of impending success. We have developed a metacognitive evaluation mechanism, similar to that which Nakagawa (1987) used to explain avoidance behavior by maximizing the evaluation of a psychological measure,

$$E_v = \sum_k \gamma_k \log \|\mathbf{u}_k - \mathbf{u}\|^2 \quad (9)$$

where  $\gamma_k > 0$ ,  $\mathbf{u}_k$  is an already retrieved target vector, and  $\mathbf{u}$  is the internal state vector in memory space. The value of the evaluation gets larger as the state in the memory space gets further from targets already retrieved.

As a result, we obtain the dynamics of the insightful retrieval, which can go from one memory to others as maximizing the evaluation,

$$\begin{aligned} m\ddot{x}_i + g_i(\dot{x}_i, x_i, t) &= -\epsilon \frac{\partial E}{\partial u_i} \\ g_i(\dot{x}_i, x_i, t) &= f_i(\dot{x}_i, t) + \beta \frac{\partial E_v}{\partial x_i} \end{aligned} \quad (10)$$

where  $\beta > 0$ .

## Simulation

### Method

The threshold vectors that represent the object and relational constraints at  $t$  are expressed as:

$$\Theta(t) = [\theta_v \ \theta_h \ \theta_s \ \theta_i \ \theta_o \ \theta_p]^T \quad (11)$$

where  $\theta\{v|h|s|i|o|p\}$  represents the strength of vertical, horizontal, slanted, inner, outer, or protruding constraints, and the notation “ $T$ ” indicates transposed.

**Simulation 1** The default constraint  $\Theta(0)$  was set as:

$$\Theta(0) = [2.0 \ 2.0 \ 1.0 \ 1.5 \ 1.0 \ 0.5]^T \quad (12)$$

In this condition, the vertical and horizontal constraints were stronger than the slanted constraint; the inner constraint was stronger than the outer constraint; and the outer constraint was stronger than the protruding constraint. The other parameters were as follows:  $\tau = 1.0$ ,

$\epsilon = 1.0 \times 10^{-3}$ ,  $\alpha = 3.0$ ,  $\zeta_f = 0.2$ ,  $\zeta_r = 0.8$ ,  $\zeta_s = 1.0$ ,  $m = 1.0$ ,  $\omega = \pi/20$ ,  $d_0 = 4.0$ ,  $d_1 = -4.0$ ,  $d_2 = 4.95 \times 10^1$ ,  $\beta = 1.0 \times 10^{-3}$ ,  $\gamma = 1.2 \times 10^1$ .

**Simulation 2** The default constraint  $\Theta(0)$  was set as:

$$\Theta(0) = [2.0 \ 0.5 \ 0.5 \ 1.5 \ 0.0 \ 3.0]^T \quad (13)$$

In this condition, the vertical constraint was stronger than the horizontal and slanted constraints; the protruding constraint was stronger than the inner constraint; and the inner constraint was stronger than the outer constraint. The other parameters were the same as in simulation 1.

**Simulation 3**  $d_2$  was set to 4.0, and the other conditions were the same as in simulation 1.

## Results and Discussion

Figure 5 shows retrieval patterns of simulations 1, 2, and 3. The horizontal axis indicates the retrieved target, the vertical axis indicates cumulative time,  $R(C_n)$  ( $n = 1, 2 \dots 8$ ) denotes the repeated retrieval of  $C_n$  ( $n = 1, 2 \dots 8$ ),  $F$  denotes retrieval failure, and  $I$  denotes the initial character. These patterns are similar to Figure 2. The retrieval patterns of the simulation reflected the typical retrieval patterns of the participants, because (1) each result could be divided into three phases, (2) targets with a figural pattern corresponding to strong constraints were retrieved first and weak ones were retrieved later, (3) targets with the same figural patterns were retrieved within a short interval, and (4) the retrieval process reached an impasse, and broke it insightfully by chaotic transition.

**Simulation 1**  $C_2$  (add inside) was retrieved at  $t = 14.3$  and  $C_1$  (add inside) was retrieved at  $t = 24.7$  because of the strong inner constraint. This period represents *Phase1* (direct retrieval).

$C_2$  was retrieved again at  $t = 35.2$  and  $C_1$  was retrieved again at  $t = 45.6$ , so the vertical, horizontal, and inner constraint were relaxed (i.e., the outer constraint became relatively stronger). As a result,  $C_3$  (add outside) was retrieved at  $t = 45.7$  and  $C_4$  (add outside) was retrieved at  $t = 45.8$ . Retrieval failed at  $t = 57.7$  and  $C_4$  was retrieved again at  $t = 57.8$ , so the horizontal, vertical, and outer constraints were relaxed. Since the protruding and slanted constraints became relatively stronger,  $C_8$  (add slanted) was retrieved at  $t = 68.8$ . This period represents *Phase2* (indirect retrieval with

repeated retrieval of some targets and retrieval failures).

After a retrieval failure at  $t = 78.9$ ,  $C_8$  was retrieved again at  $t = 79.5$ , so the slanted and protruding constraints were relaxed. Then, the strategic process failed, and the chaotic process was employed. The state in the memory space traversed  $C_4$ ,  $I$ ,  $C_1$ , and  $C_4$ . Traveling among memories,  $C_6$  (add protruding) was retrieved at  $t = 162.3$ . Having retrieved  $C_6$ , the protruding constraint became stronger. Therefore, the chaotic transition stopped and strategic process was employed again. Consequently,  $C_5$  (add protruding) was retrieved at  $t = 172.7$  and  $C_7$  (add protruding) was retrieved at  $t = 182.8$  because of the stronger protruding constraint. Retrieving all targets, the retrieval process ended. This period represents *Phase3* (impasses and insightful retrieval).

**Simulation 2**  $C_7$ ,  $C_6$ , and  $C_5$  (add protruding) were retrieved first because of the strong protruding constraint, then  $C_2$  and  $C_1$  (add inside) were retrieved. Because the outside constraint was weak,  $C_4$  and  $C_3$  were not retrieved by the strategic process in *Phase1* or *Phase2*, but were retrieved by the chaotic process in *Phase3*.

**Simulation 3** In *Phase1* and *Phase2*, the retrieval process was the same as in simulation 1, but  $C_7$ ,  $C_6$ , and  $C_5$  were retrieved earlier than in simulation 1 by the chaotic process in *Phase3*, because the retrieval process traveled among memories more often. This was caused by the smaller value of the nonlinear resistant coefficient  $d_2$ , which corresponds to the result of Tani (1996).

## General Discussion

Research into memory blocks has shown its process by setting an external stimulus (e.g., Smith 1995; Smith & Tindell 1997; Yaniv et al. 1995). We examined the dynamic spontaneous process of its resolution, and showed that the retrieval processes are similar to insight with constraint relaxation. We think that insight and memory block resolution can be explained by the same mechanism, even though representational change or information retrieval can explain the process of insight. The DIMeC (Dynamical model of Insightful Memory retrieval with Constraint relaxation) was developed, depending on constraint relaxation and a chaotic neural network. Therefore, we consider that the DIMeC architecture can be applied to the dynamic insight process.

Human memory has been explained using a neural network model based on its parallelism and distributive representation. Too much interest in this has prevented the development of a dynamical model of memory retrieval, although the human cognition process essentially consists of both parallel and sequential processes. The DIMeC model is a hybrid system that has both parallel and sequential mechanisms, enabling it to explain dynamical processes. It also refers to the conscious and un-

conscious, and can explain the priming effect, TOT phenomenon, output interference, and other memory block phenomena.

## References

- Bowers, K. S., Regehr, G., Balthazard, C., & Parker, K. (1990). Intuition in the Context of Discovery. *Cognitive Psychology*, Vol.22, 72-110.
- Finke, R. A. & Bettle, J. (1996). *Chaotic Cognition Principles and Applications*, Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hiraki, K. & Suzuki, H. (1998). Dynamic Constraints Relaxation as a Theory of Insight. *Cognitive Studies*, Vol.5, No.2, 69-79.
- Hopfield, J. J. & Tank, D. W. (1985). "Neural" Computation of Decisions in Optimization Problems. *Biological Cybernetics*, Vol.52, 141-152.
- Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint Relaxation and Chunk Decomposition in Insight Problem Solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol.25, No.6, 1534-1555.
- Nakagawa, M. (1987). A Mathematical Model of Approach and Avoidance Behavior in Psychological Field. *Japanese Psychological Research*, Vol.29, No.2, 59-70.
- Nara, S., Davis, P., & Totsuji, H. (1993). Memory Search Using Complex Dynamics in a Recurrent Neural Network Model. *Neural Networks*, Vol.6, 963-973.
- Nozaki, H., Yokoyama, S., Isomoto, Y., & Yoneda, J. (1996). A Study of Character Frequency - From the View Point of Japanese Language Education. *Educational Technology*, Vol.20, No.3, 141-150.
- Seifert, C. M., Meyer, D. E., Davidson, N. S., Patalano, A.L., & Yaniv, I. (1995). Demystification of Cognitive Insight: Opportunistic Assimilation and the Prepared-Mind Perspective. In R. J. Sternberg & J. E. Davidson (eds.) *The Nature of Insight*, Cambridge, MA: MIT press.
- Smith, S. M. (1995). Fixation, Incubation, and Insight in Memory and Creative Thinking. In Smith, S. M., Ward, T. B., & Finke, R. A. (eds.) *The Creative Cognition Approach*, Cambridge, MA: MIT press.
- Smith, S. M. & Tindell, D. R. (1997). Memory Blocks in Word Fragment Completion Caused by Involuntary Retrieval of Orthographically Related Primes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol.23, No.2, 355-370.
- Tani, J. (1996). Model-Based Learning for Mobile Robot Navigation from the Dynamical System Perspective. *IEEE Transaction on System, Man, and Cybernetics*, Part B Cybernetics, Vol.26, No.3, 421-436.
- Weisberg, R. W. & Alba, J. W. (1981). An Examination of the Alleged Role of "Fixation" in the Solution of Several "Insight" Problems. *Journal of Experimental Psychology, General*, Vol.110, No.2, 169-192.
- Yaniv, I., Meyer, D., & Davidson, S. (1995). Dynamic Memory Processes in Retrieving Answers to Questions: Recall Failures, Judgments of Knowing, and Acquisition of Information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol.21, No.6, 1509-1521.

# Declarative and Procedural Learning in Alphabetic Retrieval

Todd R. Johnson (Todd.R.Johnson@uth.tmc.edu)

Hongbin Wang (Hongbin.Wang@uth.tmc.edu)

Jiajie Zhang (Jiajie.Zhang@uth.tmc.edu)

Department of Health Informatics<sup>1</sup>

University of Texas – Houston Health Science Center

7000 Fannin Suite 600

Houston, TX 77030

## Abstract

This paper presents three experiments that study declarative and procedural learning in alphabetic retrieval. It is based on the view that speed-up during skill acquisition can result from acquiring either new procedural knowledge or new declarative knowledge, followed by speed up of both types of knowledge. In addition, both lead to different predictions of transfer due to the different retrieval characteristics of declarative and procedural knowledge. Specifically the paper uses three forms of alphabet arithmetic problems: 1)  $A + 3 = ?$ , 2)  $D - 3 = ?$ , and 3)  $? + 3 = D$ , to further examine the acquisition and use of declarative and procedural knowledge. The first two forms replicate experiments conducted by Rabinowitz and Goldberg (1995), whereas the third experiment attempts to maximally discriminate between declarative and procedural skill acquisition. The results provide further support for the hypothesis that speed-up can result from either declarative or procedural acquisition and strengthening.

## Introduction

The alphabet is a long and well-learned list. How such lists are structured in memory and how they are accessed is certainly one of the fascinating inquiries in cognitive science.

Different from the traditional serial memory tasks, where subjects are to recall a newly-learned list of items in a specified order, in alphabetic retrieval, the content retrieval is trivial. The central interest in alphabetic retrieval research is to study how the sequence information is maintained and accessed in human memory.

The techniques often used to study alphabetic retrieval can be roughly classified into two categories based on whether letters are to be actually retrieved or not. In the first category, no letters need to be retrieved. In Lovelace and Snodgrass (1971), for example, subjects were presented two letters in a pair, and had to judge if the two are in the correct alphabetic order. In the second category, one or more letters have to be retrieved. An example is the experiment by Lovelace, Powell, and Brooks (1973), in which subjects were presented a pair of letters and were

instructed to retrieve (recite) the letters between the two (see also Browman and O'Connell, 1976).

A more elegant technique in the second category is the so-called *alphabet arithmetic* task. In an alphabet arithmetic task, subjects are presented a letter, *letter1*, and a *number*. The goal is to retrieve the letter, *letter2*, that is *number* letters after (or before) *letter1*. Reaction Time (RT) is usually measured. This technique can either take the form of questions like “what comes three letters after K” (e.g., Lovelace & Spence, 1972; Hovancik, 1975; Klahr, Chase, & Lovelace, 1983), or appear in the pure algebraic form (e.g., Rabinowitz & Goldberg, 1995; Johnson, Wang, & Zhang, 1998). An example of the latter is “ $K+3=?$ ”. Subjects have to provide *N* as the answer, because *N* is 3 letters after *K*.

Various alphabet arithmetic studies all produce a more or less consistent result pattern. When RTs are plotted as a function of the serial positions of *letter1* (the stimulus), the curve ascends non-monotonically at the aggregated level, with local peaks and valleys. However, there is no general agreement upon the theoretical explanation.

Klahr, Chase, and Lovelace (1983) proposed a theory of the cognitive structure and process involved in alphabetic retrieval. According to this theory, the alphabet is represented hierarchically. At the top level, the whole list is represented as a set of groups. At the second level, each group is represented as a set of letters. Alphabetic retrieval is a search process that occurred sequentially on both levels. First, the correct group that the to-be-retrieved letter is in has to be found. Second, the letter then has to be found within that group. Both processes are conducted by self-terminating, serial searches, starting with the initial item at each level.

Obviously, based on this theory, the gradually ascending pattern of alphabetic retrieval results from this self-terminating, serial search process: it takes longer to retrieve later letters in the alphabet. In addition, since the search occurs on two levels, a sawtooth-shaped RT curve with local peaks and valleys is evident: valleys and peaks appear at the beginning (with minimum second-level search) and the end (with maximum second-level search) of each group, respectively.

---

<sup>1</sup> Portions of this research were conducted when the authors were at The Ohio State University, Departments of Pathology (Dr. Johnson) and Psychology (Drs. Wang and Zhang).

Although the idea that serial lists are represented hierarchically in memory is quite popular (e.g., Anderson & Bower, 1973; Estes, 1972; Johnson, 1991; Shiffrin & Cook, 1978; Slamecka, 1967), some researchers argue that such a structural speculation is unjustified and a simple associative model is at least equally plausible (e.g. Scharroo, Leeuwenberg, stalmeier, & Vos, 1994). According to this association idea, the alphabet is not represented hierarchically but as a single-level associative chain. In addition, alphabetic retrieval is often a direct access rather than a serial search from the very beginning. RTs in an alphabet arithmetic task are determined by the association strengths between the stimulus and the answer. The difference in association strengths is a function of past experience of how the alphabet is learned and practiced. In this view, therefore, the increasing RT curve across the alphabet is a result of the overall decreasing association strength across the alphabet. The concept of group (or chunk), which is critical in the hierarchical view and is assumed to be responsible for the fine structure of the RT curve, is nothing more a series of letters with strong associations.

Both views explain the data reasonably well (see Scharroo, Leeuwenberg, Stalmeier, & Vos, 1994). As a result, the debate continues. Fortunately, recent progress in cognitive architectures and distinction between declarative and procedural knowledge (e.g., Anderson, 1993; Anderson & Lebiere, 1998) shed new light on how alphabetic retrieval might work. Instead of treating alphabetic retrieval as either serial searching or direct associative access, it is now possible to incorporate the two views in a unified framework of declarative/procedural distinction. Specifically, the knowledge of alphabet arithmetic can be represented either declaratively (e.g., “N is 3 letters after K”) or procedurally (e.g., “To find out the letter that is 3 letters after K, count from K three times, and output the result”). While procedural knowledge is universally applicable and supports more general problem solving, such as searching, it is time consuming. On the contrary, declarative knowledge supports direct memory retrieval thus is fast, but it is conditioned on the availability of the specific declarative knowledge. As a result, when a certain problem can be solved based on stored declarative knowledge, direct retrieval is applied – no serial counting is necessary. On the other hand, when the specific knowledge necessary to solve the problem is not readily retrievable, some problem solving methods based on generally-purposed procedural knowledge, such as serial counting, have to be used. This declarative/procedural approach incorporate the hierarchical searching view and the direct association view in the sense that strong association strengths are represented by retrievable declarative knowledge, and when retrievable declarative knowledge is not available, active searching with the aid of procedural knowledge starts.

Whether alphabetic knowledge is represented declaratively or procedurally is determined by, among other things, past experience. Repeatedly solving a problem procedurally may eventually result in declarative knowledge of that problem. Rabinowitz and Goldberg (1995)

nicely illustrated this phenomenon. In one of their experiments, they asked subjects to solve 432 alphabet arithmetic problems and measured their RTs. For one group of subjects, the 432 problems include a set of 12 different problems, each repeated 36 times. For another group of subjects, the problem set consists of a set of 72 different problems, each repeated 6 times. They found that although the two groups had the same RTs at the beginning of training, the first group solved the problems much faster than the second group in the later stage of training. The reason, they argued, is that both groups solved problems procedurally (i.e., by counting) at the beginning. Since the first group solved the same set of problems over and over again, they acquired declarative knowledge about these problems and began direct retrieval in the later stage. The second group did not get enough practice for any problem, thus they kept procedurally searching in the entire session.

The idea that people solve problems by applying both declarative and procedural knowledge, whichever is appropriate, has received much support (e.g., Anderson & Lebiere, 1998; Reder & Ritter, 1992; Siegler, 1988). However, how well this framework can be applied to account for various alphabetic retrieval tasks remain unexplored. It is the purpose of this paper to report a study that empirically investigates the declarative/procedural distinctions in alphabet arithmetic.

## Experiment

For any specific alphabet arithmetic fact (e.g., C is 2 letters after A), we distinguish three different evaluation forms (see Figure 1). The first one is the standard *addition* form, in which subjects are presented “A + 2 = ?” and required to produce “C”. The second form is a *subtraction* form, in which subjects are asked to produce “A” with respect to the problem “C – 2 = ?”. The third form is called *match*. In a match form, subjects are presented “? + 2 = C”, and have to report “A” as the answer.

Addition:	A + 2 = ?
Subtraction:	C – 2 = ?
Match:	? + 2 = C

**Figure 1.** Three forms of alphabet arithmetic

Solving the three problems in Figure 1 essentially requires the same alphabet arithmetic fact. However, due to the different evaluation forms, different declarative and/or procedural representations might be applied. More specifically, if it is available and retrievable, a single piece of declarative knowledge, “C is 2 letters after A”, can be used to quickly solve both the addition and match problems. However, a different piece of declarative knowledge, “A is 2 letters before C”, has to be available and retrievable to quickly solve the subtraction problem. On the other hand, when relevant declarative knowledge is not retrievable, these problems have to be solved procedurally. Specifically, while the addition problem requires

a count forward procedure (i.e., “A, B, C..”), both the subtraction and match problems require a count backward procedure (i.e., “..C, B, A”). In addition, due to the addition format in the match problem, an extra step may be needed to convert it to a recognizable subtraction format so that the count backward procedure can be applied.

118 subjects from The Ohio State University participated in the experiment. They were divided into three groups, with at least 30 subjects in each group. A learning-transfer paradigm was adopted. In the learning phase, all three groups of subjects learned to solve alphabet arithmetic problems in the addition form. In the transfer phase, each group of subjects was instructed to solve only one type of problems, either addition, subtraction, or match.

One critical manipulation in the experiment is that each subject group was further divided into two subgroups, with each having different learning experience. Specifically, in the *consistent* subgroup, subjects solved a set of 12 problems over and over again, with each problem presented 36 times. In the *varied* subgroup, subjects solved a set of 72 problems, with each only presented 6 times. It is hypothesized that subjects in the consistent subgroups gradually developed declarative knowledge about the problems that they solved repeatedly, while subjects in the varied subgroups did not due to insufficient practice.

The experimental design is shown in Table 1. It is clear that in the transfer phase, subjects in the addition group were presented new problems that they had not seen in the learning phase, although they were in the same addition form. On the contrary, subjects in the subtraction and match groups were presented a subset of the problems they had seen in the learning phase, but in different forms. It is important to note that a portion of the experiment is essentially a replication of Rabinowitz and Goldberg (1995)’s experiment.

**Table 1.** Experimental Design\*

	Learn- ing	Transfer		
		addition “A+2=?” [30]	subtraction “C-2=?” [36]	match “?+2=C” [39]
consis- tent	$\alpha$ (36)	$\beta_2$	$\alpha$	$\alpha$
varied	$\alpha+\beta_1$ (6)	(2)	(3)	(3)

\*An alphabet arithmetic problem takes the form of letter1 +/- number = letter2. In the experiments, letter1, letter2  $\in \{A, B, \dots, Z\}$ , number  $\in \{1, 2, \dots, 6\}$ . With the constraint that the problem must be valid, we have a total of 135 problems. Let  $\alpha$  be a set of 12 such problems, where each possible number appears twice. Let  $\beta$  be a set of 96 such problems, and  $\alpha \cap \beta = \emptyset$ . In addition,  $\beta_1$  contains 60 problems of  $\beta$ , and  $\beta_2$  contains the other 36 problems. Each possible number appears in  $\beta_1$  10 times, and in  $\beta_2$  6 times. The numbers in parentheses are the number of times each problem was presented. The numbers in brackets are the numbers of subjects.

The results show that subjects could solve these alphabet arithmetic problems quite accurately. The overall error rate is 8%. There were 13 subjects who either did not follow the instruction (e.g., writing down the alphabet on a piece of paper) or had more than 20% errors. They are excluded in further analysis. More detailed error rate information, conditioned on subject groups and experimental phases, is shown in Table 2. It is clear that subjects made significantly more errors in the transfer phase, especially when they tried to solve problems in different evaluation forms.

**Table 2.** Error Rate

%	Addition	Subtraction	Match	Overall
Learning	6.0	7.8	6.4	6.8
Transfer	7.5	30.2	25.9	18.9
Overall	6.3	9.5	7.9	8.0

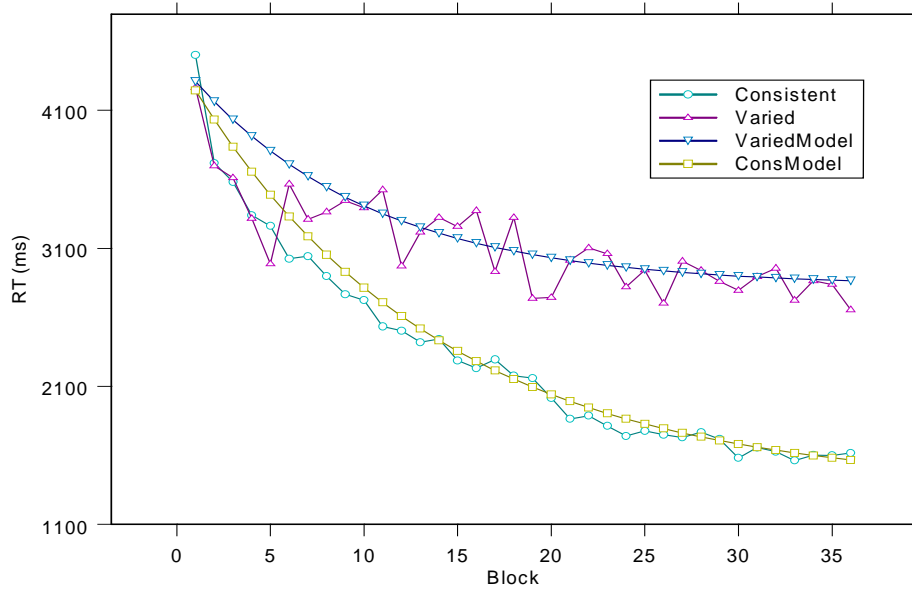
The RT results are presented separately for the learning performance and the transfer performance.

Since all three groups of subjects were trained in the same evaluation forms, the learning data is combined across the three groups. Following the practice of Rabinowitz and Goldberg (1995), to show the trend, we divided the total number of trials (432) into 36 blocks, with 12 trials in each. For each subject, we calculated the median RT of each block. Then the mean of these medians was computed across the subjects. The results are shown in Figure 2, separately for the consistent and varied conditions. It indicates that although subjects showed the same level of performance at the beginning of learning, the subjects in the consistent condition solved problems much faster toward the later stage of learning than those in the varied condition. Statistics confirmed the result. A non-linear mixed-effect exponential model was nicely fitted to the data in each condition, and the fitting curve is also shown in Figure 2. Examining the parameter estimations, it is shown that the two conditions differ significantly in terms of both the decay rate ( $z=-2.07$ ,  $p<0.05$ ) and the asymptote ( $z=14.15$ ,  $p<0.01$ ). The effect size of the asymptote difference, about 1457ms, is a strong support for the argument that different problem solving strategies were adopted in the later stage of training.

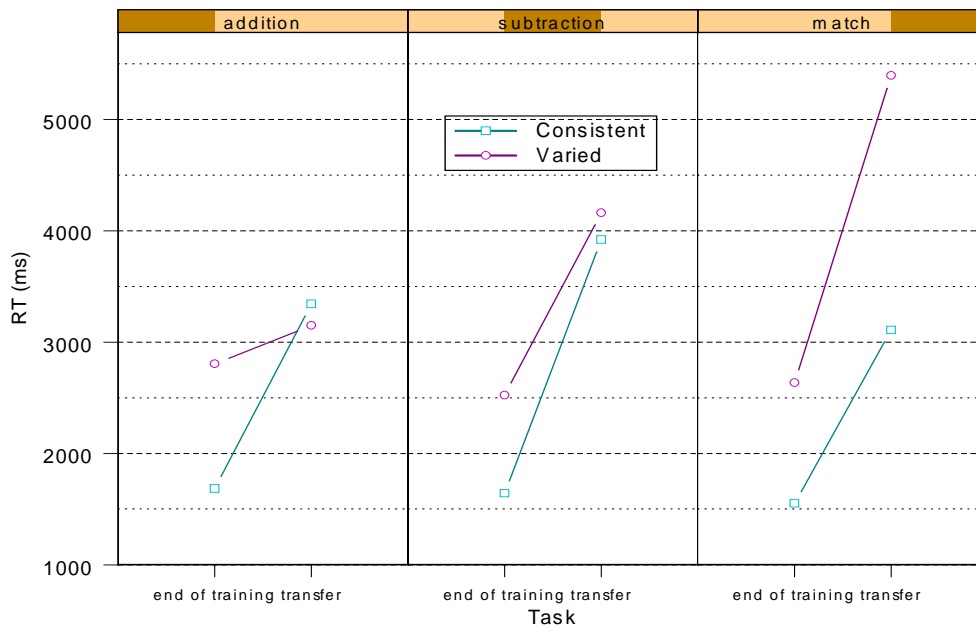
The transfer performance, conditioned on three groups of subjects and two learning conditions, is shown in Figure 3. It is easy to observe that the transfer effect (i.e., the difference between the end-of-training performance and the transfer performance) is quite different across the experimental manipulations. An overall three-way MANOVA shows that the three-way interaction among the transfer condition (addition, subtraction, or match), the learning condition (consistent or varied), and the transfer itself (the end of training performance vs the transfer performance) is significant ( $F(2,99)=6.23$ ,  $p<0.01$ ). Further analyses show that, 1) while the interaction between the learning condition and the transfer effect is significant in both the addition and match groups ( $F(1,28)=34.20$ ,  $p<0.01$ ;  $F(1,37)=5.15$ ,  $p<0.05$ , respec-

tively), it is not significant in the subtraction group; 2) In the consistent learning condition, the transfer effect is significant in all three groups ( $F(1,13)=82.63$ ,  $p<0.01$ ;  $F(1,19)=33.46$ ;  $p<0.01$ ;  $F(1,20)=64.58$ ,  $p<0.01$ ; for the addition, subtraction, and match groups, respectively). In addition, both the end-of-training performance and the transfer performance are not significantly different across the three groups; 3) In the varied learning condition, the interaction between the transfer effect and the transfer condition is significant ( $F(2,47)=7.39$ ,  $p<0.01$ ): the trans-

fer effects in the three transfer conditions are 345.3ms, 1637.7ms, and 2759.1ms, respectively; and 4) while in both the addition and subtraction transfer conditions, the performance in the transfer stage is not different between the consistent and varied learning conditions (3150ms vs 3343ms, and 4162ms vs 3922ms, respectively), in the match transfer condition, variedly-trained subjects performed much worse (i.e., longer RTs) than those consistently-trained (5396ms vs 3109ms,  $F(1,37)=23.12$ ,  $p<0.01$ ).



**Figure 2.** The Learning Performance.



**Figure 3.** The Transfer Performance.

## Conclusions and General Discussions

Overall the results are consistent with our hypothesis about declarative/procedural distinction and interaction. First, in the addition-transfer condition, it is assumed that consistent training leads to declarative knowledge about the 12 over-learned problems, and varied training results in the counting procedure being well practiced though no declarative knowledge has been acquired. In the transfer stage, since different and new addition problems were presented, no relevant declarative knowledge was available, which left the counting procedure the only appropriate means. As a result, subjects with varied training benefited and showed a transfer effect because they had practiced and speeded up their counting procedure during their extensive training. On the contrary, subjects with consistent training showed no transfer to the new addition problems, presumably because the declarative knowledge they gained was specific to the training problems thus not useful and meanwhile they did not practice enough their counting procedure in the training.

Second, in the subtraction-transfer condition, both consistently-trained and variedly-trained subjects basically faced the same new challenge – counting down the alphabet. For those with consistent training, although the transfer problems were essentially equivalent to the training problems, due to different evaluation forms, declarative knowledge about these problems was not applicable in the transfer stage. In other words, it seems likely that most subjects did not realize that they could use their memory of addition results to solve subtraction problems. As a result, they had to join those variedly-trained subjects to try to adopt the brand new counting-back procedure to solve those transfer problems. Both groups showed no transfer.

Finally, the match problem serves an excellent condition to maximally discriminate between procedural and declarative learning. According to the model of alphabet arithmetic described above, to solve a match problem, subjects could use either declarative knowledge or procedural knowledge. To solve a match problem declaratively, one need only match the problem with their declarative knowledge in order to find an answer, which would suggest a perfect transfer when the corresponding declarative knowledge is available. To solve a match problem procedurally, normally one would need to first recognize that the problem is actually a subtraction problem by doing an algebraic transformation, then adopt a procedure to count back through the alphabet, which would suggest very little transfer from the previous training.

In the current match condition, subjects with varied training in fact had no choice: they had to solve the match problems procedurally, simply because they had not acquired the relevant declarative knowledge. This explains the worst transfer performance in the varied match condition. On the contrary, for those subjects with consistent training, they actually had a choice. They could solve the match problems by either using their newly acquired de-

clarative knowledge, thus producing perfect transfer, or, they could adopt the similar converting-and-counting-back procedure, thus producing much worse transfer. The current data seems to indicate that subjects adopted neither one. In the consistent match condition, the transfer is neither perfect nor as bad as that in the varied match condition. It seems that subjects somehow combined the two approaches to solve the transfer problems. One possible scenario is that subjects used declarative knowledge to fetch the possible answer and then used a counting procedure to verify the solution. Another possibility is that some subjects used a declarative strategy and some subjects used a procedural strategy, which, when aggregated, produces the resultant pattern.

Johnson, Wang, and Zhang (1998) described an Act-R model of alphabet arithmetic that accounted for the results of the two experiments conducted by Rabinowitz and Goldberg. How this model can be applied to the match condition is of great importance to further clarify the declarative and procedural learning and application issues in alphabetic retrieval. Such a model is currently under development.

## Acknowledgements

This work is funded in part by Office of Naval Research Grant No. N00014-95-1-0241.

## References

- Anderson, J. R. (1993). *Rules of the Mind*, Hillsdale, NJ: Lawrence Erlbaum.
- Anderson, J. R. & Bower, G. H. (1973). *Human associative memory*. New York: Winston.
- Anderson, J. R. & Lebiere, C. (1998). *The Atomic Components of Thought*. Hillsdale, NJ: Lawrence Erlbaum.
- Browman, C. P. & O'Connell, D. C. (1976). Sequential phonological effects in recitation times. *Bulletin of the Psychonomic Society*, 8, 37-39.
- Estes, W. K. (1972). An associative basis for coding and organization in memory. In A. W. Melton & E. Martin (Eds.), *Coding Processes in Human Memory* (pp.161-190). New York: Winston.
- Hovancik, J. R. (1975). Reaction times for naming the first next and second next letters of the alphabet. *American Journal of Psychology*, 88, 643-647.
- Johnson, G. L. (1991). A distinctiveness model of serial learning. *Psychological Review*, 98, 204-217.
- Johnson, T. R., Wang, H., & Zhang, J. (1998). Modeling speed-up and transfer of declarative and procedural knowledge. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum
- Klahr, D., Chase, W. G., & Lovelace E. A. (1983). Structure and process in alphabetic retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(3), 462-477.
- Lovelace, E. A., Powell, C. M., & Brooks, R. L. (1973). Alphabetic position effects in covert and overt alphabetic recitation times. *Journal of Experimental Psychology*, 99, 405-408.



- Lovelace, E. A. & Snodgrass, R. D. (1971). Decision times for alphabetic order of letter pairs. *Journal of Experimental Psychology*, 88, 258-264.
- Lovelace, E. A. & Spence, W. A. (1972). Reaction times for naming successive letters of the alphabet. *Journal of Experimental Psychology*, 94, 231-233.
- Rabinowitz, M. & Goldberg, N. (1995). Evaluating the structure-process hypothesis. In F. E. Weinert & W. Schneider (Eds.), *Memory Performance and Competencies: Issues in Growth and Development* (pp. 225-242). Hillsdale, NJ: Lawrence Erlbaum.
- Reder, L. M. & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 435-451.
- Scharroo, J., Leeuwenberg, E., Stalmeier, P. F. M., & Vos, P. G. (1994). Alphabetic search: Comment on Klahr, Chase, and Lovelace (1983). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(1), 236-244.
- Shiffrin, R. M. & Cook, J. R. (1978). A model for short-term item and order retention. *Journal of Verbal Learning and Verbal Behavior*, 17, 189-218.
- Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General*, 117, 258-275.
- Slamecka, N. (1967). Serial learning and order information. *Journal of Experimental Psychology*, 74, 62-66.

# A Process Model of Children's Early Verb Use

Gary Jones (gaj@Psychology.Nottingham.AC.UK)  
Fernand Gobet (frg@Psychology.Nottingham.AC.UK)  
Julian M. Pine (jp@Psychology.Nottingham.AC.UK)  
School of Psychology, University of Nottingham,  
Nottingham, NG7 2RD, England

## Abstract

The verb-island hypothesis (Tomasello, 1992) states that children's early grammars consist of sets of lexically-specific predicate structures (or verb-islands). However, Pine, Lieven and Rowland (1998) have found that children's early language can also be built around lexical items other than verbs, such as pronouns (this contradicts a strict version of the verb-island hypothesis). This paper presents a computational model (called MOSAIC), which constructs a network of nodes and links based on a performance-limited distributional analysis of the input (mother's speech). The results show that utterances generated from MOSAIC: (1) more closely resemble the child's data than the child's mother's data on which MOSAIC is trained; and (2) can readily simulate both the verb-island and other-island phenomena which exist in the child's data.

## Introduction

One of the most influential recent constructivist accounts of early grammatical development is Tomasello's (1992) verb-island hypothesis. According to this view children start producing multi-word speech without knowledge of syntactic categories, such as noun and verb. Instead, children's early language use is based on a "functionally based distributional analysis" (Tomasello, 1992, p.28) of the language they hear. This analysis assigns predicate<sup>1</sup> status to specific words based on their function in sentences. For example, in the sentence "Adam kicks the ball", the roles of Adam and the ball are centred around "kicks", such that Adam is someone who can kick things, and the ball is something that can be kicked. The lexical item "kick" is therefore assigned a predicate role which takes as arguments a "kicker" (Adam) and a "kickee" (the ball).

The notion of "verb-island" arises because most predicates are verbs in adult language and the arguments the predicate takes are specific to that predicate (e.g., "kickers" and "kickees"). Based on this idea, children's early grammar will consist of inventories of verb-specific predicate structures (i.e., verb-islands). For example, the child will use any object which it knows has performed kicking as the antecedent to "kick". Verb-general marking (e.g., knowing that someone who *kicks* can also be someone who *hits*) does not occur until the formation of a verb category.

---

<sup>1</sup> For Tomasello, a predicate is a lexical item (typically a verb) which forms the main relational structure of a sentence. Arguments are the lexical items to which the predicate relates. Therefore in the sentence "John walks the dog", "walks" is the predicate and "John" and "dog" are the arguments.

In agreement with Ninio (1988), Tomasello argues that children will only start to construct word categories such as noun and verb when they begin to use instances of these categories as the arguments of predicates (e.g., using "ball" as an argument to the predicate "kick"). As verb-islands often use nouns as their arguments, children should form noun categories relatively early in their language development. Verb categories will only be formed later when children begin to use verbs as the arguments of other predicates (e.g., in double-verb constructions such as "Want to + V" or "Can't + V").

The verb-island hypothesis can account for a number of phenomena in children's early multi-word speech. First, it can explain the lexically-specific patterning of children's early verb use. For example, Tomasello (1992) has shown that in the early stages of grammatical development his daughter's ability to generate longer sentences built up piecemeal around particular verbs, and failed to generalise to new verbs which typically entered her speech in very simple structures. Second, it can explain the restricted nature of children's early word order rules. For example, Akhtar and Tomasello (1997) have shown that young children not only fail to generalise Subject-Verb-Object (SVO) word order knowledge from one verb to another, but are also unable to use it as a cue for sentence comprehension with novel verbs. Third, it can explain differences in the flexibility with which children use nouns and verbs in their early multi-word speech. For example, Tomasello and his colleagues have shown that young children will readily use novel *nouns* as arguments in familiar verb structures but tend to restrict their use of novel *verbs* to the structures in which they have heard those same verbs modelled in the input (Akhtar & Tomasello, 1997; Olguin & Tomasello, 1993; Tomasello & Olguin, 1993).

One weakness of the hypothesis is that there are aspects of children's early multi-word speech that do not fit a strict version of the verb-island hypothesis. For example, Pine, Lieven and Rowland (1998) have shown that many children acquire structures based around high frequency items which Tomasello would not define as predicates (e.g., case-marked pronouns such as "I" and "He" and proper-nouns such as "Mummy" and the child's name). Moreover, these pronoun and proper-noun islands not only seem to be functioning as structuring elements in children's speech, but as structuring elements which accept verbs as slot fillers. These data suggest that the lexical specificity of children's early multi-word speech is not always "verb-specificity" or even "predicate-specificity" (because verbs can be slot fillers of other structures). Verb-island effects may simply be a

special case of more general frequency effects on children's acquisition of lexically-specific structures.

This paper presents a computational model called MOSAIC (Model of Syntax Acquisition in Children), which combines naturalistic input (mother's speech) and a performance-limited distributional learning mechanism in order to produce child-like utterances as output. The results will show that MOSAIC is able to: 1) simulate verb-island phenomena that are consistent with children's early multi-word speech; 2) simulate other-island phenomena which exist in children's early multi-word speech but which are problematic for a strict version of the verb-island hypothesis; and hence 3) provide a process-based explanation of why some lexical items come to function as "islands" in the child's grammar and others do not.

### The MOSAIC model

MOSAIC is a variant of EPAM/CHREST (De Groot & Gobet, 1996; Gobet, 1996; Gobet & Simon, in press) which creates a discrimination network (a hierarchical structure of nodes which are linked together) based on a given input. Discrimination networks have a root node at the top of the hierarchy, with all other nodes cascading from the root node (see Figure 1 for an example). Nodes are connected to each other by links. This section will describe the basic working of MOSAIC, and then give an example of MOSAIC's learning mechanisms using mother's speech as input.

#### A general overview of MOSAIC

MOSAIC's discrimination network begins with a root node (which always contains no information). As in other models of the EPAM family (Feigenbaum & Simon, 1984), learning occurs in two steps. The first step involves traversing the network as far as possible with the given input, taking one feature of the input at a time. This is done by starting at the root node and examining all the test links from the root node, selecting the first test link whose test is fulfilled by the first feature in the input (when beginning learning, only the root node will exist and therefore no tests can be fulfilled). The node at the end of the test link now becomes the current node and the next feature of the input is applied to all the test links immediately below this node. The traversal continues until a node is reached where no further traversing can be done (either because the current input feature fulfilled none of the tests of the node's test links, or the current node has no test links below it). Traversing the network in this way is also how information can be output from the network (this will be explained later).

The second step involves adding new information, nodes, and test links. The full input is compared to the information at the final node that was reached by traversal. Based on this comparison, learning can arise in two ways:

1. *Discrimination*. When the input information mismatches the information given at the node (the *image*), a new test link and node are added to the tree below the node that has just been reached. The new test

will relate to the next immediate mismatched feature in the input.

2. *Familiarisation*. When the input information is under-represented by the image (the information given at the node), additional feature(s) from the input are added to the image. The information in the node will contain all information that led to the node during traversal, plus any additional feature(s).

Discrimination therefore creates nodes and test links, and familiarisation creates or modifies the information contained in nodes. The amount of information stored at nodes increases with their distance from the root, because each node contains the accumulation of information of all the nodes that were accessed in traversing to the node.

There are two constraints that are imposed when learning by discrimination and familiarisation. First, before creating a node containing more than one input feature (i.e., a sequence of features), the individual features in the sequence must have been learnt (each input element is said to be a *primitive*). Second, all nodes containing just one input feature are linked to the root node (i.e., all primitives are immediately below the root node; in this way all sequences of input features are below the node which represents the initial feature in the sequence).

Learning can also occur *whilst* traversing the network. MOSAIC compares each node traversed with other nodes in the network to see if they have a similar usage. Similar usage means that there are common test links below each of the two nodes. When this is the case, a lateral link is created between the nodes (this is explained further in the following section).

#### An example of MOSAIC learning an utterance

The input given to MOSAIC consists of a set of mother's utterances. Each line of input corresponds to a single utterance (delimited by an END marker which signifies the end of the utterance), and each word in the utterance is an input feature. The example utterance "Who came to see you on the train?" will be used as input to illustrate how MOSAIC learns.

The first input feature ("who") is applied to all of the root node's test links in the network. As the network is empty, there are no test links. At this point MOSAIC must discriminate because the input feature "who" mismatches the information at the root node (the root node information is null). The discrimination process creates a new node, and a test link from the root node to the new node with the test "who" (see Figure 1). MOSAIC must then familiarise itself with the input feature, in order to create the "who" information in the image of the node.

When encountering the same input for a second time, the test link "who" can be taken, and the input can move to the next feature, "came". As the node "who" does not have any test links below it, then under normal circumstances discrimination would occur below the "who" node. However, MOSAIC has not yet learnt the input feature

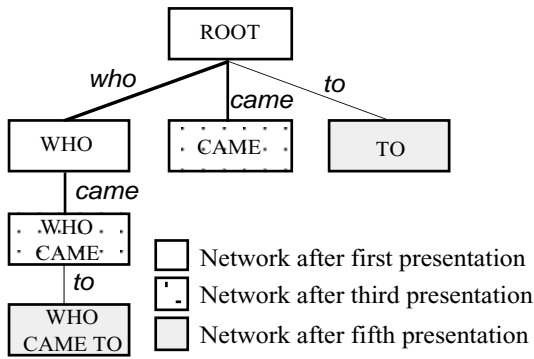


Figure 1: Structure of the MOSAIC net after five presentations of the input “Who came to see you on the train”.

“came”, and so discrimination occurs below the root node. Familiarisation will then fill the image of the new node with “came”. The third time the input is seen, the “who” test link can be taken, and the input can move onto the next feature (“came”). No further test links are available, but the input “who came” mismatches the information at node “who” and so discrimination occurs. A new node “who came” is created (see Figure 1). Familiarisation will fill in the image of the new node.

After a total of five presentations of “Who came to see you on the train?”, the network will have learnt the phrase “Who came to” (see Figure 1). This simple example serves to illustrate how MOSAIC works; in the actual learning phase each utterance is only used once, encouraging a diverse network of nodes to be built.

During traversal of the network, lateral links can be created. A lateral link is a link between any two nodes in a MOSAIC network (excepting the root node). Lateral links are designed to link together nodes which are used in the same manner. Usage is based on the test links that are immediately below a particular node. The way that MOSAIC creates nodes and test links means that all the test links that are below a particular node will consist of the word or words that follow that node in the input (as shown in the previous section). For example, in Figure 2, the words “moves”, “sits”, “walks”, and “chases” must have followed “cat” in the input, meaning sentences such as “cat sits down” have been seen in the input.

When there is a significant amount of overlap between words or phrases that follow a particular word in the input (i.e., there is significant overlap between the test links that are below two particular nodes) then the two nodes can be linked by a lateral link. The minimum number of test links which must overlap for a lateral link to be created is determined by an *overlap parameter*. Using the network in Figure 2 as an example, “cat” and “dog” will have a lateral link between them when the overlap parameter is set to 3 because at least 3 of the test links below “cat” and “dog” are shared. The next section shows how lateral links are used when generating output from MOSAIC.

### Generating utterances from a MOSAIC network

Utterances can be generated from MOSAIC by beginning at the root node and traversing down until encountering a node which contains an END marker (i.e., the last word in the utterance must be one which ended an utterance in the input). Whilst traversing down the network, both test links and lateral links can be taken. To help explain how utterances are generated from the network, test links will be called *rote links* hereafter, and lateral links will be called *generative links*. This is because test links are created from rote learning, and lateral links are created from overlap in node use. When traversing the network, if only rote links are taken then the resulting utterance must have been present in the input (because of the dynamics of the creation of the discrimination network, traversing down from the root node will always produce a phrase that existed as a full utterance or part of an utterance in the input). However, when a generative link is taken, the resulting utterance may never have been seen before in the input.

When generative links exist, MOSAIC can take these links as part of the traversal of the network. For example, in the network shown in Figure 2, the generated utterance could begin with “cat”, take the generative link across to “dog”, and then continue the utterance with any phrase that follows “dog” (i.e., the remainder of the phrase is built up by traversing the nodes below “dog”). This produces novel utterances that were not seen before in the input, such as “cat runs” and “dog moves”. Currently, only one generative link is taken per traversal of the network in order to limit the number of generated utterances (the next section

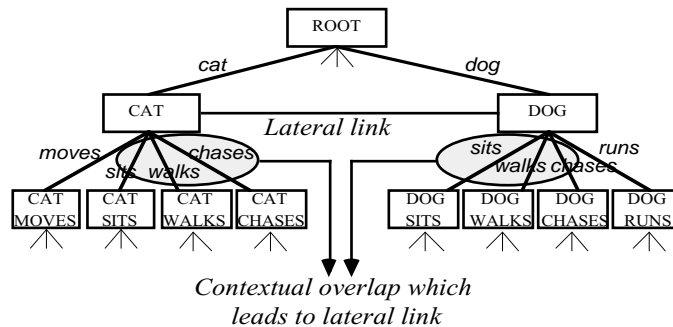


Figure 2: Example of how lateral links are created.

shows that taking only one generative link enables the network to produce over seven generated utterances to every one rote learned utterance).

### Modelling verb-island phenomena

The verb-island hypothesis states that children’s early language consists of lexical items (typically verbs) existing as predicates, which take other lexical items as arguments. As lexical items such as pronouns cannot, in Tomasello’s terms, be predicates, then for flexibility the terms frame and slot filler will be used in place of predicate and argument. A frame is therefore a relational structure of a sentence and the slot fillers to the frame are the lexical items which relate to the frame. For example, the sentence “Daddy moves the chair” has “moves” as the frame and “Daddy” and “chair” as the slot fillers.

The verb-island hypothesis can be confirmed if the language data contain verbs which exist as frames (i.e., verbs which take several different lexical items as slot fillers), and contain *very few other lexical items* which exist as frames. To examine this, the language data will be analysed by extracting verb+common-noun and common-noun+verb sequences. Common-nouns, rather than all lexical items, are examined because: 1) they are the most common category in children’s speech; 2) Tomasello (1992) predicts that children form noun categories earlier than verb categories based on their use as slot fillers (i.e., they should be used often as the slot fillers of verb frames); and 3) the analysis is more tractable with only two types of lexical item.

To investigate whether other-island phenomena exists, pronoun+verb and proper-noun+verb combinations will be extracted and analysed. Pronouns are used because a strict version of the verb-island hypothesis does not allow pronouns to act as islands. Also, pronouns occur with high frequency in the child’s data and are often followed by a verb (i.e., they may show verbs being used as slot fillers to other frames). Proper-nouns are used for an additional test of other-islands.

### Method

#### Subject data

Three sets of data are compared for the verb-island phenomena: the utterances from one child, Anne; the utterances from Anne’s mother; and the utterances from

Table 1: Sample of the utterances generated from MOSAIC.

MOSAIC utterance
I forgotten
That’s my toes again
Where’s the magic bag
And she like them
Baby put the sheep in the farmyard
What about the camel
All on the settee
Who can you see on here
He didn’t catch me

MOSAIC when trained using Anne’s mother’s utterances as input. The utterances for Anne and her mother were taken from the Manchester corpus (Theakston, Lieven, Pine & Rowland, in press) of the CHILDES database (MacWhinney & Snow, 1990). The corpus consists of transcripts of the mother-child interactions of twelve children over a period of twelve months. The transcripts contain both the utterances and the syntactic categories (e.g., noun, verb) of all words in the utterances. The child focused on here, Anne, began at age 1;10.7 and completed the study at age 2;9.10. Her starting MLU (Mean Length of Utterance) was 1.62 with a vocabulary size of 180.

For Anne there were 17,967 utterances (i.e., utterance tokens), of which 8,257 utterances were unique (i.e., utterance types). There were 7,331 multi-word utterance types. For Anne’s mother, there were 33,390 utterance tokens, 19,358 utterance types, and 18,684 multi-word utterance types. A random sample of 7,331 of Anne’s mother’s multi-word utterance types were taken to match Anne for quantity of data.

#### MOSAIC data

MOSAIC was trained on the full 33,390 utterance tokens of Anne’s mother in chronological order, one utterance at a time (as a list of words). MOSAIC’s overlap parameter was set to 15. The input to MOSAIC did *not* contain any coding information. This means that MOSAIC was not presented with any information about the categories of words (e.g., that “dog” was a noun or “go” was a verb) or about noun or verb morphology (e.g., “going” was seen rather than the morpheme “-ing” attached to the root form of the verb “go”).

After MOSAIC had seen all of the input utterances, every possible utterance that could be generated was output. This resulted in 178,068 utterance types (21,510 produced by rote and 156,558 produced by generation). Examples of the utterances generated from MOSAIC are shown in Table 1. The analyses of the data from MOSAIC are based on a random sample of 7,331 (i.e., matching Anne for quantity) of the multi-word utterance types produced by *generation*, because these are the novel utterances that will not have existed as part of the mother’s input.

#### Procedure

The utterances for both the child and mother included the syntactic category for each word in an utterance. The codings for the child’s utterances were used to determine the categories of words in the utterances of the child; the codings for the mother’s utterances were used to determine the categories of words in the utterances of the mother. Some words (such as “fire”) belong to more than one category. In these cases, a category was only assigned if the word was used as that category in at least 80% of the instances in which the word was used. For MOSAIC’s utterances, the categories were calculated based on the codings from the mother’s utterances.

The three sets of data were analysed in the same way. The method of extracting verb+common-noun combinations is detailed here but the method is the same

Table 2: Percentage of the 7,331 multi-word utterances from Anne, Anne’s mother, and MOSAIC that contain nominal+verb or verb+nominal combinations. The nominals are broken down into pronoun, proper-noun, and common-noun combinations.

Pair distribution	Anne		Anne’s mother		MOSAIC	
	Nominal+ Verb	Verb+ Nominal	Nominal+ Verb	Verb+ Nominal	Nominal+ Verb	Verb+ Nominal
<b>Pronouns</b>	4.73%	4.60%	8.83%	6.15%	5.16%	2.58%
<b>Proper-nouns</b>	1.31%	0.61%	1.94%	1.49%	0.55%	0.64%
<b>Common-nouns</b>	1.91%	7.41%	5.65%	10.42%	1.16%	5.18%

for the extraction of common-noun+verb, pronoun+verb, and proper-noun+verb combinations.

Each utterance was searched for a word which was categorised as a verb. The two words following the verb-category word were examined to see if either occurred as a common-noun. If so, the verb+common-noun pair was stored for analysis. Verbs were then converted to their root form (e.g., “going” and “goes” both become “go”) and common-nouns to their singular form (e.g., “dogs” becomes “dog”), and any duplicate pairs were removed. Analysis was therefore conducted on types, not tokens. The number of slot fillers for a verb is the number of different common-noun types that were paired with that verb.

### How well does the output of MOSAIC match the subject data?

Table 2 shows the percentage of each set of 7,331 multi-word utterances from Anne, Anne’s mother, and MOSAIC that contained verb+nominal and nominal+verb combinations (the label “nominal” refers to the group of all pronouns, proper-nouns, and common-nouns).

The data show that the utterances from MOSAIC match

Anne much more closely than they do Anne’s mother (on whose utterances MOSAIC was trained). For example, 5.16% of MOSAIC’s utterances and 4.73% of Anne’s utterances contain pronoun+verb combinations, compared with 8.83% for Anne’s mother. In fact, despite all three datasets having been matched for overall sample size, Anne’s mother produces many more instances of *every* combination shown in Table 2 (e.g., producing over twice as many different nominal+verb combinations [16.42%] as Anne [7.95%] and MOSAIC [6.87%]).

### Verb-islands exist in the data

As explained earlier, the data are expected to show that verbs act as frames (taking lots of different common-nouns as slot fillers) whereas common-nouns are not expected to act as frames. Whether this is true can be examined by looking at the number of common-noun types that follow verb types, and vice versa. We operationalise the concept of an “island” as a lexical item which acts as a frame for at least ten different slot fillers (e.g., a verb type would have to have ten different common-noun types as slot fillers). For example, for Anne, the verb “Find” is an island because it

Table 3: Verb-island data for Anne, Anne’s mother, and MOSAIC (mean=mean number of slot fillers for each frame type; islands=number of frames that have 10 or more slot fillers).

Data source	Mean	Islands	Islands having the most slot fillers
<b>VERB+COMMON-NOUN (frame=verb; slot filler=common-noun)</b>			
Anne	6.24	10	Get, Put, Want, Go, Need, Make
Mother	5.97	13	Get, Put, Want, Need, Have, Find
MOSAIC	9.74	10	Get, Put, Eat, Think, Want, Find
<b>COMMON-NOUN+VERB (frame=common-noun; slot filler=verb)</b>			
Anne	1.51	1	Baby
Mother	2.08	4	Baby, Animal, Dolly, Penguin
MOSAIC	1.57	1	Baby
<b>PRONOUN+VERB (frame=pronoun; slot filler=verb)</b>			
Anne	21.69	10	I, You, He, It, That, They, We
Mother	27.65	11	You, I, He, We, She, They, It
MOSAIC	25.20	12	You, It, That, I, He, We, She
<b>PROPER-NOUN+VERB (frame=proper-noun; slot filler=verb)</b>			
Anne	5.65	3	Anne, Mummy, Daddy
Mother	3.23	3	Anne, Mummy, Daddy
MOSAIC	6.67	2	Anne, Mummy

is followed by ten common-noun types (“Dolly”, “Plate”, “Seat”, “Welly-boot”, “Baby”, “Ribbon”, “Hat”, “Duck”, “Pen”, and “Bird”). Table 3 shows these data for Anne, Anne’s mother, and MOSAIC. This shows that there are many verb-islands for all three sources of data, but very few common-noun islands. In both cases, MOSAIC provides an identical match to Anne for number of islands.

#### **Other-islands exist in the data**

Table 3 shows that both pronoun-islands and proper-noun islands exist for Anne, Anne’s mother, and MOSAIC. The pronoun-islands are particularly strong (the mean number of slot fillers for pronouns is more than 20 for all three sets of data) and because pronouns take verbs as slot fillers, these islands are problematic for a strict version of the verb-island hypothesis which predicts that only verbs are initially used as frames. The other-islands, as Table 3 shows, are readily simulated by MOSAIC.

#### **Discussion**

The output from MOSAIC more closely resembles the child than the child’s mother, demonstrating that MOSAIC is doing more than just a straightforward distributional analysis of its input. In fact, it is a combination of the performance-limitations imposed on the model (e.g., learning one word at a time), and the frequency of occurrence of items in the input, that enable MOSAIC to match the child data. MOSAIC seeks to maximise the information held at nodes in the network, but can only do so for input sequences that occur frequently (e.g., due to limitations in only learning one item at a time). MOSAIC therefore offers a process-based explanation of why some lexical items come to function as “islands” in children’s grammar and others do not: children are maximally sensitive to the high frequency lexical items that exist in their input.

The results presented here show that when combined with naturalistic input, a simple distributional learning mechanism is able to provide an effective simulation of child language data. The simulations show that first, it is possible to model verb-island phenomena as the product of a frequency-sensitive distributional analysis of the child’s input, and, second, that the same mechanism can also simulate other-island patterns which are problematic for a strict version of the verb-island hypothesis.

#### **Acknowledgements**

This research was funded by the Leverhulme Trust under grant number F/114/BK.

#### **References**

- Akhtar, N., & Tomasello, M. (1997). Young children’s productivity with word order and verb morphology. *Developmental Psychology*, 33, 952-965.
- De Groot, A. D., & Gobet, F. (1996). Perception and memory in chess: Studies in the heuristics of the professional eye. Assen: Van Gorcum.
- Feigenbaum, E. A., & Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science*, 8, 305-336.

- Gobet, F. (1996). Discrimination nets, production systems and semantic networks: Elements of a unified framework. In *Proceedings of the 2nd International Conference on the Learning Sciences*, 398-403. Evanston, IL: Northwestern.
- Gobet, F., & Simon, H. A. (in press). Five seconds or sixty? Presentation time in expert memory. *Cognitive Science*.
- MacWhinney, B., & Snow, C. (1990). The Child Language Data Exchange System: An update. *Journal of Child Language*, 17, 457-472.
- Ninio, A. (1988). On formal grammatical categories in early child language. In Y. Levy, I. M. Schlesinger, & M. D. S. Braine (Eds.), *Categories and processes in language acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Olguin, R., & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development*, 8, 245-272.
- Pine, J. M., Lieven, E. V. M., & Rowland, C. F. (1998). Comparing different models of the development of the English verb category. *Linguistics*, 36, 807-830.
- Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (in press). The role of performance limitations in the acquisition of ‘mixed’ verb-argument structure at stage 1. In M. Perkins & S. Howard (Eds.), *New directions in language development and disorders*. Plenum.
- Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge: CUP.
- Tomasello, M., & Olguin, R. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development*, 8, 451-464.

# Handedness and Heterogeneity in Cognitive Science

**Gregory V. Jones** (g.v.jones@warwick.ac.uk)  
Department of Psychology; University of Warwick  
Coventry CV4 7AL, UK

**Maryanne Martin** (maryanne.martin@psy.ox.ac.uk)  
Department of Experimental Psychology; University of Oxford  
South Parks Road, Oxford OX1 3UD, UK

## Abstract

Outside of cognitive neuropsychology, it is often assumed that differences among individuals in cognitive activity may be adequately represented theoretically in terms only of quantitative variation across a population. A possible exception to the presumption of homogeneity within cognitive processing is explored here. It is shown that left-handed and right-handed populations exhibit consistent, qualitative differences in their remembering of orientational information. It is concluded that the subject matter of cognitive science may be more heterogeneous than is commonly assumed.

## Assumption of Homogeneity

A widespread implicit assumption in cognitive science is that its subject matter is homogeneous, in the sense that differences in cognitive processing among individuals may generally be expressed in terms of merely quantitative variation. A major exception to this assumption is commonly recognised in the field of cognitive neuropsychology, where the cognitive processes of individuals have been shown to exhibit a range of idiosyncrasies associated with different forms of physical damage to the brain (e.g., Jones & MacAndrew, 1990). For this field, the occurrence of a double dissociation of function is generally held to indicate heterogeneity of population (e.g., Jones, 1983), although the validity even of this inference has been challenged within cognitive science (e.g., Juola & Plunkett, 1998). But for those without brain damage, differences among individuals are often viewed within cognitive science as essentially random, with variation in cognitive activity among individuals represented theoretically in terms of dispersion around a central tendency. If this is a correct characterisation of much of cognitive science,

then it raises the question of whether the assumption, outside cognitive neuropsychology, of homogeneity among individuals is justified in particular circumstances.

Where may qualitative differences in cognitive processes among individuals be manifested? A classical area of investigation is that of personality (e.g., Martin, 1985). However, the domain of handedness is also an appropriate area to consider. Can the models and descriptions of cognitive science be applied indifferently, as is generally assumed, to the right-handed majority and to the left-handed minority? Or do fields of heterogeneity exist in which people's handedness influences their cognitive performance? Empirical evidence that allows these questions to be addressed is considered here. First, however, it is appropriate to consider briefly the distinctive characteristics of handedness itself.

## Handedness Populations

Most human beings exhibit a preference for the use of one or other hand. This preference is not evenly distributed between left and right, as it is for most animals. Instead, the predominant pattern of limb preference is for use of the right hand.

Although hand preference can be influenced by social pressures (e.g., Harris, 1990), it has a number of features which suggest that it is also under genetic influence (e.g., Corballis, 1997; Laval, Dann, Butler, Loftus, Rue, Leask, Bass, Comazzi, Vita, Nanko, Shaw, Peterson, Shields, Smith, Stewart, DeLisi, & Crow, 1998). For example, Klar (1996) has reported that the likelihood of a person being left-handed is increased if one of the parents of the person, although right-handed, in turn had two left-handed parents. We have shown (Jones & Martin, 2000) that a genetic model for



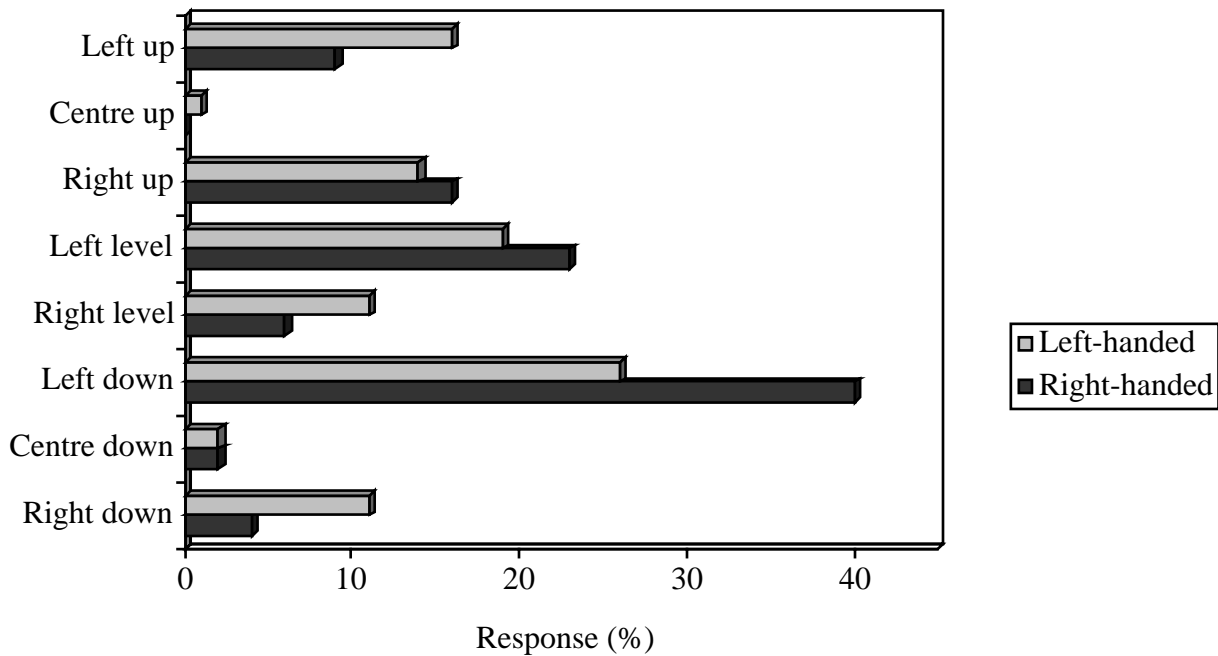


Figure 1. Orientation responses

handedness may be formulated which accounts satisfactorily for this and a number of other similar effects.

It has frequently been suggested (e.g., Day & MacNeilage, 1996) that asymmetry in limb use, via an accompanying specialisation of function in the cerebral hemispheres, played a decisive role in the evolution of language. Similarly, the degree of lateralization of language function between the hemispheres is known to differ between left-handed and right-handed populations (e.g., McManus, 1999). However, it is less clear that cognitively based performance itself differs between the left-handed and right-handed populations. Thus, despite considerable research, it has proven difficult to establish reliable associations between handedness and either developmental reading disorder (e.g., Bishop, 1990) or a variety of symptoms linked to immune disorders (cf. Geschwind & Galaburda, 1987; Bryden, McManus, & Bulman-Fleming, 1994). Indeed, until recently there has been suggestive evidence of a heterogeneity in cognitive function between left-handed and right-handed populations in only one field, that of chimeric perception.

### Heterogeneity for Cognitive Function

Chimeric faces may be constructed by artificially pairing their left and right halves. If people are asked to match a control face to either a chimeric face composed only of the left half (and its mirror image) or a chimeric face composed only of the right half (and its mirror image), it has been reliably demonstrated (e.g., Levy, Heller, Banich, & Burton, 1983; Luh, Redl, & Levy, 1994) that right-handed people, but not left-handed people, have a significant tendency to select the left half (plus its mirror image) face as the better match.

The chimeric finding appears to represent a genuinely cognitive, not hemispheric, effect since it occurs with unrestricted fixation and therefore is not related to visual field (and hence hemisphericity). However, it may also be noted that the effect is a relatively narrow one. If the assumption of homogeneity of cognitive processing broke down only in this limited field, then the case for a wider consideration of heterogeneity in cognitive science would be relatively weak. It is now becoming apparent, however, that heterogeneity is demonstrable in the wider area of memory for orientation (e.g., McKelvie & Aikins, 1993; Martin & Jones, 1998). Two further studies of memory for orientation, which are described next, confirm this finding. The first also investigates whether the heterogeneity extends

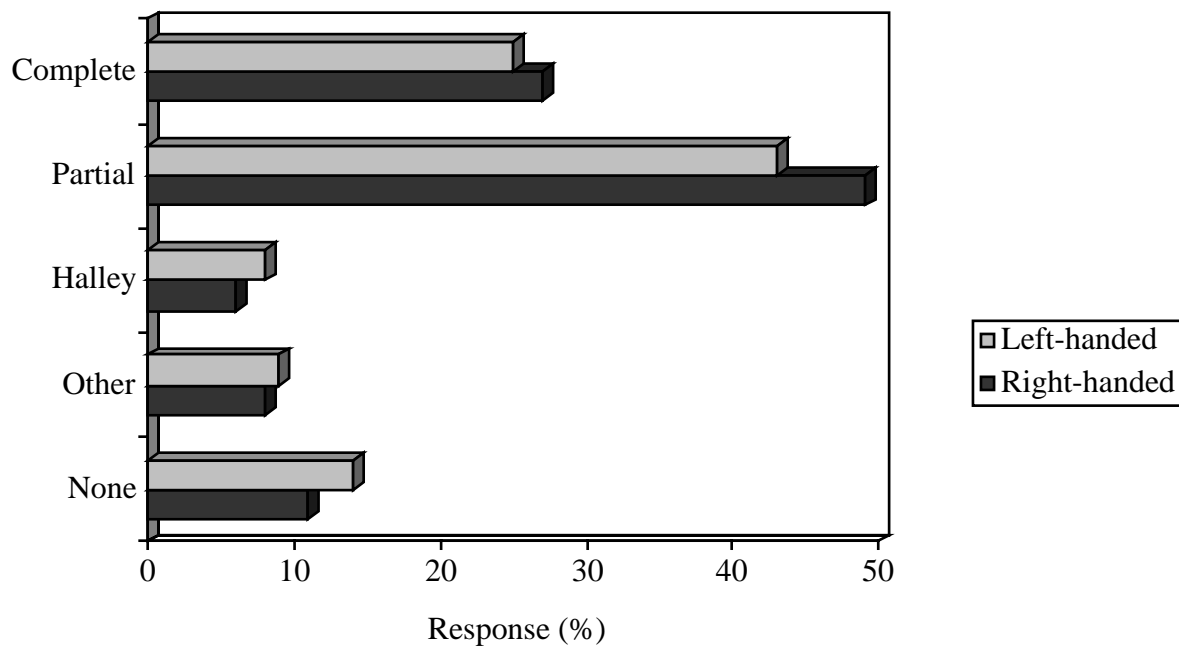


Figure 2. Name responses

to memory for other types of information, and the second also investigates whether it resides genuinely in memory or alternatively in strategic behavior.

### Extent of Heterogeneity

Are differences between left-handed and right-handed populations in cognitive functioning confined to memory for orientation, or do they extend to memory for abstract information? This question was investigated by examining people's memory for Comet Hale-Bopp, selected as subject because of the long history of popular interest in cometary appearances (see Schechner Genuth, 1997).

Approximately equal numbers of left-handed and right-handed participants were tested (N = 401). Testing occurred approximately six months after the comet's visit. Participants were tested on a series of items "about the comet which was visible to the naked eye over the Easter period." Questions probed both abstract and concrete knowledge. Binary handedness classifications were made on the basis of the hand which the participant used for drawing.

Figure 1 shows memory for orientation for left-handed and for right-handed participants. Recall of the direction of the head of the comet was classified into eight different sectors,

defined by the combination of it pointing leftwards, centrally, or rightwards; and downwards, level, or upwards. There was a significant difference between the frequency distributions of responses for left-handed and right-handed participants,  $\chi^2(7) = 20.29$ ,  $p < .01$ . In particular, right-handed participants produced a significantly greater number than left-handed participants of responses with the comet facing down to the left (the orientation most frequently encountered),  $\chi^2(1) = 7.86$ ,  $p < .01$ . Similarly, considering downward and level responses overall, it can be seen from Figure 1 that there was a contralateral tendency which associated right-handed participants with left-facing responses, and vice versa; this tendency also was significant,  $\chi^2(1) = 12.97$ ,  $p < .001$ . Similar results were found with recognition rather than recall responses.

Figure 2 shows the distributions of written name responses which were made by left-handed and by right-handed participants. Recall was classified as either (a) completely accurate (both Hale and Bopp), (b) partially accurate (incomplete or misspelled), (c) Halley (either a semantic error or an approximation to Hale), (d) unrelated name, or (e) no response. There was no significant difference between the two frequency distributions,  $\chi^2(4) = 3.08$ . Similar results were found for the recall of other

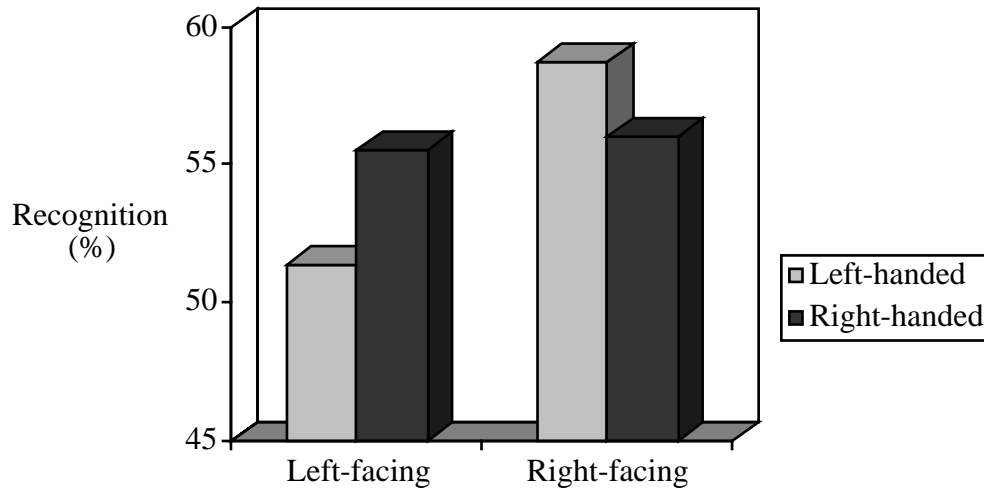


Figure 3. Contralateral handedness effect in recognition

abstract information, such as the length of time since the comet's last visit to Earth (about four thousand years).

The present findings suggest therefore that the assumption of homogeneity, which breaks down in the case of memory for orientation, continues to hold in the case of memory for more abstract information.

### Heterogeneity for Memory or for Strategy?

Although overt responses concerning memory for information have been shown to differ for left-handed and right-handed populations, it is possible in principle that the underlying difference between these populations relates not to their memory processing but instead to their strategic behavior. That is, it is possible that left-handed and right-handed populations differ not in their likelihoods of retrieving information about orientation, but instead in their strategies of producing responses when memory fails. To investigate this possibility, the confidence with which responses are produced can be examined. If heterogeneity is confined to strategic behavior, then differences between populations should arise only for responses that are made with relatively low confidence. But if heterogeneity applies to memory itself, then differences should be observed in those responses which are made with high confidence.

Approximately equal numbers of left-handed and right-handed participants were tested ( $N = 230$ ). Each participant was shown a sequence of

40 different black-and-white photographs for 3 sec each. In half of the photographs a person faced to the left of the viewer and in half a person faced to the right. Subsequently, each photograph was shown alongside its mirror image (reflected in a vertical plane) in a two-alternative forced-choice recognition task. In addition, for each recognition response the participant assigned a confidence level on a scale between 1 (guess) and 5 (certain).

Figure 3 shows the overall levels of recognition. There was a significant interaction between the effects of the direction in which the stimulus faced and the handedness of the participant,  $F(1, 228) = 9.18, p < .01$ . It can be seen that the effect was a contralateral one, in that left-facing stimuli were recognised better by right-handed than by left-handed participants, whereas right-facing stimuli were recognised better by left-handed than by right-handed participants.

To examine the possible influence of confidence, further analyses were carried out on the recognition responses that were made with the lowest level of confidence (1) and those made with the highest confidence (4 or 5). Confidence level was found to modify the two-way interaction, yielding a significant three-way interaction,  $F(1, 174) = 4.19, p < .05$ . Decomposing the three-way interaction, it was found that for those responses made with high confidence there was again a significant interaction between the effects of stimulus direction and of handedness,  $F(1, 201) = 5.06, p < .05$ . In contrast, for those responses made

with low confidence, there was no significant interaction.

Thus it is in memory processing itself, rather than in strategic behavior, that the assumption of homogeneity appears to break down. Left-handed and right-handed populations differ in how they remember orientation, not in how they guess.

### **Origin of Population Effect for Memory**

It is important to note that the results do not suggest that either left-handed or right-handed populations enjoy a general advantage in memory. Rather, the finding is one of contralaterality, in that left-handed people were more accurate than right-handed people when recalling right-facing stimuli, but less accurate when recalling left-facing stimuli. This zero-sum finding of contralaterality presents a problem for any theorists (e.g., Luh, Redl, & Levy, 1994; McKelvie & Aikins, 1993) who attempt to explain the influence of handedness upon cognition in terms of possible correlated differences in hemispheric specialization of function, because such a theory would predict that either left-handed or right-handed people should show a consistent advantage in performance.

In contrast, we have recently proposed (Martin & Jones, 1999) that the consistent differences among people in patterns of overt motor activity which categorise them as either left-handed or right-handed are accompanied by correlated differences in motor imagery. It is well established that extensive motor activation occurs in the cortex in the absence of physical movement (e.g., Decety, Grezes, Costes, Perani, Jeannerod, Procyk, Grassi, & Fazio, 1997; Jeannerod, 1994; Logie, 1995). Characteristic patterns of motor activation for left-handed people differ from those for right-handed people, partly in response to the asymmetric nature of the everyday environment (e.g., left-to-right writing). The present results suggest that, depending upon the precise details of a cognitive task, either left-handed or right-handed motor imagery may prove to be the more effective in assisting memory for orientation.

### **Conclusions**

Outside of cognitive neuropsychology, it is often assumed that differences in cognitive activity between individuals may be adequately represented theoretically in terms of random variation around a central tendency. One

exception to the implicit assumption of homogeneity within cognitive science has been characterised here. Subtle differences can be detected in the remembering of orientation by left-handed and right-handed populations. It remains to be investigated, however, how widespread is the occurrence of such heterogeneity.

### **Acknowledgement**

Research support has been provided by ESRC (UK) Grant R000236216.

### **References**

- Bishop, D. V. M. (1990). *Handedness and developmental disorder*. Oxford, UK: Blackwell.
- Bryden, M. P., McManus, I. C., & Bulman-Fleming, M. B. (1994). Evaluating the empirical support for the Geschwind-Behan-Galaburda model of cerebral lateralization. *Brain and Cognition*, 26, 103-167.
- Corballis, M. C. (1997). The genetics and evolution of handedness. *Psychological Review*, 104, 714-727.
- Day, L. B., & MacNeilage, P. F. (1996). Postural asymmetries and language lateralization in humans (*Homo sapiens*). *Journal of Comparative Psychology*, 110, 88-96.
- Decety, J., Grezes, J., Costes, N., Perani, D., Jeannerod, M., Procyk, E., Grassi, F., & Fazio, F. (1997). Brain activity during observation of actions - influence of action content and subject's strategy. *Brain*, 120, 1763-1777.
- Geschwind, N., & Galaburda, N. (1987). *Cerebral lateralization: Biological mechanisms, associations and pathology*. Cambridge, MA: MIT Press.
- Harris, L. J. (1990). Cultural influences on handedness: Historical and contemporary theory and evidence. In S. Coren (Ed.), *Left-handedness: Implications and anomalies* (pp. 195-258). Amsterdam: North-Holland.
- Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences*, 17, 187-245.
- Jones, G. V. (1983). On double dissociation of function. *Neuropsychologia*, 21, 397-400.
- Jones, G. V., & MacAndrew, S. B. G. (1990). Uniformity of associative impairment in amnesia. *Proceedings of the twelfth annual conference of the Cognitive Science Society* (pp. 750-756). Hillsdale, NJ: Erlbaum.

- Jones, G. V., & Martin, M. (2000). A note on Corballis (1997) and the genetics and evolution of handedness: Developing a unified distributional model from the sex-chromosomes gene hypothesis. *Psychological Review*, *107*, 213-218.
- Juola, P., & Plunkett, K. (1998). Why double dissociations don't mean much. *Proceedings of the twentieth annual conference of the Cognitive Science Society* (pp. 561-566). Mahwah, NJ: Erlbaum.
- Klar, A. J. S. (1996). A single locus, *RGHT*, specifies preference for hand utilization in humans. In *Cold Spring Harbor Symposia on Quantitative Biology: Vol. 61. Function and dysfunction in the nervous system* (pp. 59-65). Plainview, NY: Cold Spring Harbor Laboratory Press.
- Laval, S. H., Dann, J. C., Butler, R. J., Loftus, J., Rue, J., Leask, S. J., Bass, N., Comazzi, M., Vita, A., Nanko, S., Shaw, S., Peterson, P., Shields, G., Smith, A. B., Stewart, J., DeLisi, L. E., & Crow, T. J. (1998). Evidence for linkage to psychosis and cerebral asymmetry (relative hand skill) on the X chromosome. *American Journal of Medical Genetics*, *81*, 420-427.
- Levy, J., Heller, W., Banich, M. T., & Burton, L. (1983). Asymmetry of perception in free viewing of chimeric faces. *Brain and Cognition*, *2*, 404-419.
- Logie, R. H. (1995). *Visuo-spatial working memory*. Hillsdale, NJ: Erlbaum.
- Luh, K. E., Redl, J., & Levy, J. (1994). Left- and right-handers see people differently: Free-vision perceptual asymmetries for chimeric stimuli. *Brain and Cognition*, *25*, 141-160.
- Martin, M. (1985). Neuroticism as predisposition toward depression: A cognitive mechanism. *Personality and Individual Differences*, *6*, 353-365.
- Martin, M., & Jones, G. V. (1998). Generalizing everyday memory: Signs and handedness. *Memory & Cognition*, *26*, 193-200.
- Martin, M., & Jones, G. V. (1999). Motor imagery theory of a contralateral handedness effect in recognition memory: Toward a chiral psychology of cognition. *Journal of Experimental Psychology: General*, *128*, 265-282.
- McKelvie, S. J., & Aikins, S. (1993). Why is coin head orientation misremembered? Tests of schema interference and handedness hypotheses. *British Journal of Psychology*, *84*, 355-363.
- McManus, I. C. (1999). Handedness, cerebral lateralization, and the evolution of language. In M. C. Corballis & S. E. G. Lea (Eds.), *The descent of mind: Psychological perspectives on hominid evolution* (pp. 194-217). Oxford, UK: Oxford University Press.
- Schechner Genuth, S. (1997). *Comets, popular culture, and the birth of modern cosmology*. Princeton, NJ: Princeton University Press.

# Using Referential Communication to Study Mental Models

Julia Kalmanson (kalmanso@psych.columbia.edu)  
Department of Psychology; Columbia University, 406 Schermerhorn Hall  
New York, NY 10027

Arthur B. Markman (markman@psy.utexas.edu)  
Department of Psychology, University of Texas, Mezes Hall 330  
Austin, TX 78712

## Abstract

In this paper we evaluate the utility of a referential communication paradigm to study the content and use of mental models. In this task pairs of people collaborate to determine which of a set of infra-red images depicts a physically possible situation. We demonstrate that the referential communication task provides insight into the interaction between the content and use of naive theories of physics in a problem-solving domain.

## Introduction

Causal explanation is critical for our daily existence. Causal connections support our perception of the world as coherent, and they give us a sense of mechanism,- a sense of how things work. Our control of the physical world is to a great extent dependent on the accuracy of our understanding of the mechanics of the world.

A prominent suggestion is that causal knowledge is organized into theories that people use to reason about the world. The term "theory" has been used in a number of different ways by psychologists (see e.g., Gopnik & Meltzoff, 1997, for a discussion). One view assumes that theories are large bodies of knowledge that are coherently organized according to a few well-defined principles, so that all explanations can be deductively derived (e.g., Kuhn, 1991). A more local view of theories is presented by Murphy and Medin (1985), who define theories as "any host of mental explanations, rather than a complete organized scientific account" (p.290). Schemas and scripts contain implicit theories of causality that allow us to explicate the world, although may not possess properties of coherence and consistency.

One area where theories have received extensive attention is in naive physics, which is concerned with understanding how knowledge and experience are integrated to create an understanding of the mechanics of the world. Despite the fact that physical principles describe properties of objects with which we interact daily, people have serious difficulties understanding formal principles of physics. People's understanding of fundamental physical principles has been described as incoherent and full of misconceptions (diSessa, 1993; McCloskey et al., 1980; Clement, 1982; Cooke and Breedin, 1994).

Various theories have been proposed to account for these difficulties. Some researchers have suggested that these misconceptions arise from basic misunderstandings of physical systems that are formed prior to any formal training in physics (Caramazza and Green, 1980; McCloskey, 1983a, 1983b). The systematic nature of some errors has led researchers to suggest that certain misconceptions are not idiosyncratic, but instead are based on a more general system of beliefs or a naive theory (Clement 1982; McCloskey, 1983a, b). These theories are described as systematic, general, coherent, well-developed and well articulated conceptions that conflict with basic principles of physics, but that nonetheless adequately explain observed events in the world (McCloskey, 1983b).

An alternative view contends that people's understanding of physical phenomena is a collection of fragmented and loosely connected ideas about the world that can be used to generate situation specific explanations (diSessa, 1988, 1993). In his view, naive theories are nothing more than ad hoc explanations that are invented for particular situations. Ueno (1993) in his re-interpretation of diSessa's theory points out that these explanations are socially formed and shared. They are maintained through communication and are to a great extent guided by conversational pragmatics. For example it would be anomalous to cast a simple sentence like "Susie slapped Tom" in it's Newtonian physics equivalent of "Tom's face slaps the palm of Susie's hand, while the force of Susie's slapping is the same as the force of Tom's slapping". In everyday discourse the latter sentence would be judged nonsensical.

This reinterpretation is in line with current research on the role of communication in category acquisition. Category representations are structured in a manner that facilitates communication. People typically learn categories in the process of communicating with others. Further, people are constrained to form categories that are shared by other members of their culture if they are to use them effectively (Garrod & Doherty, 1994; Malt, 1995). In this view naive theories of physics are pragmatically motivated explanations of complex phenomena that are socially constructed to support our simplistic categorizations of the physical world.

Ultimately, of course, we would like to understand the structure of people's naive theories. In the present paper,

we begin to address this issue by examining a novel method that can elicit people's naive theories of the physical world and to explore the causal relationships that make up those theories. This methodology draws on recent findings on the role of communication in category acquisition, and attempts to elicit and explore naive theories in a communicative setting.

A popular method of eliciting people's theories of physical phenomena is to ask people to explain their predictions or decisions in interviews or other verbal protocols. These verbal protocols permit a systematic examination of explanations people generate for their own errors in reasoning. Chi and her colleagues have successfully used this method to study learning in a variety of problem-solving tasks (Chi, 1989, 1983). The method we propose is novel in that it incorporates a referential communication design into the study of naive theories. In this task, pairs of people are presented with four infra-red pictures that show the heat emanating from objects. One of the pictures is an actual infra-red image, and the other three are doctored images that have been altered to contain systematic errors. Pairs of people are shown these images and are asked to determine which image is correct.

In order to perform this task, dyads must talk about the heat pictured in the image. In this way, they must use their naive understanding of thermodynamics. Because the task involves two people, many aspects of people's beliefs about heat are stated explicitly in the conversations. People have extensive experience with heat and have a naive theory of heat-flow.

Referential communication tasks are quite data-intensive, as full transcripts of conversations are developed and must be coded. In this paper, we limit our focus to three issues. First, does communicating about this task improve performance? This question is useful for understanding whether theory development occurs through communication. If the performance of dyads improves with communication, then it is plausible to think that theories develop when people communicate. Second, we are interested in whether people use correct theories when discussing thermodynamics. Finally, we will address the relationship between theories and topics of discussion, as well as the qualitative change in discussions over time.

## Method

**Participants:** Participants were 70 members of the Columbia University community (50 in the dyad task and 20 in the control task). Six dyads involved two male participants, five involved two female participants, and the rest were mixed sex dyads. All participants were native speakers of English who did not know each other before the session. Data from one dyad had to be eliminated due to mechanical failure leaving a total of 24 dyads for analysis.

**Materials and Procedure:** The stimuli were 12 sets of false-color infra-red images of familiar objects and scenes such as plants, kitchen appliances, and street scenes. Each

set consisted of one actual image and three variants of it. The actual image was a picture of the infra-red (i.e., thermal) energy at the surfaces of the image. The color scheme involved 10 colors that were each assigned to a range of temperatures. The resulting image is called a false-color image, because it appears in colors that differ from the colors of those objects in visible light. To complete each set three additional versions of each picture were created using Adobe Photoshop by changing some of the colors in the image to create thermal inconsistencies that are highly unlikely to occur naturally. For example, the nose of a dog might be made to appear cooler than the fur-covered skin, or the pattern of heat diffusion from a heat source might be changed so that temperature did not decrease monotonically with distance.

During the experiment the sets were presented in a random order. Each pair of participants was instructed to collaborate to figure out which image in the set was the actual thermal image. Both subjects had to agree on their response before the trial was completed. To encourage discussion rather than pointing, a divider was placed between the subjects, and each subject was given an identical set of 4 images. Thus, subjects were free to refer to pictures verbally but could not point to pictures or their elements to establish reference. The discussions were videotaped and later transcribed. All subjects were aware of being videotaped. The control group consisted of 20 subjects who performed the same picture selection task alone and without verbalizing their reasons.

**General Coding:** Each utterance from the transcript was coded along six dimensions. An utterance was defined as a turn each subject took when speaking. Thus, an utterance could contain as little as a sentence fragment or as much as a paragraph. Because of space limitations, we will focus on two codes: 1). the correctness of the theory and action taken by the dyad and 2). the topic of the discussion. To assess the reliability of the coding, ten of the transcripts were scored by both coders. Correlations ranging between .9 and .98 were obtained for all codes.

The correctness of the theory and action code focused on utterances where the dyad took an action (either selecting a particular picture as the correct one or rejecting a picture). First, the action was coded as correct or incorrect. A correct action was either rejecting a picture that was not an actual thermal image, or selecting the valid image. An incorrect action was either rejecting the correct image or accepting an incorrect one. Actions were typically justified in some way, and the theory part of the code assessed whether the justification was in accord with basic principles of physics. Thus, this code had four levels:

1. Correct action considered on the basis of incorrect theory
2. Correct action considered on the basis of correct theory.
3. Incorrect action considered on the basis of incorrect theory.
4. Incorrect action considered on the basis of correct theory.

The discussion topic code distinguished between five different topics including discussions of abstract physical principles, discussions of the thermal conductivity of materials, and discussions of the internal mechanics of an object depicted. Of these codes only two yielded enough observations to warrant further analysis: 1). discussion of temperature and 2). discussion of heat diffusion. Temperature referred to explicit discussions of the temperature at particular points in the image or to relative temperatures at neighboring points. Discussions of heat diffusion were cases in which people talked about the flow of heat from one location to another or to the dissipation of heat. Because naive theories often treat temperature as a physical quantity (rather than a measure of mean molecular kinetic energy), discussions of temperature are likely to be associated with poor reasoning about thermal images (Wiser & Carey, 1983). In contrast, discussions of heat flow and thermodynamics are more likely to be related to an accurate theory of thermodynamics, and so they should be associated with good reasoning about thermal images.

### Predictions

In this paper, we focus on four aspects of the present task:

1. Communication: The first question to be explored is whether communication influences performance accuracy on this task. To address this issue, we test to see if dyads have higher accuracy than do people who perform the task alone. In addition to examining overall accuracy, we look at performance curves over the course of the experiment. Related to this issue, we can explore how the performance of dyads changes over time.
2. Correctness of theories: Expertise is typically characterized by the presence of a fully integrated representation of the domain of expertise. Experts in domains like physics reason better than novices, because they are able to focus on deep relational aspects of the situation rather than being derailed by surface aspects of the task (e.g., Chi, Feltovich, & Glaser, 1981). In the present task, we expect that subjects who exhibit evidence of having a correct theory of thermodynamics will perform the picture selection task more accurately than those with fragmentary knowledge of this domain.
3. Considered actions: When discussing an image, a dyad could decide to classify it as one of the transformed images or to retain the picture in the set that could be classified as the unaltered thermal image. Considering a particular action in relation to a particular picture singles that picture out and temporarily makes it more salient than others. It may be that mere consideration of an action, be it based on a correct or incorrect theory, affects the decision by anchoring people on a particular picture. That is, if a dyad starts out considering a particular picture and spends extensive time and energy discussing it, that investment alone may be sufficient to influence the final choice.
4. Topics of discussion: As discussed above, temperature and heat differ in their relationship to a correct theory of

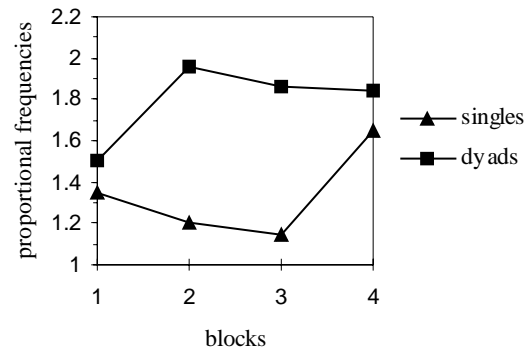
thermodynamics. We will explore the relationship between the topics discussed and people's accuracy in classifying the pictures to see whether talking about terms that are more relevant to correct theories is related to accurate performance.

### Analysis and Results

**Communication:** A key part of the present task was that dyads worked together to find the correct thermal image. As a way of exploring accuracy on this task, we explore differences between the dyads and the control group on their accuracy in selecting the actual thermal image. This analysis addressed the following questions: 1. Are dyads performing better than singles? 2. Are there differences in performance during the experiment? and 3. How does performance of dyads change over time? An examination of overall accuracy revealed that dyads were significantly more accurate ( $M=7.3$  (out of 12)) than were the people in the control group who worked alone ( $M=5.3$ ),  $t=3.1$ ,  $p<.05$ . Chance performance would be 3 correct out of 12. Both groups performed reliably above chance.

To explore how the dyads differed from the control group more carefully, we broke down the performance data into four blocks of three trials. The accuracy for each block for both conditions is shown in Figure 1.

Figure 1: Accuracy by Blocks of Three Trials



As we would expect if the people in each condition were approximately equivalent in their expertise, performance in the first block of trials is about the same in each group. The groups diverge after the first block. By the second block, the dyads are significantly more accurate ( $M=1.96$ ) than the singles ( $M=1.2$ ),  $t= -2.66$ ,  $p<.05$ . The performance of singles does improve with practice, but this does not occur until after the last block ( $M=1.65$ ).

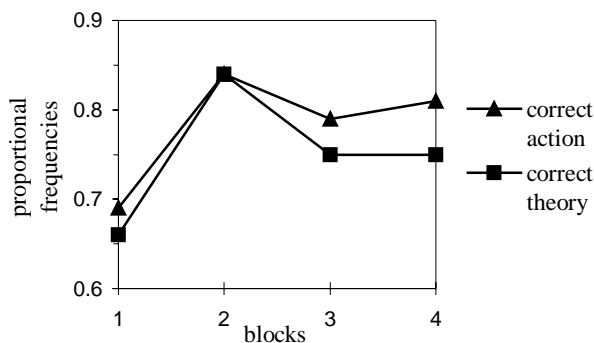
**Theories and Actions:** We now turn to the relationship between type of action, correctness of theory and overall accuracy. For this analysis, we first converted the frequency of each combination of accuracy and theory correctness for each dyad to a proportion. This conversion allowed us to control for individual differences in the length and content of dyads' discussions. We expected that correctness of



theories would be the major factor that determined accuracy. However, this prediction was not borne out. As expected, consideration of a correct action was positively related to accuracy ( $r = .58, p < .01$ ). However, in contrast to our expectations, consideration of an incorrect action was negatively correlated with accuracy regardless of whether the statement was accompanied by a correct or an incorrect theory ( $r(\text{incorrect action/incorrect theory, accuracy}) = -.67$ ;  $r(\text{incorrect action/correct theory, accuracy}) = -.47$ , both  $p < .05$ ). If we examine correctness of action and correctness of theory separately, we also find that correct actions are a better predictor of accuracy than correct theories ( $r = .74, p < .001$  and  $r = .48, p < .05$  respectively).<sup>1</sup>

The changes in performance during the experiment may be related to the consideration of correct actions and theories over the course of the study. Figure 2 shows the frequency with which correct actions and theories are considered as a function of four performance blocks, each consisting of three trials.

Figure 2 Correct Theories and Actions by Blocks of Three Trials

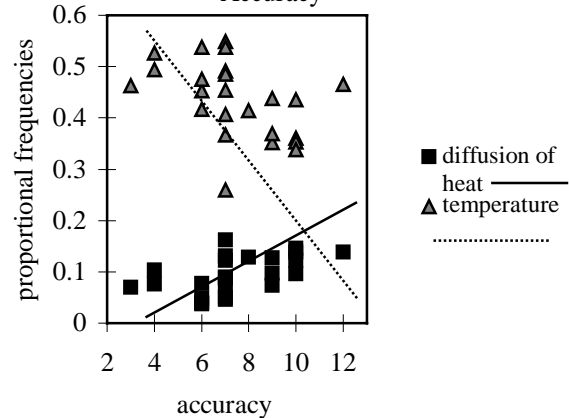


As shown in the graph, the pattern of change in the consideration of correct actions and theories closely follows that of the accuracy improvement pattern. The greatest increase in the frequency of correct actions and theories occurs between the first and the second quarters, followed by a plateau thereafter. Looking at the proportional frequency of correct actions considered on the basis of correct theories and incorrect actions on the basis of incorrect theories we see a similar separation.

**Topics of Discussion:** Finally, we can examine the relationship between accuracy and topics discussed. As described above, we are most interested in discussions of temperature and heat flow. The relationship between the use of these topics and accuracy is shown in Figure 3. As expected, discussion of temperature is negatively related to accuracy ( $r = -.43, p < .05$ ) and discussion of diffusion of heat is positively related to accuracy ( $r = .68, p < .05$ ). This finding reflects that temperature is a static quality of an image that

may signal an incorrect theory, while heat flow is a property that is central to a correct theory of thermodynamics.

Figure 3: Topics of Discussion and Accuracy



We can also look at the relationship between discussion topic and the use of correct actions and correct theories. We expect the discussion of temperature to be negatively related to the use of the correct theory and to the selection of correct actions. Consistent with this prediction, there is a relationship between the frequency of discussion of temperature and the use of incorrect theories ( $r = .56, p < .05$ ). There is also a positive relationship between the frequency of discussion of temperature and the consideration of incorrect actions ( $r = .59, p < .05$ ). In contrast, the frequency of discussions of heat diffusion are related to the consideration of correct actions ( $r = .52, p < .01$ ). Contrary to our expectation, no significant relationship was observed between the frequency of discussion of heat diffusion and the use of correct theories. This unexpected finding will be discussed in more detail below.

Finally, the use of these topics was explored across the four blocks of the experiment. While the frequency of discussion of temperatures did not change over the course of the experiment, there was an increase in the frequency of discussion of diffusion of heat during the experiment.

## Discussion

This research is the beginning of a line of research aimed at exploring the development and use of mental models. In order to get a view of people's mental models (in this case mental models of heat flow), we examined their interactions in a task in which they selected an actual thermal image from a set of doctored images of the same object. The dyadic design allows us to get people to talk about their mental models without having to resort to an unnatural task like a concurrent verbal protocol.

Any study involving transcripts of communication creates a large amount of data, and we have necessarily covered only a small fraction of what could be extracted from this work. We have focused on four main issues here: 1) the

<sup>1</sup> Correct action/incorrect theory code is not represented here because it yielded very low frequencies as compared to frequencies of other codes, and while the results are in the right direction, they were not significant.

influence of causal theories on performance accuracy in a problem solving task; 2) the relationship of considered actions to causal theories, and their influence on accuracy in a problem solving task; 3) the relationship between topics of conversation and causal theories; 4) the qualitative change in discussions and theories over time.

**Theories and Actions:** What sort of theories do people tend to have about thermodynamics? The mean accuracy in this task was 60.8%, which was well above chance. Thus, people who had probably never seen a thermal image before were pretty good at identifying actual images. While this performance must have been based on some knowledge of thermodynamics, it was not based on an accurate physical theory. The data suggest that correct performance in the task was influenced far more strongly by whether the dyad considered correct actions (properly accepting or rejecting a picture) than it was by the presence of accurate discussions of physical principles.

There are (at least) two important reasons for this finding. First, we defined the use of a correct theory as a discussion that was internally consistent and did not contradict basic principles of physics. It is quite likely, however, that this definition is too restrictive. Many people's naive physical models are successful at predicting performance in the world without necessarily embodying principles from the science of physics. McCloskey (1983b) points out that people's naive theories are well-developed conceptions that are useful for predicting the behavior of objects in the world. However, these theories often conflict with basic principles of physics. Further research must explore ways of characterizing people's naive theories of thermodynamics. One task that we have begun to use that has some promise is to give people a blank picture of a scene (such as an outdoor scene during the winter at night) and ask people to color the scene as if they were looking at the heat coming off surfaces. Pilot research with this technique suggests that it is capable of uncovering situations where people's mental model of heat differs from scientifically accepted principles.

A second reason why the presence of a correct model of heat did not always lead to correct performance is that people may have a correct model of thermodynamics, but may have some difficulty translating that model to thermal images. For example, people often correctly recognize that heat will escape from an open window if the room is warmer than the surrounding area. However, they may mistakenly expect the room to look cold in a thermal image, because it would feel cold to be standing in this room with the open window. In fact, such a room would look warm, as an infra-change in the frequency of discussions of diffusion of heat followed the same pattern.

One factor that may account for the difference in the learning curves has to do with the learning benefit associated with constantly verbalizing one's thoughts in a collaborative process. Chi and her colleagues have successfully used talk

red camera would be seeing the heat energy escaping from the room. Thus, people may understand principles of thermodynamics but have difficulty transferring this knowledge to thermal images.

**Topics of Discussion:** Another way to explore people's mental models is to look at the topics that get considered for discussion. A key distinction involves differences between discussions of temperature and discussions of heat flow. Discussions of temperature were associated with lower accuracy and less use of correct theories in this study. There are two reasons why this relationship makes sense. First, to the extent that people are treating temperature as a property of objects rather than as a measure of heat energy, they are subscribing to a mental model that is not in accord with physical principles (Wiser & Carey, 1983). Second, even if they recognize that temperature is a measure of mean molecular kinetic energy, they are still focusing on an attribute of an object. Reasoning about physical principles also requires consideration of relational properties (e.g., Gentner, et al., 1997). Discussions of heat flow, in contrast, reflect a discussion of relational properties of the domain. An important aspect of heat is that it flows from high temperature regions to low temperature regions. Focusing on these relations is often useful for understanding how thermal images are in error. In many cases, errors in thermal images reflect situations in which heat is flowing in an impossible way. The relations between locations are critical for finding errors in images.

Heat flow should also generally be related to the use of a correct theory of thermodynamics. Contrary to this expectation, there was no significant relationship between discussions of heat flow in our data. This discrepancy probably reflects the same problem raised above that our coding scheme focused on theories that were both internally consistent and in accord with physical principles. It is possible that people's models are fragmentary, and thus prone to exhibit inconsistencies. Further work must address this issue.

**Communication:** Another striking aspect of the data was that dyads were significantly more accurate than were people who performed the task alone. This difference in accuracy manifested itself in a difference in performance across blocks. The dyads showed the greatest improvement in performance accuracy in the shift from the first block of three trials to the next. In contrast, singles did improve until the final block of three trials. The frequency of correct actions and theories for dyads closely followed the pattern of the performance curve. Similarly, the pattern of out-loud protocols to study problem solving strategies in a variety of tasks. One finding that emerged from this methodology is the learning benefit of self-generated explanations (Chi, 1989; 1993). Chi argues that learning requires integration of existing knowledge with new information and that the process of self-explanations

facilitates this integration. Self-explanations derived from talk out-loud protocols have been shown to improve understanding and to enhance learning (Chi, 1989, 1994; Webb, 1989). High self-explainers display deeper understanding and more complete mental models than low self-explainers as assessed by ability to answer complex questions. Chi argues that the beneficial effect of self-explanations is partly due to the fact that self-explaining is essentially a constructive activity. Self explanations provide an opportunity to construct new declarative knowledge and to generate new rules that can subsequently be used to solve complex problems. In our study, dyads are forced by the nature of the task to engage in explanatory activity from the very outset. Since the task itself is novel, there is a strong demand to integrate and adopt an existing knowledge and to construct new rules appropriate for the task at hand. To the extent that self-explanation is a constructive activity, the construction of the new knowledge structure needed to succeed on the task is started from the very onset of the task through self-explanations and explanations designed for the partner. No similar demand was placed on the singles performing the task. They were not required to verbalize their strategies, although it is interesting to note that a few subjects had spontaneously attempted to think out-loud in the course of the study, and had to be stopped by the experimenter. Thus, while the same mental process of self-explaining may be going on in the minds of the singles, there is no experimental constraint to facilitate it. This may account for the delay in improvement among singles. The learning benefit of self-explanations and explanations generated for the partner has not been explored in the context of referential communication design. We believe that it offers a potent medium to explore these issues.

### Acknowledgments

This research was supported by AFOSR grant F49620-97-1-0155 given to the second author.

### References

- Chi, M., & Koeske, R. (1983). Network representation of a child's dinosaur knowledge. *Developmental Psychology*, 19, 29-39.
- Chi, M., Hutchinson, J. E., & Robin A. F. (1989). How inferences about novel domain-related concepts can be constrained by structured knowledge. *Merrill-Palmer Quarterly*, 35(1), 27-62.
- Chi, M. & Slotta, J. (1993). The ontological coherence of intuitive physics. *Cognition & Instruction*, 10, 249-260.
- Chi, M., Slotta, J. D. & de Leeuw, N. (1994). From things to processes: A theory of conceptual change for learning science concepts. *Learning & Instruction*, 4(1), 27-43.
- Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics*, 50, 66-70.
- Cooke, N. J. & Breedin, S. D. (1994). Constructing naïve theories of motion on the fly. *Memory & Cognition*, 22(4), 474-493.
- diSessa, A.A. (1988). Knowledge in pieces. In G. Forman & P. Pufall (Eds.), *Constructivism in computer age*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- diSessa, A. A. (1993) Toward an epistemology of physics. *Cognition & Instruction*, 10(2-3), 105-225.
- Garrod, S., & Doherty, G. (1994). Conversation, coordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, 53(3), 181-215.
- Gentner, D., Brem, S., Ferguson, R., Markman, A.; Levidow, B. ; Wolff, P. Forbus, K. Analogical reasoning and conceptual change: A case study of Johannes Kepler. *Journal of the Learning Sciences*. 6(1), 1997, 3-40.
- Gopnik, A., & Meltzoff, A.N. (1997). Words, thoughts, and theories. Cambridge, MA: The MIT Press.
- Kuhn, D., (1991) *The skills of argument*. Cambridge University Press.
- Malt, B.C. (1995). Category coherence in cross-cultural perspective. *Cognitive Psychology*. 29(2), 85-148.
- McCloskey, M. (1983a). Intuitive physics. *Scientific American*, pp. 122-130.
- McCloskey, M. (1983b). Naïve theories of motion. In D. Gentner and A. L. Stevens (Eds.), *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naïve beliefs about the motion of objects. *Science*, 210, 1139-1141.
- Murphy, G. L. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92(3), 289-316.
- Ueno, N. (1993). Reconstructing P-prim theory from the viewpoint of situated cognition. *Cognition and Instruction*, 10(2-3), 239-248.
- Webb, N.M. (1989). Peer interactions and learning in small groups. *International Journal of Education Research*, 13, 21-39.
- Wiser, M., & Carey, S. (1983). When heat and temperature were one. In D. Gentner & A.L. Stevens (Eds.) *Mental models*. Hillsdale, NJ: Lawrence Erlbaum Associates.

# Was Apatosaurus a Vegan? Dinosaur Knowledge Rocks When Learning About Evolution

David R. Kaufman (davek@socrates.berkeley.edu), Michael Ranney (ranney@cogsci.berkeley.edu),  
Eric Lewis (eslewis@aol.com), and Anna Thanukos (thanukos@socrates.berkeley.edu)  
Cognition and Development, Graduate School of Education 4533 Tolman Hall  
UC Berkeley, Berkeley, CA 94720 USA

Sarah Brem (sarah.brem@asu.edu)  
Division of Psychology in Education; P.O. Box 870611  
Arizona State University; Tempe, AZ 85287

## Abstract

We present the results of an experiment involving a curriculum designed to foster conceptual changes, generative understandings, and coherent evolutionary explanations. This middle school curriculum highlights dinosaur knowledge due to its intrinsic interest to students and its compatibility with the objectives of integrating several concepts (e.g., variation and heredity) into a coherent natural selection schema. The domain also allows one to communicate an understanding of the process of evolutionary change across geologic time. Students in a class that received the curriculum exhibited significantly greater gains than did a control class, across a range of problem types. Further, the subjects in the conceptual change classroom appear to be less prone to generating the kinds of explanations that directly conflict with Darwinian patterns of reasoning.

## Introduction

Transfer is one of the most widely investigated phenomena in both cognitive science and science education. Arguably, the most important goal of education is to foster the transfer of knowledge and skills. Similarly, any learning theory worth its salt must include mechanisms of transfer. From some perspectives, studies of transfer have often yielded dismal results (Detterman, 1993). However, as one might expect, increasing any salient similarity between training and the transfer materials increases the probability that transfer will occur (Bassok & Holyoak, 1993) as does selecting a source domain in which the subjects have substantial prior knowledge (Kaufman, Patel, & Magder, 1996).

Evolution is a central unifying theory in modern biology, contributes to a foundation for learning across biological sciences, and provides a basis for understanding the interrelationships among all organisms. However, evolution remains a polarizingly controversial and poorly understood subject that typically receives a minimal amount of class time (Working Group on Teaching Evolution, 1998). In addition, evolutionary concepts are often taught as a set of discrete ideas, rather than as a central integrative topic (Sharman, 1994). Several recent studies have documented a range of misconceptions and erroneous beliefs in students' understandings of evolutionary concepts and their resistance

to instructional effects. Many individuals regard evolution as a need-driven, adaptive teleological process, whereby organisms change traits in response to some environmental pressure (Bishop & Anderson, 1990). Some see evolution as Lamarckian, in which an organism passes on to its offspring characteristics that are acquired during its lifetime (Samarapungavan & Wiers, 1997). These teleological and Lamarckian beliefs conflict with Darwinian theory of evolution by natural selection theory, an essential component of modern biological understanding.

Why is evolution such a difficult topic for students to master? In our view, the development of evolutionary competence is predicated on the following four factors (Kaufman, Ranney, Thanukos, & Brem, 1999). 1) Conceptual Knowledge: evolutionary knowledge involves the complex integration of concepts from several biological disciplines, including genetics, ecology, and paleontology. 2) Reasoning and argumentation: evolutionary reasoning makes formidable demands on the process of coordinating evidence and hypotheses; part of the problem concerns the unique nature of evolutionary explanation, which often requires reasoning about historical narratives rather than proximate (Kaufman, et al, 1998). 3) Epistemological commitments: students' views concerning the nature of science and of the biological world affect their understandings of evolution (Rudolph & Stewart, 1998). 4) Discourse practices: as in all sciences, there are ways of constructing explanations and using communication that are sanctioned or eschewed by the domain; students have considerable difficulty mastering mechanistic causal explanations and often use scientific terms, such as adaptation, inappropriately.

What are the ultimate objectives of an introductory evolutionary curriculum designed for middle school students? As with other sciences, the goals are to promote robust conceptual understanding and durable transfer. That is, we do not want students to learn merely about the evolution of dinosaurs, insects or other sets of organisms. We want students to begin to "own" Darwinian patterns of reasoning and apply them flexibly in multiple contexts. Ohlsson's (1993) notion of an abstract schema allows us to sharpen our intuitions about transfer. Such a schema encodes the structure of an explanation, rather than its content. The

following schema, adapted from Ohlsson, illustrates the notion of a Darwinian explanation pattern:

- There exists a species that varies randomly on a set of heritable characteristics.
- An environmental pressure (from imperceptible to catastrophic) will favor individuals (regarding survival) with certain traits.
- The selection mechanism operates such that these individuals are more likely to reproduce and pass on their traits to offspring.
- Therefore, more individuals in the next generation will possess the favored trait and the relative distribution of the trait will increase.
- Over many generations (i.e., hundreds or thousands), these small changes in traits accumulate and may eventually substantially modify the characteristics of the species.

This natural selection schema is potentially applicable to any organism and can be articulated by instantiating the appropriate slots (e.g., favored traits, and environmental pressure). The schema embodies both core conceptual knowledge and the relational argument structure that constitutes natural selection. Mastery of this explanation form across several domains would constitute strong evidence for transfer. Of course, a mere syntactic mapping is not all that is required; the use of this abstract schema requires substantial biological knowledge and development of the aforementioned factors that comprise evolutionary competency.

We sought to develop a curriculum that (a) specifically targets conceptual knowledge and reasoning/argumentation, and (b) engages students' prior knowledge in a domain of student interest. Chi and colleagues (Gobbo & Chi, 1986; Chi, Hutchinson, & Robbins, 1989) demonstrated that young children have substantial dinosaur knowledge and can employ this knowledge to make inferences about the organisms' diets, habitats, and locomotion. In addition, many middle school students have a basic mastery of the concepts required to learn natural selection (e.g., inheritance, biodiversity, variation, and prey/predation), but they lack an organizing schema for understanding evolution (Ash & Brown, 1996). It was hypothesized that knowledge of dinosaurs would represent a generative source domain in order to impart a robust understanding of evolution. Further, the study of dinosaurs exemplifies the historical/narrative dimension of evolutionary reasoning and the process of evidence gathering from the fossil record. In general, greater subject matter knowledge increases the likelihood of transfer since both entities (i.e., dinosaurs) and ecological processes (e.g., predator/prey relations) are familiar in this context. There is less need for negotiating new terminology and other unfamiliar surface features. This domain may also serve to foster epistemological commitments regarding the transitional state of knowledge, since new fossil finds and concomitant hypotheses are regularly brought to the public's attention.

## Method

### Participants

Two seventh grade classes from an urban, ethnically diverse, public school participated. The experimental and control classes included 21 and 27 students, respectively.

### Procedure

**Pretest:** In the first of the study's three phases, each class was given a dinosaur knowledge test followed by two evolutionary knowledge tests. The dinosaur test consisted of 39 questions that evaluated students' abilities to identify dinosaurs from pictures, draw inferences about the dinosaurs' diets from both pictures and dinosaur names, match dinosaur names to descriptions of dinosaurs, order events on a timeline, and, respond to Likert items about dinosaurs. The evolution tests assessed students' understandings of related concepts. The two tests respectively consisted of seven Likert items, followed by eight short essay questions (/problems). The evolution tests assessed students' knowledge of heredity, variation, selection pressure, survival advantage, and mutation. The test questions involved a variety of animal contexts including birds, humans, and dinosaurs. There were three types of essay questions, involving natural selection, conditions of adaptational change, and common ancestry. An example of an essay problem addressing natural selection is "Apatosaurus was a dinosaur that had a long neck (longer than modern giraffes). The ancestors of all Apatosaurs had short necks (similar to necks of horses). Please explain how Apatosaurs came to have long necks." The conditions of adaptational change questions can best be answered by discussing functional adaptation and the time scale for such adaptations to appear. An example of this type of problem is: "If there were a sudden drought that killed off most edible plants, could a cow start to eat other animals instead of plants? Explain why or why not." The common ancestry questions addressed the salient similarity and differences among ancestral and contemporary species. The following problem is an example: Ostriches are large birds that cannot fly. The Rhea and the Emu are in the same family of birds – they are very closely related genetically. Interestingly, Ostriches are found in Africa, the Rhea live in South America, and the Emu live in Australia. How can you explain that these birds, which cannot fly, are found on different continents? "

**Instruction:** In the next phase of the study, the two classes participated in divergent eight-day evolution units. In the experimental class, hands-on activities, illustrations, and lectures were constructed to illustrate scientists' conceptions of dinosaur life. (Each lesson included at least one hands-on activity, an interactive discussion, and independent thinking assignments.) The curriculum addressed a range of topics: heredity, variation in the environment, mutation, extinction, and variation among individuals in a population. Explicit examples were provided to model how students could transfer evolution concepts to other animal species. This curriculum was created and taught by the teacher-researcher, a Masters

student who had prior experience teaching the students in this classroom (Lewis, 1999). This instructor used a constructivist pedagogy, largely modeled after Minstrell's instructional approach (van Zee & Minstrell, 1997). Minstrell is noted for introducing a new topic with a "benchmark lesson." He attempts to discover what students know about a topic, and tries to evaluate which of the different facets of the larger concept are understood or misunderstood.

The control class was taught by its regular teacher. He had over 25 years of teaching experience and taught in a traditional didactic manner while relying on the textbook. Students were responsible for taking accurate notes and answering the questions that appeared in the text. The control class drew on a range of organisms to illustrate the process of how life changes over time and evidence for these changes.

**Posttest:** The final phase of the study measured what students learned by again assessing dinosaur and evolution knowledge. The dinosaur test was essentially identical to the original test except for the order of questions. The evolution posttest used analogous (structurally isomorphic) and questions that were identical to those on the pretest. The evolution posttest was thus designed to assess the students' basic learning and their ability to transfer evolutionary knowledge to novel contexts.

**Analyses:** The Likert questions were initially scored on a seven point scale, based on the "correctness" of answers, and then scaled to fractions of a single point. The essay questions were scored and weighted for difficulty, according to a modified version of a rubric created by Kaufman, et al (1999). The coding criteria are similar to those used by Ferrari and Chi (1997) and Ohlsson (1990). For example, on the natural selection questions, explanations were coded for clear expressions (i.e., not merely jargon usage) of 1) variation, 2) selection pressure (environmental contingencies), 3) survival advantage (adaptive characteristics) and 4) heredity. A subset of the 16 questions was rescored by a second reader, resulting in an interrater reliability of 94%.

## Results

The results indicate that both control and experimental classes exhibited various gains. A multivariate repeated measures analysis of variance was performed, with the three tests serving as the dependant variables and class as the independent variable. The analysis revealed a main effect for class ( $F(1,46)=6.24$ ,  $p<.05$ ) with the experimental class performing better than the control class. There was also a significant temporal effect, indicating that subjects performed better on the posttests ( $F(1,46)=100.79$ ,  $p<.001$ ). In addition, there was a significant time by class interaction ( $F(1,46)=19.46$ ,  $p<.001$ ) with the experimental class exhibiting a larger gain.

The overall results, presented in Table 1, reveal that both classes had considerable and comparable prior dinosaur knowledge, averaging 70% ( $F(1,46)=1.19$ , n.s.) on the

dinosaur pretest. The evolutionary knowledge pretest indicated even more similarity between the experimental and control classes, averaging 67% ( $F(1,46)=.28$ , n.s.) over the Likert questions and 23% ( $F(1,46)=.002$ , n.s.) over the essay questions. Both classes improved on the evolution posttests, averaging 77% of the Likert questions' points ( $F(1,46)=37.26$ ,  $p<.001$ ) and 39% of the essay questions' points ( $F(1,46)=67.70$ ,  $p<.001$ )

Not surprisingly, the experimental class demonstrated a significantly greater increase on the dinosaur posttest than the control group ( $F(1,46)=9.660$ ,  $p<.005$ ). More importantly, the experimental class showed a greater gain on both the Likert ( $F(1,46)=7.60$ ;  $p<.01$ ) and the essay ( $F(1,46)=6.43$ ;  $p<.02$ ) evolution tests. In concert with the view of dinosaur knowledge as an anchor for learning, an exploratory regression analysis to determine the predictors of the evolution essay posttest showed the dinosaur pretest to be the best predictor, accounting for over 30% of the variance.

Further exploration (Table 2) of the three essay questions that most involved natural selection reveal very modest pretest performance; the mean score for the experimental class was 12% (SD = 9%), while that for the control class was 16% (SD = 10%). However, during the posttest, the mean for the experimental class grew to 35% (SD = 16%), while the mean for the control class was 22% (SD = 19%). Table 2 also illustrates the breakdown of these three essay responses into the four natural selection criteria. The results indicate that students generated responses that accounted for selection pressures 36% of the time, whereas students only discussed the role of heredity in natural selection 8% of the time. The experimental class demonstrated several notable gains regarding the criteria (particularly, mutation/variation, which grew from 11% to 54%), whereas the gains of the control class were generally more modest. Consider the following student responses, regarding why apatosaurs/giraffes had longer necks than their ancestors:

### A7 Pretest

They need to reach the food at the top of trees and they evolved with longer and longer necks.

### A7 Posttest

"There was a random mutation and one baby had a long neck some of its baby will have long neck too. Then soemthing in there environment or surrounds change i.e. Food is higher in the trees making it good to have a long neck because food is harder to get the ones with the short necks die leving only ones with long necks. They mate and then there are more long neck and this keeps happing."

### A13 Pretest

I think that apatosaurs came to have longer necks by evolution. Over time, they got bigger and bigger.

### A13 Posttest

RANDOM MUTATION! Giraffes may have had offspring that, purely by luck, had long necks. Maybe food on the floor of the forest was

diminishing and the long necked giraffes got food from high up. The short neck giraffes probably died of starvation. Then when only long necks were left, long necks had to reproduce. If longnecks mated they'd produce other long necks, until present day giraffes were known for their long necks.

The pretest responses often invoked the notion of "need" with no real sense of mechanisms. On the posttest, the subjects expressed more sophisticated understandings of

evolutionary concepts and at least rudimentary mastery of the appropriate form of a natural selection explanation. In spite of the differential learning successes exhibited by the experimental class, their explanations were still rather modest or inconsistent, as evidenced by their evolution posttest essays and their natural selection question responses. These results are consistent with other studies (e.g., Ohlsson, 1990; Bishop & Anderson, 1990) that documented persistent difficulties in students' (from middle school to college) reasoning about natural selection.

**Table 1.** Mean percentages and standard deviations (parentheses) for all tests and classes.

	Dinosaur Pretest	Dinosaur Posttest	Evolution Pretest Likert	Evolution Posttest Likert	Evolution Pretest Essay	Evolution Posttest Essay
Exp. n=21	73 (17)	85 (14)	68 (8)	84 (10)	23 (12)	45 (18)
Cont. n=27	68 (17)	70 (17)	66 (12)	72 (14)	23 (11)	35 (16)
<i>Total</i>	<i>70 (16)</i>	<i>76 (16)</i>	<i>67 (11)</i>	<i>77 (13)</i>	<i>23 (11)</i>	<i>39 (17)</i>

**Table 2.** Percentages of natural selection essay responses with respect to aspects of the coding criteria.

	<b>Experimental</b>		<b>Control</b>		<b>Totals</b>
	Pretest	Posttest	Pretest	Posttest	
M/V	11	54	20	15	25
SP	25	41	28	48	36
SA	6	27	12	20	16
HE	6	19	4	4	8
<i>Mean</i>	<i>12 (9)</i>	<i>35 (16)</i>	<i>16 (10)</i>	<i>22 (19)</i>	<i>21(12)</i>

*Code: M/V: Mutation/Variation, SP: Selection Pressure, SA: Survival Advantage, HE: Heredity.*

**Table 3.** Percentages of some non-Darwinian essay responses.

	<b>Experimental</b>		<b>Control</b>	
	Pretest	Posttest	Pretest	Posttest
Lamarckian	4	3	4	7
Teleological	10	7	10	18
Amechanistic	24	14	21	24
<i>Total</i>	<i>38</i>	<i>24</i>	<i>35</i>	<i>49</i>

Students often exhibit patterns of reasoning that are inconsistent with Darwinian explanations. When possible, non-Darwinian response patterns were classified as Lamarckian, Teleological, or Amechanistic (absence of mechanism), as shown in Table 3. Note that there were many responses that were not fully consonant with a Darwinian explanation, yet were not classifiable according to

this coding scheme. Lamarckian explanations implying the passing of acquired traits to progeny. For example, a student from the experimental class explained, "I think that maybe the cheetahs hunted animals that started to get fast and run away. The cheetahs had to adapt and run faster to catch their food. As their prey began to lead them on chases, their speed increased. Over time, their muscles probably just got bigger

and stronger (because they worked them so much). Now, cheetahs run very fast, and can catch gazelles and impalas and zebras and antelope." Teleological explanations suggest that need causes evolutionary change. A control class student envisioned changes in the eating habits of a cow if there were suddenly no grass to eat: "I think that they would just get so hungry they would start with insects and move their way up to fish." Another student, during the control class's posttest, explained that giraffe necks are so long "Because the giraffes had to stretch their necks to reach the trees for food." A mechanistic explanation indicates that evolution simply happens. In explaining why cheetahs became faster, one control class student stated "they had adapted to the prey getting faster. Through evolution." While some non-Darwinian explanations merely reflect an inability to express ideas maturely or the absence of specific biological knowledge, they may also indicate an inability to construct reasonable Darwinian arguments with their existing knowledge of evolution.

Table 3 also illustrates ways in which the two classes, while performing rather similarly on the pretest, differ when it comes to posttest response patterns. The experimental class's non-Darwinian responding went from 38% to 24%, while the control class's non-Darwinian responding moved in the opposite direction, from 35% to 49%. One of the central goals of the experimental curriculum was to foster effective Darwinian reasoning, and the results suggest modest success in that regard. Further, the results suggest that instruction may even foster more problematic patterns of reasoning.

## Discussion

Recently, innovative curricula have targeted different facets of student difficulty regarding understanding evolution. The present study suggests that a conceptual change evolution curriculum anchored in the domain of dinosaur knowledge can promote the integration of core concepts and foster more effective Darwinian reasoning. Dinosaurs seem to be a good choice as an anchor for a contextualized curriculum. Dinosaur knowledge has been established to be relatively high among middle school students, and the results of this study suggest that having dinosaur knowledge may provide students with an advantage in learning about evolution.

Although the results of this curriculum are promising, the gains are still modest. Further research is needed to exclude the possibility that the differences between groups are not the result of extraneous factors. For example, the gains may be explained by the novelty of dinosaurs, the experimental teacher's enthusiasm, or the relative advantage of constructivist teaching methods over conventional didactic instruction. Nevertheless, this study suggests that employing an intrinsically motivating curricular source domain that engages a student's prior knowledge can facilitate the development of evolutionary competence. The dinosaur curriculum was designed to foster generative conceptual knowledge and coherent evolutionary reasoning. The other two pieces of the evolutionary competence puzzle,

epistemological commitments and discourse practices, were less central in the curriculum. These core features of evolutionary competence are clearly interdependent. For example, a student who appreciates the "correct form" of a natural selection explanation, but lacks a suitable descriptive vocabulary (i.e., one who can't "talk the talk") is unlikely to generate a coherent explanation. The standards of coherence for evolutionary explanations are particularly exacting, and coherence-building interventions are worthy pursuits in fostering critical thinking (Ranney & Schank, 1998).

Teleological reasoning was noted in many of the students' explanations. Teleological causation in explanations is hardly unique to evolution. It may underlie intuitive theories of biology in children as well as adults (Carey, 1995). Biological processes can be thought about in mechanistic or teleological terms. While it is advantageous for students to have a principled, mechanistic, understanding of scientific concepts, teleological or goal-oriented explanations are often presented in textbooks and lectures to orient students to the functions of a particular bodily mechanism. Teleological explanations are also commonplace in everyday discourse. Considerable research indicates that young children develop rudimentary theories in which biological functions are often expressed in intentionalist terms, such as striving to fulfill "wants" and "needs" (Hatano & Inagaki, 1996; Carey, 1995). This phenomenon is not unique to children. Medical students sometimes generate teleological explanations in reasoning about the function of the heart and circulatory system (Kaufman, Patel, & Magder, 1996).

Teleological thought is rooted in productive forms of knowledge and provide coherent explanations for nonintuitive phenomena that surrounds us. It is a challenge is to effectively exploit this knowledge in formal education in order to develop mechanistic understandings of biological processes. For example, teleological reasoning may promote an understanding of structure-function relations in young children; Ash and Brown (1996) developed a curriculum that fosters transitions from more rudimentary forms of teleological thought towards adaptationist reasoning that approximates mature natural selection explanations.

Our results suggest that some students began to demonstrate proficient Darwinian explanation patterns. However, most students continued to experience difficulties incorporating notions of variation and heredity into their responses—and subjects were somewhat inconsistent across problems. Anchoring in a given domain represents a starting point, but experience applying the schema in different domains is likely a prerequisite to mastery.

Learning about evolution in a familiar domain can certainly facilitate the development of disciplinary discourse, though more needs to be done to foster proficient "evolution speak". Tabak and Reiser (1999), using BGUILLE (the Biology Guided Inquiry Learning Environment) and working with middle school teachers, also try to advance productive discourse strategies in learning about natural selection, striving to scaffold students so that they can progress from lay explanations to increasingly sophisticated scientific explanations. The process involved establishing scientific norms, providing specific prompts (e.g., to elaborate incomplete explanations) and reshaping response patterns in



a manner that approximates scientific discourse. Fostering effective disciplinary discourse practices is essential in the development of evolutionary competence.

The concept of biological evolution represents a critical challenge for students to master, and given that it is a foundational concept in the biological sciences, it warrants special attention. A growing body of empirical work has systematically diagnosed a range of student difficulties pertaining to evolution, and researchers will certainly continue to develop ever more promising instructional strategies to support coherent evolutionary reasoning and argumentation.

### Acknowledgments

We thank the participants in the Evolution and Biology Education research groups. We are especially grateful to Anli Liu, Jennifer Schindel, and Franz Cheng for their comments and assistance in the data analyses.

### References

- Ash, D. & Brown, A. (1996). *Thematic Communities guide shifts in biological reasoning: children's transition towards deep principles of evolution*. Paper presented at the Annual Meeting of the Educational Research Association, New York, April 1996.
- Bassok, M. & Holyoak, K.J. (1993). Pragmatic knowledge and conceptual structure: Determinants of transfer between quantitative domains. In: D. K. Detterman and R. J. Sternberg (eds.), *Transfer on trial: Intelligence, cognition, and instruction..* Norwood, NJ: Ablex.
- Bishop, B. & C. Anderson (1990). Student Conceptions of Natural Selection and its role in Evolution. *Journal of Research in Science Teaching*, 27, 415-427.
- Carey, S. (1995). On the origin of causal understanding. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate*. (pp. 268-308). Oxford: Clarendon Press.
- Chi, M., Hutchinson, J., & Robin, A. (1989). How Inferences About Novel Domain-Related Concepts Can be Constrained by Structured Knowledge. *Merrill-Palmer Quarterly*, 35, 27-62.
- Detterman, D.K. (1993). The case for the prosecution: Transfer as an epiphenomenon. In D.K. Detterman and R.J. Sternberg (eds.), *Transfer on trial: Intelligence, cognition, and instruction*. Norwood, NJ: Ablex.
- Ferrari, M & Chi, M.T.H. (1998). The nature of naive explanations of natural selection. *International Journal of Science Education*, 20 1231-1256.
- Gobbo, C. & Chi, M. (1986). How knowledge is structured and used by expert and novice children. *Cognitive Development*, 1, 221-237.
- Hatano, G & Inagaki, K. (1996). Cognitive and cultural factors in the acquisition of intuitive biology. In D. R. Olson & Torrance, N. (eds.) *The Handbook of Education and Human Development: New models of learning, teaching, and schooling*. New York: Blackwell.
- Kaufman, D.R. Patel, V.L., & Magder, S (1996). The Explanatory Role of Spontaneously Generated Analogies in a Reasoning about Physiological Concepts. *International Journal of Science Education*, 18, 369-386.
- Kaufman, D.R., Ranney, M, Ohlsson, S. Reiser, B.J., & Shapiro, L. (1998). Symposium: Multidisciplinary Perspectives on Evolutionary Reasoning. Presented at the Twentieth Annual Meeting of the Cognitive Science Society. Madison, WC. August 1-4, 1998.
- Kaufman, D.R., Ranney, M, Thanukos, A., & Brem, S. (1999). How to Spin an Evolutionary Tale. Presented at the Twenty-First Annual Meeting of the Cognitive Science Society. Vancouver, B.C. August 19-22, 1999.
- Lewis, E. (1999). *Dinosaur Knowledge as a Fruitful Anchor for Learning Evolution Proto-Concepts*. Masters Thesis, Cognition and Development, Graduate School of Education. University of California, Berkeley. Berkeley, CA.
- Ohlsson, S. (1990). *Young Adults' Understanding of Evolutionary Explanations: Preliminary Observations*. (Office of Educational Research and Improvement Rep.). Pittsburgh: University of Pittsburgh, The Learning Research and Development Center.
- Ohlsson, S. (1993). Abstract schemas. *Educational Psychologist*, 28, 51-66.
- Ranney, M., & Schank, P. (1998). Toward an integration of the social and the scientific: Observing, modeling, and promoting the explanatory coherence of reasoning. In S. Read & L. Miller (Eds.), *Connectionist models of social reasoning and social behavior*. Mahwah, NJ: Lawrence Erlbaum
- Reiser, B. J., Sandoval, W. A., & Tabak, I. (1998). *Teachers' support of students' biological inquiry: Making use of artifacts of students' reasoning*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.
- Rudolph, J. & Stewart, J. (1998). Evolution and the Nature of Science: On the Historical Discord and Its Implications for Education. *Journal of Research in Science Teaching*, 35 (10), 1069-1089.
- Samarapungavan, A. & Wiers, R. (1997). Children's Thoughts on the Origin of Species: A Study of Explanatory Coherence. *Cognitive Science*, 21 (2), 147-177.
- Shankar, G., & Skoog, G. D. (1993). Emphasis given evolution and creationism by Texas high school biology teachers. *Science Education*, 77, 221-233.
- Sharman, L. C. (1994). Teaching evolution: Designing successful instruction. *Journal of Science Teacher Education*, 5, 122-129.
- Tabak, I. & Reiser, B.J. (1999). *Steering the course of dialogue in inquiry-based science classrooms*. Paper presented at the 1999 Annual Meeting of the American Educational Research Association. Montreal, Canada.
- van Zee, E. & Minstrell, J. (1997). Using Questioning to Guide Student Thinking. *The Journal of the Learning Sciences*, 6, 227-269.
- Working Group on Teaching Evolution, National Academy of Science (1998). *Teaching about evolution and the nature of science*. Washington, DC: National Academy Press.

# Evaluating Competition-based Models of Word Order

Frank Keller

keller@cogsci.ed.ac.uk

Institute for Communicating and Collaborative Systems

Division of Informatics, University of Edinburgh

2 Buccleuch Place, Edinburgh EH8 9LW, UK

## Abstract

The ordering of constituents in semi-free word order languages has attracted considerable attention in theoretical linguistics. Three types of models have been proposed to explain word order preferences, based on (a) weighted constraints, (b) Optimality Theory (c) syntactic weight. All three models use grammatical competition to explain the interaction of word order constraints. They rely on intuitive judgments or corpus studies, but have not been evaluated against experimental data. This is the purpose of the present paper. We report the results of a magnitude estimation experiment investigating word order in German, focusing on the interaction of verb position, case marking, pronominalization, and information structure. The experimental data are compatible with models (a) and (b), indicating that relativized (ranked or weighted) constraints are essential in explaining word order preferences. Model (c), on the other hand, is not compatible with the data.

## Introduction

The languages of the world differ substantially in the degree of word order variation they allow. On the one end of the spectrum, we find languages like English, which exhibit a relatively fixed word order. On the other end, there are languages like Warlpiri (an Australian language), which allow a large degree of word order variation. Many languages exhibit a semi-free word order, i.e., the word order is fixed in some respects, but variable in others.

Word order variation typically manifests itself not in binary acceptability judgments, but in the form of word order preferences, to which a diverse set of factors contribute, including syntactic, pragmatic, and phonological factors. This poses an interesting challenge for linguistic theory, which is equipped to deal with binary ungrammaticality resulting from the violation of individual linguistic constraints, rather than with preferences (degrees of acceptability) that emerge from the interaction of a diverse set of factors.

A number of approaches have been developed to deal with this challenge, all of which diverge from conventional linguistic frameworks in assuming a relative (weighted or ranked) rather than an absolute (binary) notion of linguistic constraints. Three main types of models have been proposed, based on (a) weighted constraints (Jacobs, 1988; Uszkoreit, 1987), (b) Optimality Theory (Choi, 1996; Müller, 1999), (c) syntactic weight (Hawkins, 1992). All of these models use a notion of grammatical competition to explain the interaction of the factors that influence word order.

Models (a)–(c) rely on informal, intuitive acceptability judgments (and on corpus data in the case of (c)). It is safe

to assume that such judgments allow to determine binary acceptability reliably. However, their reliability is much less obvious with respect to degrees of acceptability like the ones that occur in word order data (Bard, Robertson, & Sorace, 1996; Cowart, 1997; Schütze, 1996). This makes it desirable to evaluate linguistic models of word order against experimentally collected acceptability data.

The purpose of the present paper is to provide a first step towards such an experimental evaluation.<sup>1</sup> The methodology we use is magnitude estimation, which has been shown to yield reliable, yet maximally delicate judgments of linguistic acceptability (Bard et al., 1996). The empirical domain of our investigation is the variation in the order of verb complements in German, a semi-free word order language. We outline the necessary linguistic background in the following section.

## Word Order in German

German has a fixed verb order. Subordinate clauses are verb final, while yes/no questions require verb initial order, and declarative main clauses have the verb in second position. In the generative literature, the subordinate clause order is generally considered the basic order from which the main clause and question orders are derived by movement (e.g., Haider, 1993). The present experiment will focus on subordinate clauses (which is also customary in the processing literature on German, e.g., Bader & Meng, 1999). Using subordinate clauses avoids potential confounds from topicalization and other phenomena that can occur in verb second clauses.

While verb order is fixed in German, the order of the complements of the verb is variable, and a number of factors have been claimed to influence complement order. These factors include case marking, thematic roles, pronominalization, information structure, intonation, definiteness, and animacy (Choi, 1996; Jacobs, 1988; Müller, 1999; Uszkoreit, 1987; Scheepers, 1997). The present study focuses on the effect of case marking, pronominalization, and information structure on word order.

We elicited acceptability judgments for four subordinate clause orders, illustrated by the examples in (1). As mentioned above, subordinate clauses in German require verb final order (see (1a), (1b)). Verb initial orders (see (1c), (1d)) give rise to strong unacceptability.

---

<sup>1</sup>Previous experimental work on word order preferences in German (Pechmann, Uszkoreit, Engelkamp, & Zerbst, 1994; Scheepers, 1997) only dealt with isolated stimuli, i.e., failed to address contextual effects on order, one of the topics of the present paper

- (1) a. **SOV:**  
 Maria glaubt, dass der Vater den Wagen kauft.  
 M.-NOM believes that the father-NOM the car-ACC buys  
 ‘Maria believes that the father will buy the car.’  
 b. **OSV:** Maria glaubt, dass den Wagen der Vater kauft.  
 c. **VSO:** Maria glaubt, dass kauft der Vater den Wagen.  
 d. **VOS:** Maria glaubt, dass kauft den Wagen der Vater.

We also examined the influence of pronominalization on word order. The experiment included sentences where none of the NPs was pronominalized (see (1)), but also sentences where the subject, object, or both the subject and the object were pronominalized (see (2)).<sup>2</sup>

- (2) a. Maria glaubt, dass er den Wagen kauft.  
 Maria-NOM believes that he-NOM the car-ACC buys  
 ‘Maria believes that he will buy the car.’  
 b. Maria glaubt, dass der Vater ihn kauft.  
 Maria-NOM believes that the father-NOM it-ACC buys  
 ‘Maria believes that the father will buy it.’  
 c. Maria glaubt, dass er ihn kauft.  
 Maria-NOM believes that he-NOM it-ACC buys  
 ‘Maria believes that he will buy it.’

Information structure figures as a determinant of complement order in the accounts of Choi (1996), Jacobs (1988), Müller (1999), and Uszkoreit (1987). Information structural effects can be studied by embedding the sentence in a question context: the *wh*-phrase marks the focussed constituent, while the other constituents are non-focussed, or ground (Vallduví, 1992). The following contexts were used in the experiment:

- (3) a. **All Focus:** Was gibt’s neues?  
 ‘What’s new?’  
 b. **S Focus:** Wer kauft den Wagen?  
 ‘Who will buy the car?’  
 c. **O Focus:** Was kauft der Vater?  
 ‘What will the father buy?’

A null context condition was included as a control, allowing us to study how subjects react in the absence of any contextual information.

## Models of Word Order

### Weighted Constraints

Uszkoreit (1987) models word order preferences using weighted constraints. In such a setting, linguistic constraints are annotated with a numeric weight that reflects their importance in determining grammaticality (for a similar proposal, see Jacobs, 1988). Uszkoreit assumes constraint competition, i.e., not all constraints are necessarily satisfiable in a given linguistic structure. This entails that grammaticality is a gradient notion; the degree of grammaticality of a linguistic structure is computed as the sum of the weights of the constraint violations the structure incurs.

Uszkoreit (1987, p. 114) proposes the following constraints on word order in German (constraints irrelevant to the data under consideration are omitted and constraint names are provided):

<sup>2</sup>Note that only masculine NPs were used, as these are unambiguous in their case marking, both as full NPs and as pronouns (while the case morphology of feminine and neuter NPs exhibits syncretism).

- (4) a. VERB:  $X \prec V[-MC]$   
 b. NOM:  $[+NOM] \prec [+ACC]$   
 c. FOC:  $[-FOCUS] \prec [+FOCUS]$   
 d. PRO:  $[+PRO] \prec [-PRO]$

These constraints are constituent order constraints, with ‘ $\prec$ ’ denoting linear precedence. The constraint VERB relies on the feature MC (main clause) to specify verb order; if this feature is negative (i.e., in a subordinate clause), then the verb has to succeed any other constituent. The constraint NOM requires that nominative precedes accusative. The information structural requirement FOC specifies that ground constituents (marked  $[-FOCUS]$ ) precede focused constituents. The constraint PRO requires pronouns to precede full NPs.

Uszkoreit does not provide weights for the constraints in (4).<sup>3</sup> Intuitively, however, we expect a violation of VERB to lead to serious unacceptability, i.e., VERB should receive a higher weight than the other constraints.

### Optimality Theory

Standard Optimality Theory (OT; Prince & Smolensky, 1993) assumes a binary notion of grammaticality; a linguistic structure is either optimal (and thus grammatical) or suboptimal (and thus ungrammatical). However, OT can be extended to model gradient grammaticality; Müller (1999) puts forward a modified version of OT based on the distinction between grammaticality (manifested in binary judgments) and markedness (associated with word order preferences). Grammaticality is handled in terms of conventional OT-style constraint competition. This competition can yield several grammatical candidates, among which further competition takes place based on markedness constraints. The markedness competition then induces a preference order on the candidates that predicts their relative acceptability. (Note that the grammaticality/markedness dichotomy is reminiscent of the distinction of hard and soft constraints proposed by Keller (1998).)

In Müller’s account, the constraints on pronoun order belong to the realm of grammaticality, while the constraints on case order and focus-ground order (among others) belong to the realm of markedness. We omit technical details and only state constraints relevant to the present data set:

- (5) a. NOM:  $[+NOM] \prec [-NOM]$   
 b. FOC:  $[-FOCUS] \prec [+FOCUS]$   
 c. AN:  $[+ANIMATE] \prec [-ANIMATE]$

Note that the constraints NOM and FOC are similar to Uszkoreit’s constraints in (4). AN is an additional constraint that requires animate NPs to precede inanimate ones. In contrast to Uszkoreit, Müller postulates an explicit constraint ranking:

- (6)  $NOM \gg AN \gg FOC$

In addition to the markedness constraints in (5), a set of grammaticality constraints is postulated (omitted here). These constraints deal with pronoun order and ensure that pronouns occur at the left periphery of the clause. All candidates that fail to meet this requirement are predicted to be (categorically) ungrammatical. In contrast to Uszkoreit, Müller does not include constraints on verb order.

<sup>3</sup>Pechmann et al. (1994) tentatively assume that all constraints have equal weights, which entails that the degree of unacceptability only depends on the number of violations.

## Syntactic Weight

Hawkins (1992) proposes an approach to word order preferences that also relies on grammatical competition, but makes very different assumptions concerning the source of this competition. Hawkins assumes that constituent order is determined by the syntactic weight of the constituents, a notion that is supposed to reflect how easily the constituents can be recognized by the human parser. According to Hawkins (1992, p. 200), relative syntactic weight explains word orders frequencies in corpora, as well as the relative acceptability of different orders in native speaker's judgments.

Hawkins proposes Immediate Constituent to Word Ratio (ICR) as a metric for syntactic weight. Intuitively, ICR measures the length of a constituent relative to its position in the clause (see Hawkins, 1992, for details). If two sentences differ in average ICR, the one with the higher average ICR is predicted to be more acceptable. The ICR for a given word is calculated as  $n/m$ , where  $n$  is the number of the constituent, while  $m$  is the number the word, counted from left to right. The average ICR for a sentence is obtained by averaging the ICRs of its words. As an example, consider the ICRs for the sentences in (1):

- (7) a. M. glaubt, dass [[der Vater] [den Wagen] kauft.] ICR  
1/1 1/2 2/3 2/4 3/5 .65  
b. M. glaubt, dass [kauft [der Vater] [den Wagen].] ICR  
1/1 2/2 2/3 3/4 .86

Provided that subject and object have the same length, SO and OS orders receive the same ICR, i.e., examples (1a) and (1b) both have an ICR of .65 (see (7a)), while examples (1c) and (1d) both receive an ICR of .86 (see (7b)). For pronominalized NPs, the following ICRs are predicted:

- (8) a. Maria glaubt, dass [[er] [den Wagen] kauft.] ICR  
1/1 2/2 2/3 3/4 .86  
b. Maria glaubt, dass [[der Vater] [ihn] kauft.] ICR  
1/1 1/2 2/3 3/4 .73  
c. Maria glaubt, dass [[er] [ihn] kauft.] ICR  
1/1 2/2 3/3 1.0

This means that Hawkins's account predicts that pronouns have to precede full NPs (if they are longer than a single word). However, if both the subject and the object are pronouns, then both SO and OS receive an ICR of 1.0, i.e., they should be equally acceptable.

Note that Hawkins predicts that information structure (focus and ground) should not play a role in determining word order preferences, contrary to claims by Müller (1999) and Uszkoreit (1987), among others.

## Experiment

### Method

**Subjects** Fifty-one native speakers of German participated in the experiment. All participants were naive to syntactic theory.

**Materials** A factorial design was used that crossed the factors verb order (*Vord*), complement order (*Cord*), pronominalization (*Pro*), and context (*Con*). The factor *Con* had four levels: null context, all focus, S focus, and O focus, as illustrated in (3). The factor *Vord* had four two levels: verb final

(see (1a), (1b)) and verb initial (see (1c), (1d)). The two levels of *Cord* were subject before object and object before subject, as in (1a), (1c) and (1b), (1d). In the null context condition, the factor *Pro* had four levels, viz., both S and O full NPs, S pronoun and O full NP, S full NP and O pronoun, and both S and O pronouns (see (2)). In the context condition, *Pro* only had two levels, viz., no pronoun and pronoun. In the all focus and S focus contexts, the object was pronominalized, while in the O focus context, the subject was pronominalized. This design ensures that the pronoun is interpreted as ground and hence is unstressed (as the sentential stress has to fall on the focussed constituent). We are only interested in the syntactic behavior of weak (i.e., unstressed) pronouns; strong (i.e., stressed) pronouns are subject to different syntactic constraints.

This yielded a total of  $Vord \times Cord \times Pro = 2 \times 2 \times 4 = 16$  cells for the null context condition, and  $Vord \times Cord \times Pro \times Con = 2 \times 2 \times 2 \times 3 = 24$  cells for the context condition. Eight lexicalizations per cell were used, which resulted in a total of 320 stimuli. A set of 24 fillers was used in the null context condition; 16 fillers were employed in the context condition. The fillers were designed to cover the whole acceptability range.

To control for possible effects from lexical frequency, the lexicalizations for subject, object, and verb were matched for frequency. Frequency counts for the verbs and the head nouns were obtained from a lemmatized version of the Frankfurter Rundschau corpus (40 million words of newspaper text) and the average frequencies were computed for subject, object, and verb lexicalizations. An ANOVA confirmed that these average frequencies were not significantly different from each other.

**Procedure** The method used was magnitude estimation as proposed by Stevens (1975) for psychophysics and extended to linguistic stimuli by Bard et al. (1996) and Cowart (1997).

Subjects first saw a set of instructions that explained the concept of numerical magnitude estimation using line length. Subjects were instructed to make length estimates relative to the first line they would see, the reference line. They were told to give the reference line an arbitrary number, and then assign a number to each following line so that it represented how long the line was in proportion to the reference line. Several example lines and corresponding numerical estimates were provided to illustrate the concept of proportionality. Then subjects were told that linguistic acceptability could be judged in the same way as line length. The concept of linguistic acceptability was not defined, but examples of acceptable and unacceptable sentences were provided.

The experiment started with a training phase designed to familiarize subjects with the magnitude estimation task. Subjects had to estimate the length of a set of lines. Then, a set of practice items (similar to the experimental items) were administered to familiarize subjects with applying magnitude estimation to linguistic stimuli. Finally, subjects had to judge the experimental items. A between subjects design was used to administer the factor *CON*: subjects in Group A judged non-contextualized stimuli, while subjects in Group B judged contextualized stimuli. The factors *Vord*, *Cord*, and *Pro* were administered within subjects. Using a Latin square design, eight lexicalizations were created for each group. The lexi-

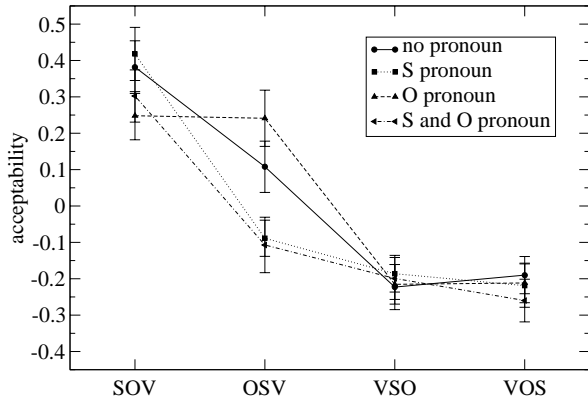


Figure 1: Interaction for word order and pronominalization, null context

calizations for Group A contained 16 items, while the ones for Group B contained 24 items.

Each subject saw one lexicalization and 24 fillers in Group A or one lexicalization and 16 fillers in Group B, i.e., a total of 40 items per group. Each subject was randomly assigned to a group and a lexicalization: 20 subjects were assigned to Group A, and 31 to Group B. Instructions, examples, training items, and fillers were adapted for Group B to take context into account.

## Results

The data were normalized by dividing each numerical judgment by the modulus value that the subject had assigned to the reference sentence. This operation creates a common scale for all subjects. All analyses were carried out on the geometric means of the normalized judgments, as is standard for magnitude estimation data (Bard et al., 1996; Cowart, 1997).

In discussing the results, we make use of the following abbreviations: SO for subject before object, OS for object before subject, XV for verb final, VX for verb initial. The indices ‘pro’ and ‘full’ indicate pronouns and full NPs, respectively. For instance,  $VS_{full}O_{pro}$  stands for an VSO order where the subject is a full NP and the object is a pronoun.

**Null Context Condition** Figure 1 graphs the average judgments for each word order. An ANOVA for the null context condition revealed a highly significant main effect of *Vord* (verb order) ( $F_1(1, 19) = 56.911, p < .0005; F_2(1, 7) = 621.924, p < .0005$ ): XV orders (mean = .1879) were more acceptable than VX orders (mean = -.2129). A highly significant main effect of *Cord* (complement order) was also obtained ( $F_1(1, 19) = 26.966, p < .0005; F_2(1, 7) = 72.610, p < .0005$ ): SO orders (mean = .0659) were more acceptable than OS orders (mean = -.0909). The main effect of *Pro* (pronominalization) was significant by subjects only ( $F_1(3, 57) = 5.150, p = .003; F_2(3, 21) = 0.647, p = .593$ ).

The ANOVA also revealed a significant interaction of *Cord* and *Pro* ( $F_1(3, 57) = 13.026, p < .0005; F_2(3, 21) = 4.663, p = .012$ ). This indicates that pronominalization has an influence on complement order preference. We also found interactions of *Cord* and *Vord* ( $F_1(1, 19) = 47.437, p < .0005; F_2(1, 7) = 17.148, p = .004$ ) and of *Vord* and *Pro* (significant

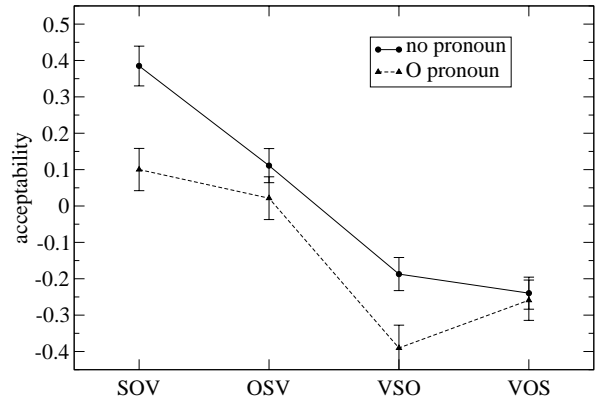


Figure 2: Interaction for word order and pronominalization, all focus context

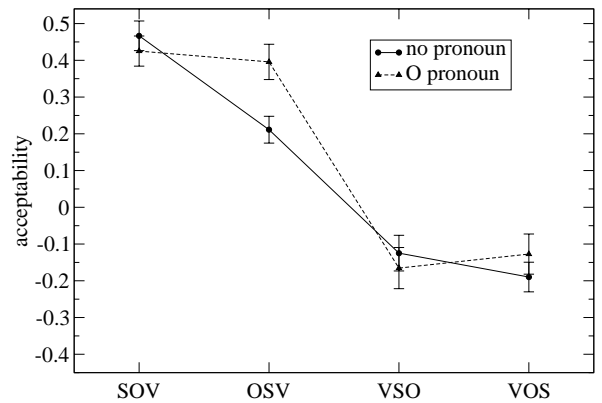


Figure 3: Interaction for word order and pronominalization, subject focus context

by subjects only,  $F_1(3, 57) = 4.223, p = .009; F_2(3, 21) = 1.107, p = .368$ ). A three-way interaction *Vord/Cord/Pro* was also present (significant by subjects only,  $F_1(3, 57) = 7.415, p = .009; F_2(3, 21) = 1.900, p = .161$ ).

The meaning of the interactions involving *Vord* becomes clear from Figure 1: the effect of pronominalization on complement order is limited to verb final orders; all verb initial orders are equally unacceptable, independent of complement order and pronominalization.

**Context Condition** Figures 2–4 graph the average judgments for each context. An ANOVA for the context condition confirmed the main effect of verb order found in the null context condition ( $F_1(1, 30) = 121.507, p < .0005; F_2(1, 7) = 225.903, p < .0005$ ): XV orders (mean = .2519) were more acceptable than VX orders (mean = -.1973). The main effect of complement order could also be replicated ( $F_1(1, 30) = 40.275, p < .0005; F_2(1, 7) = 15.359, p = .006$ ): SO orders (mean = .0785) were more acceptable than OS orders (mean = -.0239). A highly significant main effect of *Con* (context) was also present ( $F_1(2, 60) = 28.953, p < .0005; F_2(2, 14) = 54.056, p < .0005$ ), as well as a weak effect of *Pro* ( $F_1(2, 60) = 5.564, p = .025; F_2(2, 14) = 1.511, p = .259$ ).

The ANOVA uncovered an interaction of *Cord* and context,

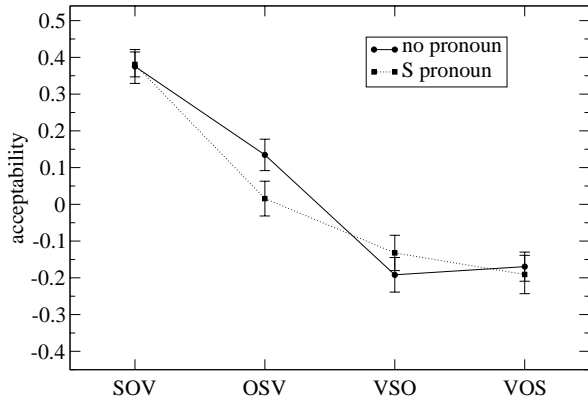


Figure 4: Interaction for word order and pronominalization, object focus context

significant by subjects and marginal by items ( $F_1(2, 60) = 6.016, p = .004; F_2(2, 14) = 3.076, p = .078$ ), which confirms that information structure (manipulated by context) has an influence on complement order preferences. We also found a marginal interaction of *Cord* and *Pro* ( $F_1(1, 30) = 4.025, p = .054; F_2(1, 7) = 3.634, p = .098$ ) and a highly significant interaction of *Pro* and context ( $F_1(2, 60) = 11.864, p < .0005; F_2(2, 14) = 16.07, p < .0005$ ). Recall that our materials were designed such that in all focus and S focus contexts, the object was pronominalized, while in an O focus context, the subject was pronominalized. This means that the *Cord/Pro* and *Pro/Con* interactions are only meaningful with respect to the three-way interaction *Cord/Pro/Con* which was also significant ( $F_1(2, 60) = 19.718, p < .0005; F_2(2, 14) = 7.73, p = .005$ ). This interaction demonstrates that the ordering of pronouns is subject to contextual effects (which will be discussed in the next section). The ANOVA also showed an interaction of *Vord* and *Cord* ( $F_1(1, 30) = 50.960, p < .0005; F_2(1, 7) = 7.221, p = .031$ ) and of *Vord* and context ( $F_1(2, 60) = 10.589, p < .0005; F_2(2, 14) = 11.945, p = .001$ ). The meaning of these interactions becomes clear from Figures 2–4: the interaction between complement order and context is limited to verb final orders; all verb initial orders are equally unacceptable, independent of context.

## Discussion

All differences referred to in the following were significant according to post-hoc Tukey tests (space limitations prevent the inclusion of the full set of Tukey results).

**Weighted Constraints** The experimental findings for the null context condition provided broad support for the ordering constraints in (4), initially proposed by Uszkoreit (1987). There was a clear preference for XV over VX, in line with the predictions of the verb ordering constraint VERB. The NOM constraint, which requires nominative to precede accusative, received support from the fact that SO orders were more acceptable than OS orders. Finally, the constraint PRO, which requires that pronouns precede full NPs, explains why  $S_{full}O_{pro}$  is less acceptable than  $S_{pro}O_{full}$ , while  $O_{full}S_{pro}$  is less acceptable than both  $O_{full}S_{full}$  and  $O_{pro}S_{full}$  (see Figure 1).

The interactions involving the factor *Vord* demonstrated that the effects of NOM and PRO disappear if the constraint VERB is violated. This indicates that a violation of VERB is more serious than violations of PRO or NOM, which in Uszkoreit’s framework means that VERB receives a higher weight than both PRO and NOM.

The behavior of VERB was replicated in the context condition. VERB leads to serious unacceptability in all contexts and blocks out all other constraint effects. Note, however, that we found an interaction of PRO and context that does not readily follow from Uszkoreit’s account. The prediction that pronouns have to precede full NPs is only born out in the all focus context. In the S focus and O focus contexts, the effect of PRO disappears, which might indicate that PRO is only valid if the context fails to provide an antecedent for the pronoun.

S focus and O focus contexts showed evidence for FOC, the constraint that requires ground elements to precede focus elements. In both contexts, SO was the preferred order, even though it violates FOC in the S focus context; in both contexts, the acceptability of OS was reduced compared to SO. However, this reduction was significantly higher in the O focus context, where OS violates FOC. The overall SO preference (even if it is disfavored by the context) indicates that the effect of FOC is weak compared to the influence of NOM, i.e., NOM should receive a higher weight than FOC. Only for OS orders, i.e., when NOM is violated, the influence of FOC becomes visible. No effects of context were found for VX orders, which indicates that FOC has a lower weight than VERB, just like NOM and PRO.

Overall, we have established the following facts about constraint weights: VERB has a higher weight than PRO, NOM, and FOC. NOM, on the other hand, has a greater weight than FOC. This is compatible with the following weight assignments:

$$(9) w(\text{VERB}) = 3, \quad w(\text{PRO}) = w(\text{NOM}) = 2, \quad w(\text{FOC}) = 1$$

To conclude, our results provide support for Uszkoreit set of word order constraints and show that his weighted constraint model is able to account for the experimental data.

**Optimality Theory** Note that the weights in (9) can also be interpreted as a set of OT-style constraint ranks:

$$(10) \text{VERB} \gg \{\text{PRO}, \text{NOM}\} \gg \text{FOC}$$

This ranking is compatible with Müller’s ranking in (6). Note that the effect of the AN (animacy) could not be tested in the present experiment: all nominative NPs were animate, while all accusative NPs were inanimate, hence a violation of NOM also entails a violation of AN.

Müller distinguishes between grammaticality and markedness, and predicts that ungrammatical candidates are categorically unacceptable, while marked structures are only dispreferred. PRO is a classified as a grammatical constraint, and hence should induce categorical unacceptability. Our data provides counterevidence to this prediction: in the null context,  $S_{full}O_{pro}V$  and  $O_{pro}S_{full}V$  are equally acceptable (see Figure 1), even though  $S_{full}O_{pro}V$  violates PRO while  $O_{pro}S_{full}V$  does not (the same pattern occurs in the all focus and S focus contexts). This is unexpected under the assumption that PRO is grammatical constraint; the data suggest that it should be reclassified as a markedness constraints.

On the other hand, VERB seems to be a genuine grammatical constraint. Its violation leads to strong ungrammaticality in all contexts, independently of which other constraints are violated (see Figures 1–4). This indicates that VERB (not explicitly dealt with by Müller) is a grammatical constraint.

Hence our data provides counterevidence for some aspects of Müller's particular account of word order in German. However, the main feature of his model, viz., the distinction between grammaticality and markedness, is supported by our experimental results.

**Syntactic Weight** Several of the order preferences found in this experiment are incompatible with Hawkins's account in terms of ICR. Most strikingly, we found that  $S_{full}O_{full}V$  is more acceptable than  $O_{full}S_{full}V$  (see Figures 1–4), even though both have the same ICR (see (7)).

As far as pronominalization is concerned, we found that in the null context,  $S_{pro}O_{full}V$  is more acceptable than  $S_{full}O_{pro}V$  and  $O_{pro}S_{full}V$  is more acceptable than  $O_{full}S_{pro}V$  (see Figure 1), consistent with the ICR predictions (see (8)). However, the predictions with respect to double pronouns were not born out: these receive the maximum ICR score of 1.0, but we found that the orders  $S_{pro}O_{pro}V$  and  $O_{pro}S_{pro}V$  are as unacceptable as  $S_{full}O_{pro}V$  and  $O_{full}S_{pro}V$ , respectively, even though these orders only have an ICR of .73 (see (8)). Also, the fact that  $S_{full}O_{pro}V$  and  $O_{pro}S_{full}V$  are equally acceptable is unexpected as these orders differ in ICR (see (8)). This observation holds across contexts, see Figures 1–4.

Also the focus effects we found are unexpected under a syntactic weight account: the acceptability of OSV is increased in an O focus context (compared to an S focus context, see Figures 3, 4), even though the ICR remains constant.<sup>4</sup> Finally, the fact that VX structures are severely unacceptable across the board does not follow from syntactic weight—in fact VX orders have a higher ICR than XV orders (see (7)).<sup>5</sup>

To summarize, while corpus data seems to support a syntactic weight account (see Hawkins, 1992, for details), the acceptability judgments in our experiment are largely incompatible with Hawkins's predictions.

## Conclusions

We reported the results of a study of word order variation in German that investigated the interaction of syntactic (complement order and verb order) and information structural constraints (pronominalization and focus). The data were used to evaluate a set of competition-based models of word order, including (a) Uszkoreit's (1987) weighted constraint model, (b) Müller's (1999) optimality theoretic account, and (c) Hawkins's (1992) syntactic weight model.

The experimental data are broadly compatible with models (a) and (b), indicating that a relativized (ranked or weighted) notion of linguistic constraints is essential for explaining word order preferences. Model (c), however, was not

<sup>4</sup>Note that Hawkins (1992, p. 196) concedes that informational concepts like focus play a limited role in 'structures for which syntactic weight makes either no predictions or weak predictions'.

<sup>5</sup>However, Hawkins argues that languages can grammaticalize word orders, which then are no longer subject to syntactic weight. This would explain the general unacceptability of VX in subordinate clauses in German.

well-supported by the data. While this model may be suitable for describing word order distributions in corpora, it does not seem to be directly applicable to contextualized acceptability judgments such as the ones reported in the present paper.

On the other hand, we found that some of the individual linguistic assumptions made by Uszkoreit and Müller were not born out in our data. This highlights the fact that informal acceptability judgments are not sufficient to clarify the intricate preference patterns that emerge from the interaction of syntactic, pragmatic, and phonological constraints on word order. Experimentally collected judgments are necessary to obtain reliable, delicate data that can inform detailed models of word order preferences.

The results of the present study have been replicated for a free word order language (Greek) and for spoken stimuli (Keller & Alexopoulou, 2000).

## References

- Bader, M., & Meng, M. (1999). Subject-object ambiguities in German embedded clauses: An across-the-board comparison. *Journal of Psycholinguistic Research*, 28(2), 121–143.
- Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72(1), 32–68.
- Choi, H.-W. (1996). *Optimizing structure in context: Scrambling and information structure*. Unpublished doctoral dissertation, Stanford University.
- Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage Publications.
- Haider, H. (1993). *Deutsche Syntax generativ: Vorstudien zur Theorie einer projektiven Grammatik*. Tübingen: Gunter Narr.
- Hawkins, J. A. (1992). Syntactic weight versus information structure in word order variation. In J. Jacobs (Ed.), *Informationsstruktur und Grammatik* (pp. 196–219). Opladen: Westdeutscher Verlag.
- Jacobs, J. (1988). Probleme der freien Wortstellung im Deutschen. In I. Rosengren (Ed.), *Sprache und Pragmatik* (Vol. 5, pp. 8–37). Department of German, Lund University.
- Keller, F. (1998). Gradient grammaticality as an effect of selective constraint re-ranking. In M. C. Gruber, D. Higgins, K. S. Olson, & T. Wysocki (Eds.), *Papers from the 34th Meeting of the Chicago Linguistic Society* (Vol. 2: The Panels, pp. 95–109). Chicago.
- Keller, F., & Alexopoulou, T. (2000). *Phonology competes with syntax: Experimental evidence for the interaction of word order and accent placement in the realization of information structure* (Rutgers Optimality Archive No. ROA-351-1099).
- Müller, G. (1999). Optimality, markedness, and word order in German. *Linguistics*, 37(5), 777–818.
- Pechmann, T., Uszkoreit, H., Engelkamp, J., & Zerbst, D. (1994). *Word order in the German middle field: Linguistic theory and psycholinguistic evidence* (CLAUS Report No. 43). Department of Computational Linguistics, Saarland University.
- Prince, A., & Smolensky, P. (1993). *Optimality Theory: Constraint interaction in generative grammar* (Technical Report No. 2). Center for Cognitive Science, Rutgers University.
- Scheepers, C. (1997). *Menschliche Satzverarbeitung: syntaktische und thematische Aspekte der Wortstellung im Deutschen*. Unpublished doctoral dissertation, University of Freiburg.
- Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.
- Stevens, S. S. (Ed.). (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York: John Wiley.
- Uszkoreit, H. (1987). *Word order and constituent structure in German*. Stanford, CA: CSLI Publications.
- Vallduví, E. (1992). *The informational component*. New York: Garland.

# The Précis of Project Nemo, Phase 2: Levels of Expertise

**Susan S. Kirschenbaum** (kirschenbaumss@csd.npt.nuwc.navy.mil)

Naval Undersea Warfare Center Division  
Code 2214, Building 1171/1  
Newport, RI 02841 USA

**Wayne D. Gray** (gray@gmu.edu)

Human Factors & Applied Cognition  
George Mason University; m/s 3f5  
Fairfax, VA 22030 USA

## Abstract

Project Nemo examines the cognitive processes and representational structures used by submarine Commanders while attempting to locate an enemy submarine hiding in deep water. In phase 2 we collected performance and protocol data from junior, mid-career, and senior submarine officers. The data support the conclusions from phase 1 (Gray, Kirschenbaum, & Ehret, 1997) that most AO actions can be characterized as a sequence of small, steps in a shallow goal hierarchy (rather than as following a detailed master plan). The nature of these successive choices vary as a function of the officer's expertise. The results are congruent with an interpretation in which the process of schema instantiation provides the control of cognition.

## Introduction

In phase 1 of Project Nemo (Gray et al., 1997) we analyzed six hours of verbal and action protocols from expert submarine Approach Officers (AOs) as they detected and localized (i.e., determined the course, speed, and range) a hostile submarine hiding in deep water.

The results of phase 1 support a description of the cognitive control structure that orchestrates the AOs' behavior as *schema-directed problem solving with shallow and adaptive subgoaling* (Ehret, Gray, & Kirschenbaum, in press). The schema is the task-relevant knowledge accumulated in over 20 years of experience as a submariner (half of it at sea). It is a set of declarative as well as procedural knowledge structures. An implication of shallow subgoaling is that the knowledge available to AOs is so rich that steps to supplement this knowledge can be shallow.

A second implication is that the AO solves a series of problems, one every 30 to 300 s. The problem is always the same; namely, "what is the state of the world – NOW." The AO is trying to find a quiet target hiding in a noisy environment while remaining undetected himself. The protocol analysis revealed that he takes a series of short steps that either (a) assess the noise from the environment or signal from the target – NOW, or (b) attempt to reduce the noise or increase the signal from the target by maneuvering ownship. As shown in Figure 1, these short steps result in shallow subgoaling. When a

subgoal pops, the schema is reassessed. The result of this reassessment directs the next step (i.e., selects the next subgoal). This step is accomplished, it returns information to the schema, the schema is reassessed, and so on.

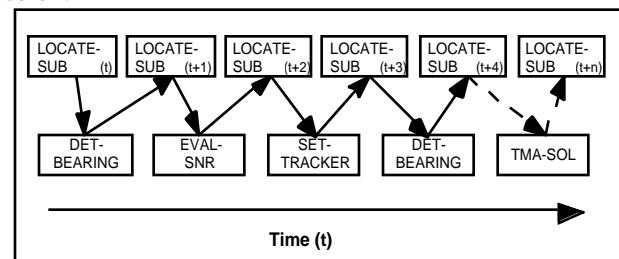


Figure 1: Schema-directed problem solving with shallow and adaptive subgoaling

The process of subgoaling is adaptive in two senses. First, the subgoal that is chosen next reflects the current reassessment of the schema. Second, this choice is sensitive to both the long-term importance of the subgoal as well as its recent history of success or failure. Regardless of a goal's long-term importance, AOs will not continue to attempt a goal if successive tries fail. Instead, they will choose another goal and return to the more important goal later.

The dynamic aspect of the AO's task plays an important role in this view of schema-directed problem solving. First, the state of the AO's world is continually changing – both ownship and target are moving at a given depth, direction, and speed. For ownship the value of these attributes can be changed, but the problem cannot be stopped. Consequently, time is an important part of the picture. Second, subgoals are not accomplished once and then discarded. In the AO's world, subgoals bring in certain types of information or accomplish certain changes to ownship. As the world changes, any given subgoal may be revisited not only to acquire the current value, but also to acquire information about the rate and direction of change (e.g., DET-BEARING in Figure 1).

Phase 1 ran 10 senior officers on a high fidelity simulation located at the Naval Undersea Warfare Center



in Newport, RI. For phase 2, we built the Ned<sup>1</sup> scaled world in Macintosh Common Lisp to run on a portable computer. (A description of the simulation can be found in Ehret et al., in press.) This portability enabled us to take Ned to U. S. Navy submarine bases in Bangor, WA and Pearl Harbor, HI. Consequently, we were able to collect data from 36 active-duty submarine officers.

In this paper we present a brief overview of the phase 1 empirical data. (More details can be found in Gray et al., 1997; and Kirschenbaum, Gray, & Ehret, 1997.) Our focus is on the data collected using the Ned scaled world, its similarities to the phase 1 data, and the variations among levels of expertise.

### **The Submariner's Task and Tools**

The job of the Approach Officer is to respond to hostile targets. He<sup>2</sup> heads the team that must detect, track, classify, localize, and if ordered, attack the target. He performs this task with the support of many special-purpose systems run by skilled operators, but is ultimately responsible for the success of the encounter.

Two features of the problem make it an especially challenging one. First, this is a dynamic problem. Both ownship and the contacts are moving, and, perhaps, changing course, etc. during the encounter. Second, there are only sparse and highly uncertain data about the contacts. The AO's expertise lies in using his knowledge of the relationships among the cues to guide information gathering over the course of the scenario and instantiate a generic "contact" schema for each contact.

Special tools are used for controlling own ship, listening to the contact, and localizing. As sound transmission is distorted, reflected and bent in the ocean by temperature, salinity, pressure, detecting, tracking, and locating the source of a passive sonar contact is highly very difficult and impacted by uncertainty. Because passive sonar only provides bearing (direction) data, target-motion-analysis (TMA) tools for localizing the targets employ statistical methods to estimate the target's course, speed, and range. As this is a mathematically under-constrained problem, submariners call this process "finding a solution."

## **Review of Phase 1**

### **Method**

All of the participants in phase 1 were highly experienced submarine officers who had served as Commanding Officers (COs) or Executive Officers (XOs) aboard U. S. Navy Submarines. The officers were presented with scenarios that required localizing an enemy submarine hiding in deep water. The scenarios were presented on the CSEAL (Combat Systems Engineering and Analysis Laboratory) high fidelity simulation. CSEAL is a submarine command-center-in-a-box. It has generic

versions of all of the essential submarine tools. As CSEAL is a developer's tool, it was run by an operator. The AOs requested information and ordered actions to be carried out by the operator. Videotapes and verbal protocols were the primary data. These were supplemented by computer-logged action protocols.

In both phases we investigated the situation assessment part of the AO's mission. Situation assessment begins with detection of the contact and ends when the AO is sufficiently confident of the solution to declare that he is ready to move to the engagement phase. Each scenario began with a status report such as an AO might receive when first taking his turn on watch. The status report provided scenario specific information including ownship course, speed, and depth as well as information on any contacts. All scenarios began with a single contact, classified as a merchant.

### **Review of Phase 1: Results**

During phase 1 we developed an encoding scheme (Gray & Kirschenbaum, in press) that included nine operators and a three-level goal structure (for detailed information, see Kirschenbaum et al., 1997). Most of the AOs' time and effort was spent in service of two top-level goals: LOCATE-MERCHANT (LOC-MERC) and LOCATE-SUBMARINE (LOC-SUB). Given that localizing the sub is clearly the higher priority, we were puzzled to find that the two goals were used with approximately equal frequency (see the left side of Figure 2). However, as the middle of Figure 2 shows, this equal frequency of use masked a large difference in the number of subgoals per LOC-MERC versus LOC-SUB.

More interesting, this disparity in number of subgoals per goal was not reflected in the number of operators per subgoal. As shown by the right-side of Figure 2, the mean number of operators per subgoal was constant. The same number of operators were used in a subgoal regardless of whether its supergoal was LOC-MERC or LOC-SUB.

Along with other analyses that we conducted, this analysis suggested that the basic unit of action was the subgoal. Formalized plans or established methods with fixed number of steps, exist at the subgoal level. At the level of LOC-MERC or LOC-SUB, each subgoal returns a discrete piece of knowledge that is added to the schema. The schema is reevaluated to determine what piece of knowledge to select next. When there is little new information to be gained by continuing working on the current goal, the goal is popped and a new top-level goal is pushed.

The question pursued below is whether the phase 2 data support the phase 1 interpretation of expert performance and in which ways intermediate and novice behavior conforms or differs from the experts.

---

<sup>1</sup> Ned Land was an able seaman and trusted assistant to Prof. Aronax aboard the Nautilus.

<sup>2</sup> In the current US Navy all submariners are men.

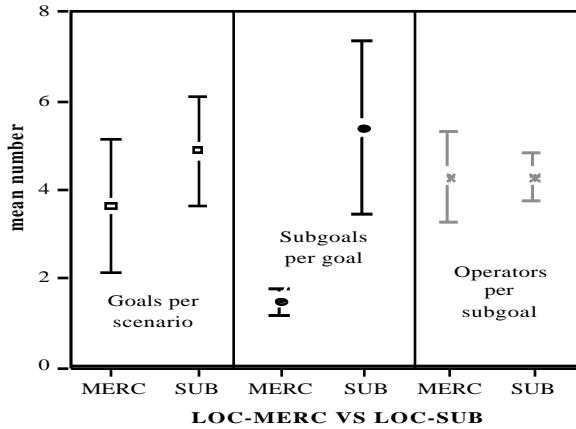


Figure 2: Phase 1. Data for the two main top-level goals, LOCALIZE-MERC and LOCALIZE-SUB. Left -- mean number of level-1 goals per AO-Trial. Middle -- mean subgoals per goal. Right -- mean number of operators per subgoal. [Error bars show the 95-percent confidence intervals for the standard error of the mean (SEM).]

### The Ned Experiment: Replication and Expansion

Table 1: Demographic data on participants.

Means for	CO/XO	DH	JO
n	15	11	10
Years at sea	8.7	6.4	3.2
Years in Navy	17.8	13.4	7.3
Age	38.9	34.4	28.3

### Participants

Three groups of current submarine officers participated in the study: Junior Officers (JOs), Department Heads (DHs), and Commanding Officers or Executive Officers (CO/XOs). The average number of years spent at sea, years in the navy, and ages can be found in Table 1. The expert participants in this study had slightly less experience than those in phase 1. This was most likely because, unlike the phase 1 AOs, all were active duty and none were post-command. (The phase 1 COs had a mean of 10.8 years at sea and 20.3 years in the Navy.)

### Ned Simulation

The Ned simulation was designed to overcome problems encountered in collecting and encoding data from CSEAL. (These problems and their solution are detailed in Ehret et al., in press.) For the current discussion, the two most relevant improvements in Ned over CSEAL were the elimination of redundant information and the control that Ned provided over access to information.

With minor exceptions, Ned's displays were specialized so that each type of information was available from one display only. For example, when an AO went to the display for the broadband spherical sonar sensor, we could be sure that he wanted one of the 10 types of information

that was available only from that display. Once an AO selected a display, all information fields for the display were covered by black boxes (as in the bottom display of Figure 3). Clicking on the field removed the black box and revealed the data until the mouse was moved from the field. (Ned consists of 10 specialized displays.)

Ned captured all AO interactions, including display navigation and viewing information (enter and exit times and information content). It also recorded truth every 20 seconds. In addition, the AOs were encouraged to think aloud and all sessions were video recorded.

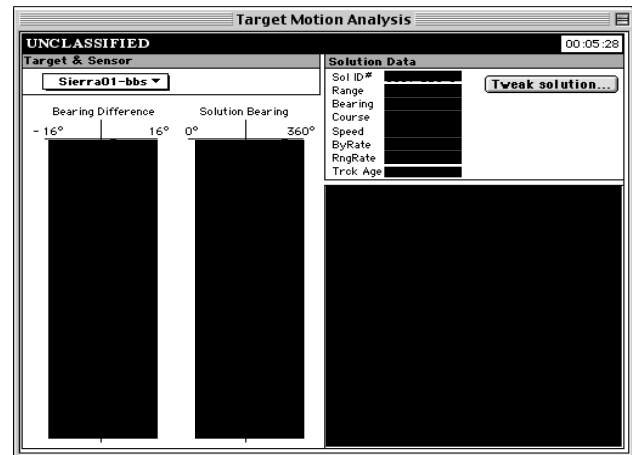
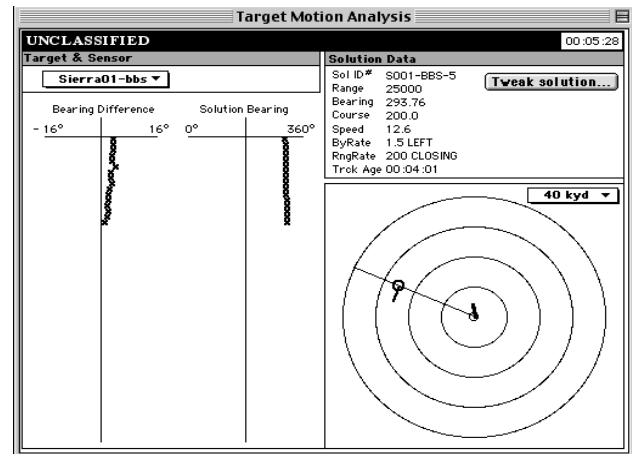


Figure 3: The Ned Target Motion Analysis screen without (above) and with (below) black boxes covering data fields.

### Scenarios

Four scenarios were used. Two were identical to those used in phase 1 and two were slightly modified versions of the phase 1 scenarios. At the beginning of each scenario the AO had ownship position (course, speed, and depth) and confirmed contact and bearing (direction from ownship) for a merchant. He also had intelligence that a "hostile" submarine was in the region.

### Procedures

Each session began with training on Ned and training in talking aloud while problem solving. Each AO solved two

scenarios. Sessions lasted approximately 2 hours.

### Protocol Encodings

Five CO/XO's were unable to complete two scenarios due to interruptions for other responsibilities. From the remaining 31 AOs, data from six AOs at each experience level (18 in all) were selected for detailed encoding. In each case, the second scenario was encoded. Protocols were selected on the basis of completeness, the lack of technical glitches, and the clarity of the recorded protocols.

### Semi-Automatic Protocol Encoding

Each click of the mouse on a menu item or a black box was saved to a log file. This enabled us to write code that encoded each action protocol and segmented groups of related actions. For example, if the AO clicked on the black box covering the range information in Figure 3 (see upper right part of the display), he was credited with two operators one for querying and one for receiving range information from the target motion analysis system.

### Operators

Operators are the lowest level encoding and represent the AO's direct interaction with the Ned simulation. Unlike phase 1, the majority of operators (approximately 99%) were encoded automatically from the computer outputted action protocols. All encodings were confirmed and/or modified by comparison with the video-taped verbal protocols. Across the three groups a total of 9,073 operators were encoded as belonging to one of nine operator categories.

Table 2: Example of goal and operator encodings.

GOAL	OP	INFO-SOURCE	SHIP	ATTR	VALUE	DUR
DETERMINE-BEARING						
	QUERY	NBT-BY-FIELD	SUB	BY		
	RECEIVE	NBT-BY-FIELD	SUB	BY	316 or 244	1.15
POP						

An example of the result of the automatic encoding of operators is provided in Table 2. Prior to this point in the scenario, the AO has called up the narrowband towed display (NBT). Here he queries the bearing (BY) information for the SUB by clicking on the black box that covers the field. The black box disappears, enabling the AO to receive the information that the narrowband towed sensor gives the ambiguous information that the hostile submarine's bearing from ownship is either 316 or 244 degrees. The bearing information is uncovered for 1.15 sec before the AO moves the cursor out of the bearing field.

Details of the operator types and categories used in phase 1 are available from Kirschenbaum, et al. (1997).. The phase 2 operator types and categories differed minimally from those used in phase 1; however, their similarity and differences from the phase 1 operators are beyond the scope of the current report.

### Goals and Subgoals<sup>3</sup>

The AO's mission, as given in the instructions, is to destroy the hostile submarine. Therefore his primary goal is to detect and localize the sub. However, these are not his only goals. He must also keep track of the merchant, avoid collision, evaluate the environment, and keep track of ownship.

Under Ned we have a precise record of the AO's information access. This record, linked by time to the verbal protocol, permitted a more detailed encoding of goals than was possible for phase 1. Hence, the goal and subgoals used in phase 2 differed from those discussed in Kirschenbaum et. al, (1997). However, the discussion of these differences will have to await a fuller report.

Of the 18 scenarios studied in phase 2, six were used to train the three encoders. These are referred to as "consensus encodings." The operators for each of the remaining 12 scenarios were encoded into goals by two independent encoders. Cohen's Kappa for interrater reliability averaged 0.84 and ranged from a low of 0.54 to a high of 0.96. All correlations are significant ( $p < .001$ ). The discrepancies between encodings were resolved by the third encoder.

Table 3: Typical goal, subgoal, operator sequence. (This is a truncated sequence and is for illustrative purposes only).

L-1	L-2	L-3	OPERATOR
LOC-SUB			
EVALUATE-TRACE			Query
			Receive
			Derive
			DisplayNav
TMA-SOLUTION			
EVAL-SOLUTION-BEARING			Query
			Receive
			Derive
TWEAK-SOLUTION-BEARING			Tweak
EVALUATE-SOLUTION-RANGE			Query
			Receive
			Derive
TWEAK-SOLUTION-RANGE			Tweak
DETERMINE-SOLUTION-QUALITY			Query
			Receive

### Goals and Schema

The schema instantiation process that we hypothesize controls cognition during situation assessment proceeds by pushing and popping a series of largely independent subgoals (see Figure 1). When a goal pops, information is returned to the schema being instantiated. The amended instantiation selects the next goal to push. For example, a typical sequence of goals, subgoals, and operators might read like that in Table 3

<sup>3</sup> For ease of exposition, level-1 goals will be referred to simply as *goals*; level-2 and level-3 subgoals collectively as *subgoals*.

In this sequence the AO first evaluates the sonar trace. This may return information to his schema regarding the target's bearing and bearing rate. He then switches to the display shown in Figure 3 and examines the TMA solution, alternately evaluating and tweaking the values of different parameters. As the values are returned to his schema he can compare them with his knowledge of how targets and the TMA algorithms work to derive better values to test. At the end of the sequence, he determines the solution quality by examining how closely the dots in the bottom-left section of the TMA screen (Figure 3a) stack on the central line.

### Summary

Ned records AO actions with greater specificity and accuracy than permitted by CSEAL. Consequently, we revised the goal types and categories to take advantage of this greater detail. However, the phase 2 revisions are elaborations on the goal categories and types used in phase 1. Thus, the phase 1 goal structure, with minor modifications, can support the detailed analysis of Ned data.

### Data Analysis and Results

The 9,073 operators collected in phase 2 can be aggregated and examined for many different purposes. In the current paper we limit our purposes to three. First we generally compare the goal and subgoal structure used in phase 2 with that of phase 1. For our second and third purpose, we limit ourselves to the three measures used in Figure 2: LOC-MERC and LOC-SUB goals per scenario, number of subgoals per LOC-MERC and LOC-SUB, and number of operators per subgoal. We begin by using these measures to compare the experts in phase 2 (i.e., the CO/XO's) to those in phase 1. We then use these same measures to look across levels of expertise for phase 2.

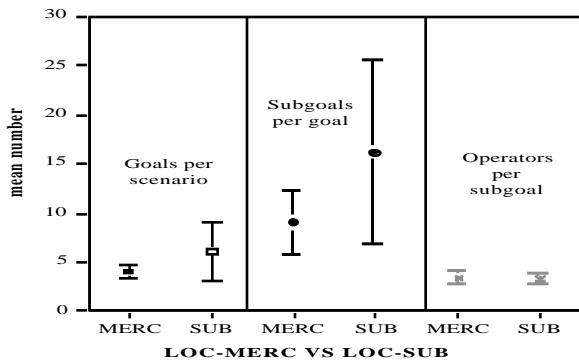


Figure 4: Phase 2 : Data for the two main top-level goals, LOCALIZE-MERC and LOCALIZE-SUB. Left -- mean number of level-1 goals per AO-Trial. Middle -- mean subgoals per goal. Right -- mean number of operators per subgoal. (Error bars show the 95% CI for SEM.) Compare with Figure 2.

### Comparison with CSEAL Data

The Ned data replicated all of the major findings reported in phase 1. In phase 1 we reported a relatively flat goal

hierarchy of 2-3 levels. This is confirmed by the more precise Ned data. Level-3 goals were confined to three level-2 goals, and the large majority 62.1% of all level 3 goals, were subgoals of a single, high-frequency level-2 goal, TMA-SOLUTION.

Secondly, in phase 1 we were able to encode the protocols using only 9 operators. Nine operators worked well for phase 2. The only notable difference in operator sets was exchanging the N/A category from the verbal protocol encodings of phase 1 for a display-manipulation category (i.e., clicking on menu item or black-box) in phase 2. Also, as in phase 1, we found relatively few operators per goal with a mean of 6.0 operators per Level 2 subgoal and 3.4 operators per Level 3 subgoal (though this varied by subgoal).

### CO/XO: Phase 1 versus Phase 2 Comparisons of Expert Level Performers

Comparing the three frames of Figure 2 with those of Figure 4 yields a qualitatively similar picture. In both phases, although the differences in numbers are small, the CO/XOs push more LOC-SUB than LOC-MERC goals. However, these small differences at the goal level are countered by large differences at the subgoals level (middle frame of Figure 4). As in phase 1, for phase 2 the number of operators per terminal subgoal (right frame of Figure 4) does not differ as a function of the top-level goal.

These comparisons are consistent with our phase 1 conclusions that the subgoal level captures a basic unit of AO expertise. The goal level, LOC-MERC and LOC-SUB, divides the world into episodes. An episode requires a varying number of subgoals. The exact number depends on features of the current scenario. Merchants are noisy and easy to localize. Hence, most LOC-MERC episodes occur between attempts to detect the submarine and most end with the AO obtaining a good solution on the merchant.

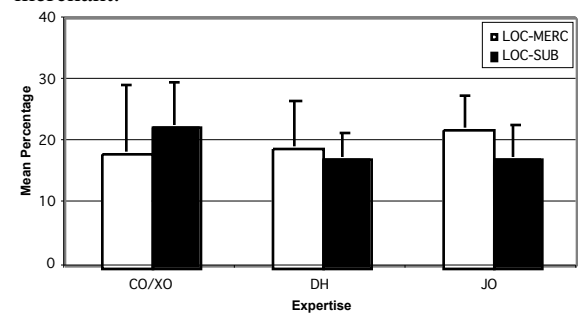


Figure 5: Mean frequency of LOC-MERC and LOC-SUB goals as a percentage of total Level-1 goal usage for three experience levels.

In contrast, enemy submarines are quiet and trying to avoid detection. Hence they are difficult to localize. Most LOC-SUB episodes end after the AO concludes that the current data are not very good and will not get better unless he can take some action to reduce noise or to collect data that will disambiguate data already collected. This decision to halt the current attempt to localize the submarine is never cut-and-dried.

These characterizations of the differences between LOC-MERC and LOC-SUB provide an explanation for the large differences in variance (see the error bars in the middle frame of Figure 2 and Figure 4) in number of subgoals per level 1 goal. For LOC-MERC, localizing is routine. In contrast, LOC-SUB requires flexibility to determine either that the current data are inadequate to enable the target to be localized or that the current best solution is such-and-such.

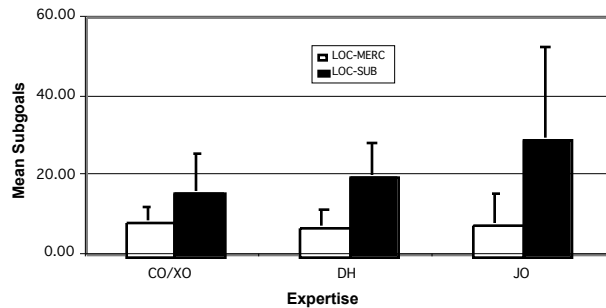


Figure 6: Mean total time spent in LOC-MERC and LOC-SUB goals for the three levels of expertise.

### Expertise Effects

All expertise groups pushed LOC-MERC and LOC-SUB goals with approximately the same frequency (see Figure 5). For all groups, within-group variability overshadows the apparent difference between the goal frequencies. Despite the approximately equal number of LOC-MERC and LOC-SUB goals, across expertise levels there were large differences in the amount of time spent trying to localize the merchant as opposed to the submarine (see Figure 6). The inequality in total time spent pursuing the two goals increases with inexperience.

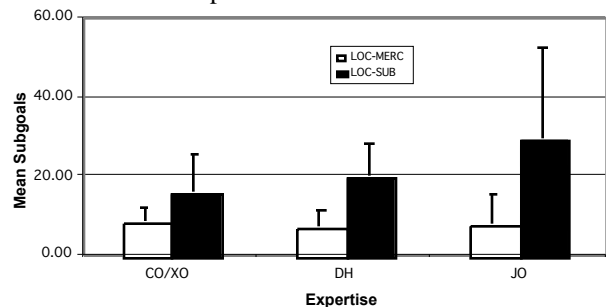


Figure 7: Mean number of subgoals for LOC-MERC and LOC-SUB goals.

As suggested by Figure 7, this difference in time as a function of expertise can be largely accounted for by differences in the number of subgoals. The Junior Officers use almost twice as many subgoals as the most experienced officers. Analyses not reported show that the number of operators per subgoal does not vary with expertise.

### Summary and Conclusions

The similarity between the CO/XO's in the two phases of Project Nemo support our characterization of performance

as schema-directed problem solving with shallow and adaptive subgoaling. The top-level goals, LOC-MERC and LOC-SUB, do not involve a fixed number of steps; rather, progress on a goal continues until a reevaluation of the schema determines that further effort would be wasted. What is fixed are the number of steps (operators) required for the terminal subgoals.

The phase 2 differences in expertise enrich our hypotheses. The most junior officers use the same building blocks as the most senior officers; that is, the same terminal subgoals are used with the same number of operators per subgoal. In contrast to the typical study of expertise, our "novices" were experienced (see Table 1). Very few officers switch branches of the Navy. Hence, our novices had spent 7.3 years in the Navy with 3.2 years at sea. All of their sea time was spent in submarines.

Where our novices differ from our experts is in their facility at schema-directed problem-solving. Simply put, the less experienced officers pursue bad data longer than the more experienced ones. The experienced ones know when it is time to give up on the current data set and do something to obtain more or better data.

### Acknowledgments

Susan S. Kirschenbaum's work has been jointly sponsored by Office of Naval Research (ONR) (Program element 61153N) and by Naval Undersea Warfare Center's Independent Research Program, as Project A10328. The work at George Mason University was supported by a grant from ONR (#N00014-95-1-0175) to Wayne D. Gray.

We thank Brian Ehret for his knowledgeable encoding, for his programming skills, and for his development of the Ned simulation. We also thank LT David Soldow for collecting the data for the Ned phase of the project and the officers of the U.S. Navy Submarine force for their participation, both as AOs and as consultants.

### References

- Ehret, B. D., Gray, W. D., & Kirschenbaum, S. S. (in press). Contending with complexity: Developing and using a scaled world in applied cognitive research. *Human Factors*.
- Gray, W. D., & Kirschenbaum, S. S. (in press). Analyzing a novel expertise: An unmarked road. In J. M. C. Schraagen, S. F. Chipman, & V. L. Shalin (Eds.), *Cognitive task analysis*. Mahwah, NJ: Erlbaum.
- Gray, W. D., Kirschenbaum, S. S., & Ehret, B. D. (1997). The précis of Project Nemo, phase 1: Subgoaling and subschemas for submariners. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 283-288). Hillsdale, NJ: Erlbaum.
- Kirschenbaum, S. S., Gray, W. D., & Ehret, B. D. (1997). *Subgoaling and subschemas for submariners: Cognitive models of situation assessment* (Technical Report 10,764-1). Newport, RI: NUWC-NPT.

# Visual and Spatial Representations in Relational Reasoning

**Markus Knauff** (mknauff@princeton.edu)

Princeton University, Department of Psychology;  
Green Hall, Princeton, NJ 08544 USA  
and

University of Freiburg, Center for Cognitive Science  
Friedrichstr. 50, 79098 Freiburg, Germany

**P. N. Johnson-Laird** (phil@princeton.edu)

Princeton University, Department of Psychology;  
Green Hall, Princeton, NJ 08544 USA

## Abstract

Psychologists have argued that visual imagery plays a vital role in human reasoning. If so, then reasoning with materials that are easy to visualize should be better than reasoning with materials that are hard to visualize. The literature, however, reports inconsistent results. Our starting point was that the inconsistencies arise from confounding imageability with the spatial nature of the materials. Hence, we manipulated the ease of envisaging the materials as visual images and also as spatial layouts. An experiment showed that materials that are easy to visualize impair reasoning unless they are also easy to envisage spatially.

## Introduction

“I am by the sea and I have a picture. This is a picture of a picture. I am – ” She screwed up her face and scowled – “thinking.” . . . She paused, frustrated by the vivid detail of her picture, not knowing how to extract from it the significance she felt was there.

– *The Inheritors*, William Golding, 1955, p. 62

Speculations about the role of visual imagery in human reasoning have a long history, and have recently surfaced again in the claims of computer scientists that reasoning based on diagrams has advantages from a computational point of view (Glasgow, Narayanan, & Chandrasekaran, 1995). Yet, the situation is not so clear as it should be from either a psychological or computational standpoint. In psychology, Kosslyn (e.g. 1994) and his colleagues have no doubt that visual imagery plays a key role in reasoning. The origins of this idea are the pioneering studies of DeSoto, London, & Handel (1965) and Huttenlocher (1968), who investigated so-called three-term series problems, such as:

Ann is taller than Beth.

Cath is shorter than Beth.

Who is tallest?

DeSoto et al. argued that reasoners imagine the three individuals on the vertical axis of a visual image, and then read off the answer by inspecting the image. Various sorts of

evidence support this hypothesis, including the well-known effects of mental rotation (Shepard & Cooper, 1982) and mental scanning (Kosslyn, 1980). Indeed, metrical information, which is often posited as the main characteristic of mental images, affects reasoning performance (Kelter & Kaup, 1995; Rinck, Hähnel, Bower, & Glowalla, 1997). Likewise, Pearson, Logie, & Gillhooly (1999) studied mental synthesis tasks, which elicit reasoning, and detected interference from visual secondary tasks.

In contrast, several studies have failed to find any effect of imageability on reasoning (Mynatt & Smith, 1977; Sternberg, 1980; Newstead, Pollard, & Griggs, 1986; Richardson, 1987; Johnson-Laird, Byrne, & Tabossi, 1989). Furthermore, Sternberg (1980) did not find a reliable correlation between reasoning ability and scores on imageability items of IQ-tests (Sternberg, 1980). Knauff and his colleagues found interference between relational reasoning and spatial secondary tasks but no such effects of visual secondary tasks (Knauff, Rauh, Schlieder, & Jola, 1999; Knauff, Jola, Strube, Rauh, & Schlieder, 2000).

From a computational point of view, the situation is similar. Researchers into diagrammatic reasoning have argued that diagrams are useful in solving problems, ranging from the analysis of molecular structure (Glasgow & Papadias, 1992) to the navigation of robots (Stein, 1995). Reasoning based on such analog representations can be more powerful than traditional propositionally based reasoning (Glasgow *et al.*, 1995). This approach, however, appears to conflict with theories of qualitative spatial reasoning. Their proponents argue that abstract representations of spatial relations together with an appropriate reasoning engine are a better way to enable computers to make predictions, diagnoses, and plans, when quantitative knowledge is unavailable or leads to computationally intractable inferences (Hernández, 1994; Cohn, 1997).

The aim of our research was to clarify the role of mental images in human reasoning. Our basic assumption is that the inconsistent psychological effects of imageability arise from

a failure to distinguish between visual characteristics and spatial characteristics of mental representations. On the one hand, if reasoning relies on mental images, then the easier it is to visualize the information in the premises, the better performance should be. On the other hand, if reasoning relies on spatial models, then the easier it is to envisage a spatial array, the better performance should be. We carried out a preliminary study of various relational terms to assess the ease of imagining assertions based on them as visual images and as spatial arrays. We then carried out an experiment to investigate the effects of both these factors on relational reasoning.

### A preliminary study

In order to generate the materials for our main experiment, 10 students at Princeton University, who were native speakers of English, filled out a questionnaire about the ease of forming visual images and spatial arrays for a set of thirty relational assertions, such as:

The cat was above the dog.

The assertions were based on such relations as *cleaner-dirtier*, *uglier-prettier*, *heavier-lighter*, *smarter-dumber*, and *above-below*. The participants rated the ease of forming visual images and of forming spatial arrays of the assertions on separate seven-point scales ranging from “very easy” to “very difficult”. The frequencies of usage of the relational terms were controlled word frequencies were controlled (Francis & Kucera, 1982), and the order of the assertions was counter-balanced across the participants.

Table 1: Three sorts of relational terms from the preliminary study and their mean ratings for ease of forming a visual image and a spatial array. The scales ranged from 7 (very easy) to 1 (very difficult)

	Visual image ratings	Spatial ratings
<u>Spatio-visual relations</u>		
above-below	5.3	5.4
front-back	5.2	5.3
<u>Visual relations</u>		
cleaner-dirtier	5.1	1.6
fatter-thinner	4.8	2.0
<u>Control relations</u>		
better-worse	2.1	1.1
smarter-dumber	2.8	1.2

The ratings of assertions based on a relation and its converse did not differ reliably, and so we pooled the results. The ratings enabled us to select three sorts of pairs of relations from the set as a whole. These pairs and their mean ratings are shown in Table 1. The three sorts of relations are: 1. relations such as *above-below* that were easy to envisage spatially and visually, which we henceforth

refer to as spatio-visual relations; 2. relations such as *cleaner-dirtier* that were hard to envisage spatially but easy to envisage visually, which we henceforth refer to as visual relations; and pairs such as *better-worse* that were hard to envisage either spatially or visually, which we henceforth refer to as control relations.

The differences between the three groups were statistically reliable, whereas there were no significant differences within the groups. None of the relations in the preliminary study were easy to envisage spatially but difficult to envisage visually.

### The Experiment

**Design.** The aim of the experiment was to investigate the effects of the three sorts of relational terms (visuo-spatial, visual, and controls) on relational reasoning. The participants acted as their own controls and evaluated inferences of all three sorts in 12 three-term series problems and 12 four-term series problems. The relations in these problems were those in Table 1. There were two valid and two invalid problems of each of the three sorts in both the three-term and four-term series problems, making a total of 24 problems. The problems were presented in a counterbalanced order over the set of participants.

**Participants.** We tested 22 undergraduate students of Princeton University (mean age 19.5; 12 female, 10 male), who received a course credit for their participation.

**Materials.** The three-term and four-term series problems all concerned the same terms (*dog, cat, ape* and *bird*). Here is an example of a problem with a valid conclusion:

The dog is cleaner than the cat.

The ape is dirtier than the cat.

Does it follow:

The dog is cleaner than the ape?

**Procedure.** The participants were tested individually in a quiet room, and they sat at a laptop computer that administered the experiment in separate stages (Potts & Scholz, 1977). The premises were presented one at a time on a sequence of screens (in black letters) followed by a putative conclusion (in red letters). The participants were told to evaluate whether or not the conclusion followed necessarily from the premises. They made their response by pressing the appropriate key on the keyboard, and the computer recorded their response and latency. Prior to the experiment, there were eight practice trials.

**Results.** The problems were easy, and 89 percent of the responses were correct. Furthermore, there were no significant differences in error rates for the three sorts of problems. Figure 1 shows the mean latencies for the correct responses to the three sorts of relational problems. As there was no reliable difference between the three-term and four-term series, we have pooled the results. The participants responded faster to the visuo-spatial problems (2200 ms) than to the control problems (2384 ms), though this difference was not significant, but slower to the visual problems (2654 ms) than to the control problems (Wilcoxon test  $z = 3.07$ ;  $p < .002$ ). Overall, the difference over the three groups was reliable (Friedman analysis of variance,  $F = 8.08$ ;  $p < .02$ ).

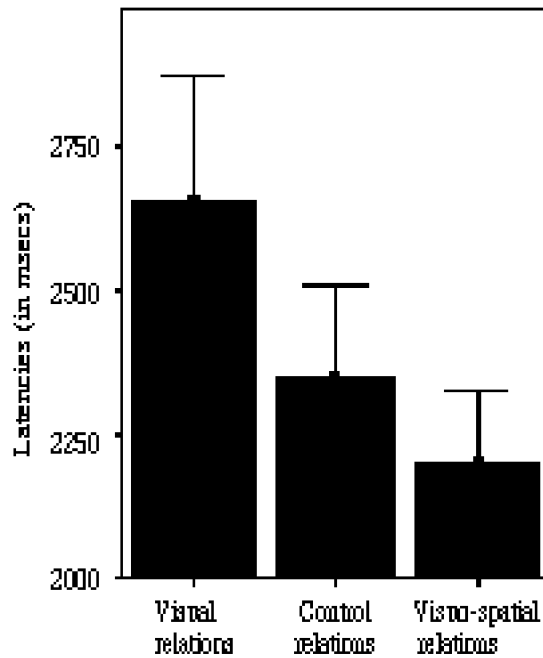


Figure 1: Mean reaction latencies [in milliseconds] and standard errors in the relational reasoning with the three sorts of relation: visual relations, control relations and visuo-spatial relations.

The differences are also reflected in the premise reading times. Because for all three premises we obtained a similar pattern of results, we pooled together all three premises. The mean reading times were 6.6s for the visual premises, 6.2s for the control premises, and 6.0s for the visuo-spatial tasks. The trend over the three groups was reliable (Page's  $L = 284$ ;  $p < .05$ ). Likewise, the difference between the visual and visuo-spatial premises was reliable (Wilcoxon test  $z = 2.07$ ;  $p < .05$ ).

## General Discussion

Our starting point was the conjecture that the conflicting results in the literature on imagery and relational reasoning arose from a failure to distinguish between visual images and spatial representations. Our preliminary study enabled us to identify (a) visuo-spatial relations, such as *above-below*, which are easy to envisage both visually and spatially, (b) visual relations, such as *cleaner-dirtier*, which are easy to envisage visually but hard to envisage spatially, and (c) control relations, such as *better-worse*, which are hard to envisage both visually and spatially. Unfortunately, we were unable to identify relations that were easy to envisage spatially but hard to imagine visually; and some colleagues doubt the existence of such relations. Nevertheless, the results of our experiment established the importance of distinguishing between visual and spatial representations. Visual relations such as *fatter* and *thinner*

significantly impede the process of reasoning in comparison with control relations such as *smarter* and *dumber*. In contrast, visuo-spatial relations, such as *front* and *back*, which are easy to envisage visually and spatially, speed up the process of reasoning in comparison with control relations (though the difference did not reach significance).

What causes the trend in our results? One possible explanation is suggested by the theory of mental models (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). It postulates that people make transitive inferences by constructing models of the situations that the premises describe. They possess neither axioms nor rules of inference for transitivity, but merely construct an appropriate model. For example, given the premises:

The cat is above the ape.

The dog is below the ape.

they construct a spatial model representing the relative positions of the three individuals:

cat

ape

dog

They evaluate a putative conclusion by checking whether it holds in the model. If it does, they search for a counterexample, i.e., a model that satisfies the premises but refutes the conclusion. Given that no such counterexample exists, the conclusion is valid (see Byrne and Johnson-Laird, 1989). Perhaps the ability to envisage spatial models is a precursor to many forms of abstract reasoning (Johnson-Laird, 1996). Likewise, relational terms that lead naturally to spatial models should speed up the process of reasoning. In contrast, a visual relation, such as *dirtier*, may elicit irrelevant visual detail. One imagines, say, a cat caked with mud, but such a representation is irrelevant to the transitive inference. It takes additional time to replace this vivid image with one in which dirtiness is represented in degrees. In other words, the visual relations, which are hard to envisage spatially, lead to a mental picture. But, the vivid details in this picture interfere with the process of thinking – much as they did for the character in our epigraph from William Golding's novel.

This interpretation is consistent with Logie's (1995) distinction between the visual and spatial subsystems in Baddeley's conception of working memory (Baddeley & Hitch, 1974; Baddeley, 1986). One subsystem (the visual cache) is linked to visual perception and the "visual buffer" (Kosslyn, 1994), and the other subsystem (the inner scribe) is amodal and handles spatial information for use by different cognitive and motor systems (Logie, 1995). Knauff and his colleagues have carried out a series of experiments in which the participants evaluated three-term series inferences as primary tasks together with visual and spatial secondary tasks (Knauff et al., 1999, 2000). The results showed that the spatial tasks interfered with reasoning, whereas the visual tasks did not interfere with reasoning.

A theoretical argument corroborating our hypothesis comes from a comparison of computational accounts of spatial reasoning. Schlieder (1999) compared two computational models of empirical data from Knauff and his colleagues (Knauff et al., 1995, 1998). One model was based on visual images with metrical information (Berendt,



1996), and the other model was based on diagrams that represent only the characteristic points of objects with no metrical information (Schleider's own model). This second spatial account yielded a better account of the empirical results.

An alternative account of our results, however, makes no appeal to the nature of mental representations. It is conceivable that the critical difference between the three sorts of relations is that they differ in the extent to which they suggest transitive relations over the individuals in our problems. Spatial relations among them are unequivocal, whereas the visual relations are more dubious. Given, say, the following premises:

The cat is fatter than the ape.

The ape is fatter than the dog.

reasoners may wonder whether the *fatness* of cats, apes, and dogs, is commensurate. Thus, when one asserts that an elephant is thin, the claim is relative, and so it is perfectly sensible to assert that a thin elephant is fatter than a fat dog. Hence, the criteria of fatness shift from one animal to another. This factor might have confused reasoners in our experiment momentarily, and accordingly lead to longer latencies with the visual relations. One strong argument against this account, however, is that the reading times for the individual premises also showed an advantage for visuospatial relations over visual relations. There remains one other possibility: the visuospatial relations were expressed by prepositions whereas the other relations in our experiment were expressed by comparative adjectives. It is conceivable that this factor, or some other unknown confound, might be responsible for our experimental results. Our next task is to examine in more detail our explanation in terms of irrelevant visual data.

### Acknowledgments

The first author is a visiting research fellow at the Department of Psychology, Princeton University, and the holder of a post-doctoral scholarship from the German National Research Foundation (DFG; Kn465/2-1). Both authors are grateful for the helpful comments of Uri Hasson, Juan Garcia Madruga, Vladimir Sloutsky, Yingrui Yang, and Lauren Ziskind on an earlier draft of this paper.

### References

- Baddeley, A.D. (1986). *Working memory*. Oxford: Oxford University Press.
- Baddeley, A.D. & Hitch, G. (1974). Working memory. In G.H. Bower (Eds.), *The psychology of learning and motivation* (vol. 8, pp. 47-89). New York: Academic Press.
- Byrne, R. M. J., & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory and Language*, 28, 564-575.
- Cohn, A. G. (1997). Qualitative spatial representation and reasoning techniques In: *KI-97: Advances in Artificial Intelligence* (pp. 1-30). Berlin: Springer.
- DeSoto, L B., London, M., & Handel, M.S. (1965). Social reasoning and spatial paralogue. *Journal of Personality and Social Psychology*, 2, 513-521.
- Girotto, V., Mazzocco, A. and Tasso, A. (1997). The effect of premise order in conditional reasoning: a test of the mental model theory. *Cognition*, 63, 1-28.
- Francis W. N. & Kucera, H. (1982). *Frequency Analysis of English Usage*. Boston: Houghton Mifflin Company.
- Glasgow, J.I. & Papadias, D. (1992). Computational Imagery, *Cognitive Science*, 17, 355-394.
- Glasgow, J. Narayanan, N. H. & Chandrasekaran, B. (1995) (Eds.). *Diagrammatic reasoning*. Cambridge: AAAI Press.
- Golding, W. (1955) *The Inheritors*. New York: Harcourt Brace.
- Hernández, D. (1994). *Qualitative representation of spatial knowledge*. Berlin: Springer-Verlag.
- Huttenlocher, J. (1968). Constructing spatial images: A strategy in reasoning. *Psychological Review*, 75, 550-560.
- Johnson-Laird, P.N. (1983). *Mental Models*. Cambridge: Cambridge University Press.
- Johnson-Laird, P. N. (1996). Space to think. In P. Bloom et al., *Language and Space* (pp. 437-462). Cambridge, MA: MIT Press.
- Johnson-Laird, P. N. & Byrne, R. (1991). *Deduction*. Hove (UK): Erlbaum.
- Johnson-Laird, P. N., Byrne, R. & Tabossi, P. (1989). Reasoning by model: the case of multiple quantifiers. *Psychological Review*, 96, 658-673.
- Kelter, S. & Kaup, B. (1995). Räumliche Vorstellungen und Textverstehen. Neuere Entwicklungen der Theorie mentaler Modelle. In B. Spillner (Eds.), *Sprache und Verständlichkeit. Kongressbeiträge zur 25. Jahrestagung der Gesellschaft für angewandte Linguistik* (S. 70-82). Frankfurt a. M.: Lang.
- Knauff, M., Rauh, R. Schlieder, C., & Jola, C. (1999). Räumliches Denken unter Arbeitsgedächtnisbelastung. [Spatial Thinking under working memory load]. In E. Schröger, A. Mecklinger, & A. Widmann (Eds.), *Experimentelle Psychologie. Beiträge zur 41. Tagung experimentell arbeitender Psychologen* (pp. 283-284). Lengerich: Pabst Science Publishers.
- Knauff, M., Jola, C., Strube, G., Rauh, R., & Schlieder, C. (2000). Visuo-spatial working memory involvement in spatial thinking. In preparation.
- Knauff, M., Rauh, R., & Schlieder, C. (1995). Preferred mental models in qualitative spatial reasoning: A cognitive assessment of Allen's calculus. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 200-205). Mahwah, NJ: Lawrence Erlbaum Associates.
- Knauff, M., Rauh, R., Schlieder, C., & Strube, G. (1998). Mental models in spatial reasoning. In Freksa, C., Habel, C., & Wender, K. F. (Eds.), *Spatial Cognition – An interdisciplinary approach to representing and processing spatial knowledge* (pp. 267- 291). Lecture Notes in Computer Science, Bd. 1404. Berlin: Springer.
- Kosslyn, S.M. (1980). *Image and mind*. Cambridge, MA: Harvard University Press.
- Kosslyn, S. M. (1994). *Image and brain*. Cambridge, MA: MIT Press.
- Logie, R.H. (1995). *Visuo-spatial working memory*. Hove: Lawrence Erlbaum.

- Mynatt, B. T. & Smith, K.H. (1977). Constructive processes in linear ordering problems revealed by sentence study times. *Journal of Experimental Psychology: Human Learning and Memory*, 3, 357-374.
- Newstead, S.E., Pollard P., & Griggs, R. A. (1986). Response bias in relational reasoning. *Bulletin of the Psychonomic Society*, 2, 95-98.
- Pearson, D. G., Logie, R., H., & Gillhooly, K.J. (1999). Verbal representations and spatial manipulation during mental synthesis. *European Journal of Experimental Psychology*, 11, 295-314.
- Potts, G.R. & Scholz, K.W. (1975). The internal representation of a three-term series problem. *Journal of Verbal Learning and Verbal Behavior*, 14, 439-452.
- Richardson, J.T.E. (1987). The role of mental imagery in models of transitive inference. *British Journal of Psychology*, 78, 189-203.
- Rinck, M. Hähnel, A, Bower, G. & Glowalla, U. (1997). The metrics of spatial situation models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 622-637.
- Schlieder, C. (1999). The construction of preferred mental models in reasoning with interval relations. In G. Rickheit & C. Habel (Eds.), *Mental models in discourse processing and reasoning*. Amsterdam: North-Holland.
- Shepard, R.N. & Cooper, L.A. (1982). *Mental images and their transformations*. Cambridge, MA: MIT Press.
- Stein, L. A. (1995). Imagination and situated Cognition. *Journal of Experimental and Theoretical Intelligence*.
- Sternberg, R. J., (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology: General*, 109, 119-159.

# Constraints of Embodiment on Action Coordination

**Günther Knoblich (knoblich@mpipf-muenchen.mpg.de)**

Cognition and Action; Max Planck Institute for Psychological Research; Amalienstr. 33  
80799 Munich, Germany

**J. Scott Jordan (jordan@sxu.edu)**

Department of Psychology; Saint-Xavier-University; 3700 West 103<sup>rd</sup> Street  
Chicago, IL 60655 USA

## Abstract

One consequence of the embodiment of cognition is that a single cognitive system may use fast internal mechanisms to coordinate conflicting actions in real time performance. In contrast, two different cognitive systems engaged in joint action have to resolve similar conflicts via the environment. A tracking paradigm was used to investigate the coordination of conflicting actions in individuals and groups. The main question was whether and how persons engaged in joint action would exploit the perceivable environmental outcomes of their partner's actions to adjust their own actions with respect to a jointly desired state. Groups performed worse than individuals, initially, but they achieved the same level of performance after some training. Groups improved because conflicting results of the partner's actions were taken into account when members of the pair produced their own actions. This led to the emergence of an agreed-upon environmental location, around which, group members coordinated their action effects. The results are consistent with the view that the special requirements of social interaction may have fostered the development of higher cognitive functions.

## Varieties of Embodiment

During the past decade, more and more researchers have become interested in the notion of embodied cognition (A. Clark, 1997; Port & van Gelder, 1995; Varela, Thompson, & Rosch, 1991). This approach has arisen largely out of dissatisfaction with the earlier notion of a central, disembodied symbol-manipulation system that is buffered from the environment via sensorimotor systems. In contrast, the Embodied approach stresses both, the importance of sensorimotor processes in cognitive functioning, and the close, dynamically supportive couplings that exist at all times between organisms and their environments.

Despite their common ground, different versions of the Embodied approach take issue with different aspects of the symbolic approach. In its most radical form, which is advocated by proponents of Dynamical Systems Theory (Port & van Gelder, 1995; Thelen & Smith, 1994), embodied cognition constitutes a rejection of representationalism as a whole, and conceptualizes cognition in terms of dynamic organism-environment couplings, exclusively. Less radical versions also stress the dynamic, organism-environment couplings, yet retain the notion of internal representation in order to account for the fact that certain biological systems

are able to produce actions that are directed toward objects not currently present in the immediate environment (Ballard, Hayhoe, Pook, & Rao, 1997; A. Clark, 1997). Given that these representing systems are assumed to have emerged due to the possibilities they afforded action production, proponents of this version often claim cognitive functioning to be constrained in one way or another, by the functioning of the sensorimotor system (e.g. the formation of concepts, (Barsalou, 1999; Lakoff & Johnson, 1999)).

The present research takes this notion as its starting point, and addresses its implications for action coordination. This is because, to date, embodiment has focused primarily on the individual cognitive system and its continuous environmental couplings. Members of many species, however, especially humans, often engage in joint action with other members of their species. Though some may consider social interaction just another example of environmental interaction, it may be the case that special requirements of joint action placed certain constraints on action production. Such constraints may have served, historically, to shape the structures and processes that came to be embodied in evolving cognitive systems (Mead, 1934; Vygotsky, 1978). The present research addresses these constraints.

## Individual and Joint Coordination of Conflicting Actions

A major function of action control in an individual organism is to select proper actions to obtain a desired impact on the environment (Prinz, 1997). If there are conflicting action alternatives, some internal mechanism may resolve the conflict (Anderson, 1990) and the motor system can be adjusted according to the action selected.

The situation is quite different when the action alternatives are distributed across two different cognitive systems that are engaged in joint action. Following a definition by H. H. Clark (1996), by a joint action we denote an action "that is carried out by an ensemble of people acting in coordination with each other" (p. 3). This implies that the individuals in the ensemble try to achieve a common goal. However, the intention to achieve a common goal does not protect the ensemble from encountering conflicts, especially when each system has only one of many action alternatives at its disposal.

To illustrate, imagine a situation in which two people drive a car together on a straight road. They can neither see

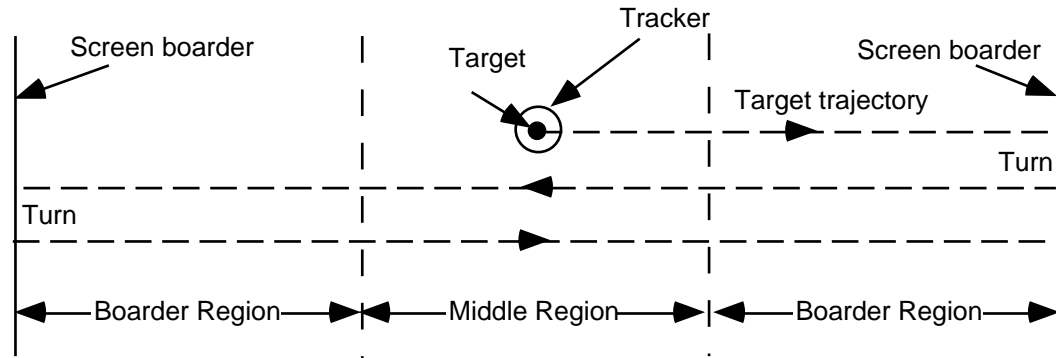


Figure 1: Illustration of experimental paradigm (vertical positions of target and tracker do not change in the actual task).

nor speak to one another. Person G controls the gas and person B, the brakes. As long as the car does not need to stop, there is no conflict, G acts and B does nothing. Now imagine a situation in which the ensemble encounters a traffic light. If the traffic light is green, no change is needed. If the traffic light is red, G has to stop acting and B has to start acting. Hence, this situation requires G and B to coordinate their actions according to an anticipated point at which they want to bring the car to a full stop. Conflicts may arise with respect to the point in time at which G stops and B starts acting, and as a consequence the ensemble may give gas and brake at the same time during a certain time interval. Hence, the car may well stop at a point that was intended neither by G, nor by B.

It is very unlikely that an individual in the same situation would carry out both actions at the same time even if different feet were used for giving gas and braking. An internal mechanism would select between the action alternatives in advance, instead of carrying out two conflicting actions at the same time. In the joint action example, conflict resolution is necessarily linked to noticeable changes in the environment, at least initially. Hence, if the situation requires braking, and B decides to start braking early, G will only know of that decision after perceiving that B has started to brake.

The aim of our research is to investigate how individuals and groups optimize their performance when conflicts arise in real time action coordination. Our main hypothesis is that persons engaged in joint action will use perceivable outcomes of the other's actions to dynamically adjust their own actions with respect to a commonly desired future state. Individual performance can be used as baseline to determine how the same conflict is dealt with within a single cognitive system.

### Experimental paradigm

We use a tracking paradigm for our studies. Generally, in tracking tasks one has to control a tracker so as to minimize the distance between the tracker and the target. The tracker is controlled by means of simple and clearly defined actions, e.g. hand movements or keypresses. The standard task requires minimal anticipation of future events and no conflict arises between alternative actions. For our study, we developed a different type of tracking task. In this task, an-

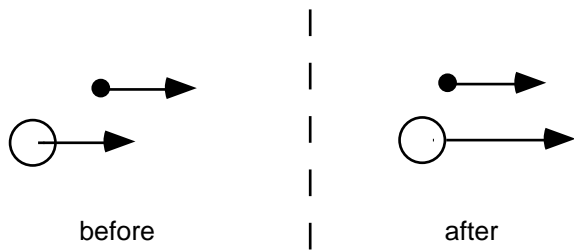
tipication of future events is crucial, and conflicts between two different action alternatives arise in a clearly defined manner. Figure 1 illustrates this task.

A target moves across the computer screen horizontally with constant velocity. As soon as it reaches a border of the screen, it changes its direction abruptly and moves back towards the other border, changes its direction again, and so on. The task is to keep a tracker on the target by controlling its velocity with two keys. When the tracker is moving to the right, hitting the right key accelerates it by a constant amount and hitting the left key decelerates it by the same amount. When the tracker is moving to the left, hitting the left key accelerates it and hitting the right key decelerates it. To illustrate, if the right key has been pressed five times, the left key will have to be pressed five times to bring the tracker to a full stop. Within the middle region, tracking performance can be optimized by decreasing the immediate error, as in most tracking paradigms. For instance, if the target is moving to the right and the tracker is left of the target, accelerating the tracker by a right keypress is the only reasonable action to be carried out (see panel a, in Figure 2).

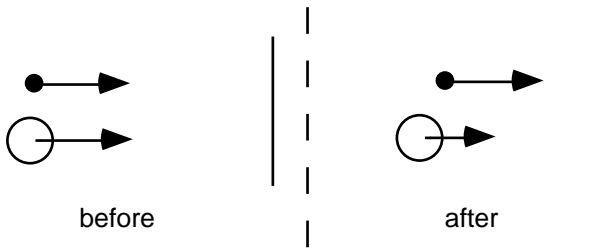
The situation is different within the border regions. In these regions, a conflict arises between two alternative strategies. The first alternative, i.e., trying to stay on target as long as possible, will minimize the immediate error up until the point at which the target changes its direction. Afterwards, a large error will arise because tracker velocity can only be changed gradually. Several keypresses will be needed to stop the tracker and more will be needed to gain velocity in the opposite direction. During this interval, the target will continue moving in the opposite direction, constantly increasing the distance between itself and the tracker. Thus, trying to minimize immediate error will create an extremely large future error.

The second alternative is to slow down the tracker before the target turns. In this case, the immediate error will be *increased* to prevent future error. This is the case because the target continues to move toward the border as the tracker decelerates with each keypress. Using the latter strategy is the only way to improve performance within the border region, especially when the target moves fast and the impact of each keypress on the velocity of the tracker is low. Within this context, we refer to keypresses that decrease immediate error as compensatory presses, and those that that

increase immediate error in order to reduce future error, as anticipatory brakes (see panel a, and b, in Figure 2).



a) Effect of a compensatory button press.



b) Effect of an anticipatory button press.

Figure 2: Illustration of the effects of compensatory and anticipatory keypresses (in the actual task target and tracker are horizontally aligned).

To investigate conflict resolution in individuals and groups we used two versions of the task that differed in one single aspect. In the individual condition each person controlled both keys, in the group condition each person controlled only one key. Hence, in the individual condition, the conflict between minimizing immediate vs. future error arises within *one* cognitive system, while in the group condition it arises between *two* cognitive systems. As a consequence, individuals may solve the conflict by use of fast internal mechanisms, while groups have to use certain aspects of the environment to act out the conflict overtly. Thus, in the group condition, the only way to better coordinate conflicting actions and thereby improve performance is to focus on changes in certain aspects of the environment that result from the other person's actions. Regularities in these changes can then be used to adjust one's own actions with respect to the commonly desired future state.

### Predictions

The nature of the present paradigm affords the measurement of several dependent variables that characterize performance, the extent and timing of the anticipatory strategy, and certain environmental anchors to which coordination can be linked. In the following, we will describe the rationale for using each of these variables, and derive predictions for the individual and group condition, in turn.

### Performance

To characterize performance, we use the absolute distance

between tracker and target at the time of each button press as an error measure. Our prediction is that the error should be lower for individuals, initially, because individuals coordinate conflicting actions by using fast internal mechanisms, whereas groups can only use perceivable changes in the environment, in their attempt to coordinate conflicting actions. Hence, groups should need more time to coordinate, which in turn should deteriorate real time performance, initially. However, if persons in a group are able to integrate some aspect of the environment that characterizes their partner's actions into their individual planning, the difference in error between individuals and groups should largely decrease.

### Extent and timing of anticipatory strategy

**Anticipatory brakes.** The extent to which the anticipatory strategy is employed within the boundary regions can be defined as the proportion of anticipatory brakes (see Figure 2, panel b) occurring in that region. We predict that the anticipatory brake rate will be greater for individuals than for groups, because once individuals have decided to prevent future error, they will be less likely to switch back to the conflicting action that reduces immediate error. In contrast, coordination requires overt action within groups. Therefore, the person who is responsible for reducing immediate error will quite likely produce actions that interfere with the anticipatory actions of the other person. The anticipatory brake rate should increase in both, individuals and groups, as they become more familiar with the task, because employing an anticipatory strategy is the only way to reduce overall error.

**Decision point.** One way in which the person responsible for anticipatory braking in a group can compensate for conflicting actions of the other is to take them into account, when timing her or his own actions. This should lead to earlier initiation of anticipatory braking in the group condition. The decision point, by which we denote the position of the tracker at the time of the first anticipatory brake, can be used to test this hypothesis. It should be further removed from the border in the group condition than in the individual condition.

### Environmental anchors

**Location of turn-around points.** By the turn-around point we denote the most extreme tracker location during each run of the target from one side of the screen to the other. If the target turns at the right border, the turn-around point is the maximal screen position of the tracker, if the target turns at the left border the turn-around point is the minimal screen position of the tracker (see Figure 1). To make turn-around points on both sides of the screen comparable they are expressed in terms of the absolute distance to the respective border. At the turn-around point the tracker comes to a full stop and is accelerated towards the other direction by the following keypresses. The turn-around point is functionally important because it can be used as an environmental anchor to which the goal of minimizing overall error can be tied. The reason is that, given a certain velocity of the target and a certain impact of each keypress, the optimal turn-around point will be relatively invariant.

If, as predicted, groups pick a decision point that is further removed from the boarder, groups may achieve a turn-around

point that is as equally removed from the border as the one achieved by individuals. Otherwise, it should be less removed from the border in the group condition. In the individual and the group condition as well, the turn-around point should become further removed from the border in later trials because overall error can be decreased by turning the tracker earlier.

**Homogeneity of turn-around points.** In the individual condition the turn-around points at the left and the right border are the result of actions taken by the same person. The situation is different in the group condition. Whenever the target approaches the right border, the person who is in charge of the left key is responsible for anticipatory braking and the person who is in charge of the right key is responsible for compensating immediate error. Whenever the target approaches the left border, each group member must assume the opposite role (the compensator becomes the anticipatory braker, the anticipatory braker becomes the compensator).

Hence, the prediction for individuals is that they will pick similar turn-around points at both borders. Therefore, the absolute difference between the left and the right turn-around point in a trial should be relatively small and not change substantially across consecutive blocks. In contrast, two persons in a group should pick more heterogeneous turn-around points initially. However, in later trials they may coordinate their actions by „agreeing“ on a certain turn-around point. Therefore, we predict a huge initial difference that substantially decreases in later blocks.

## Method

**Participants** Forty-five paid participants took part in the experiment. Fifteen participants were assigned to the individual condition. Thirty participants were assigned to the group condition.

**Material and Procedure** Upon entering the lab, participants were informed of the nature of the task. They were instructed individually in the group condition. Afterwards, they were seated in front of a computer monitor at a distance of 80 cm and were asked to put on a set of headphones. Participants in the group condition were divided by a partition. They could neither see one nor talk to one another. However, each was provided with a separate computer monitor, and all events taking place during the experiment (e.g. the movements of the tracker and the movements of the target) were presented simultaneously on both monitors. Thus, the only information shared was the task display and the acoustic feedback accompanying each keypress.

At the beginning of each trial target and tracker were displayed in the middle of the screen for 500 ms, the tracker being superimposed on the target. Thereafter, the target started moving either to the left or to the right with constant velocity. After reaching the border, it abruptly began traveling back in the opposite direction. There were three such target turns during each trial. The initial velocity of the tracker was zero. Each left keypress accelerated the tracker to the left and each right keypress accelerated it to the right. Right presses triggered a 600 Hz tone and left presses triggered a 200 Hz tone. Participants in the individual condition controlled both keys. In the group condition, each member

was given an individual control panel consisting of one key. Keypresses of the individual on the left side of the partition resulted in tracker acceleration to the left, while those of the other individual produced tracker-acceleration to the right. The experiment consisted of 3 blocks of 40 trials each.

## Results and Discussion

### Performance

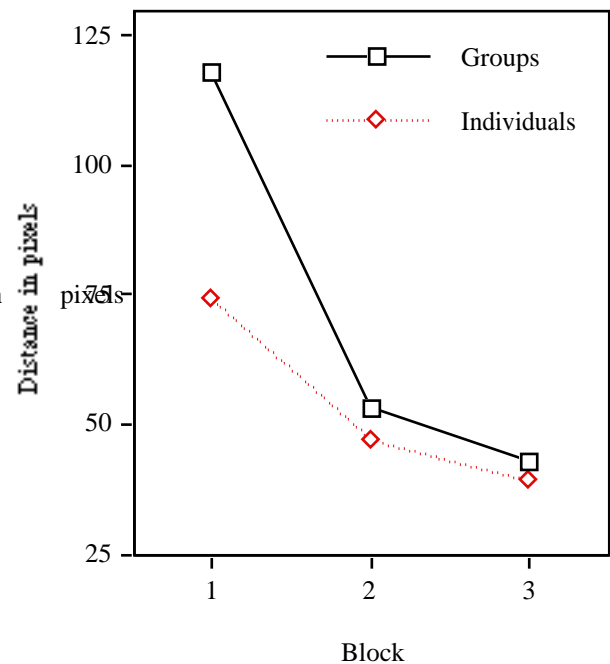


Figure 3: Individual and group performance across consecutive blocks

As can be seen in Figure 3, error decreased for individuals and groups across consecutive blocks. Hence, performance improved in individuals and groups. As expected, the error was much larger in the group condition during the first block. After the second block, group performance reached the level of individual performance. A 2 x 3 ANOVA with the factors Experimental Group (Individuals and Groups, between) and Block (1, 2, and 3, within) revealed a significant main effect for the Block factor,  $F(2, 56) = 24.2, p < .001$ , and a significant interaction between Experimental Group and Block,  $F(2, 56) = 3.5, p < .05$

### Anticipatory brakes

The anticipatory brake rate was computed as the number of anticipatory brakes in a border region divided by the overall number of button presses in that region. Figure 4 shows the results. The anticipatory brake rate increased over consecutive blocks for individuals and groups. As expected, the anticipatory brake rate was constantly lower in the group condition than in the individual condition.

A 2 x 3 ANOVA with the factors Experimental Group (Individuals and Groups, between) and Block (1, 2, and 3,

within) revealed significant main effects for the Group factor,  $F(1, 28) = 9.4, p < .01$ , and the Block factor,  $F(2, 56) = 26.1, p < .001$ . There was no significant interaction.

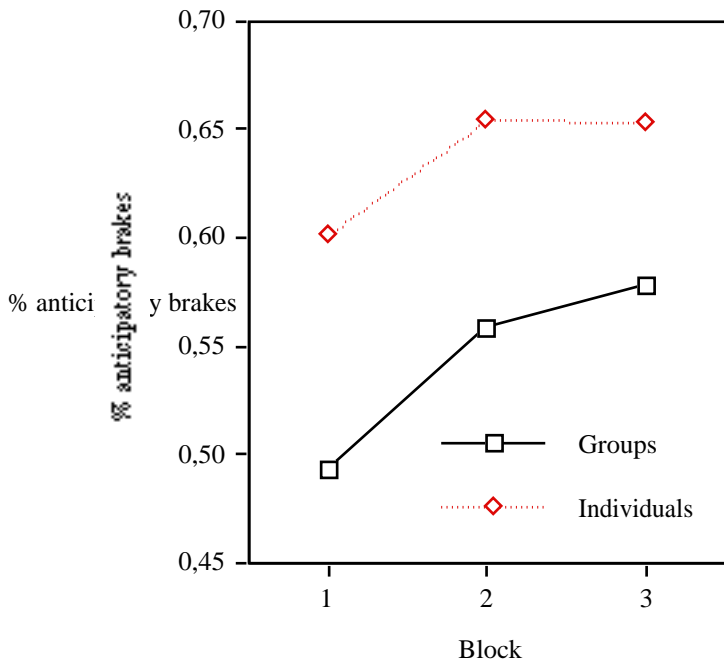


Figure 4: Anticipatory brake rate in individuals and groups across consecutive blocks

### Decision points

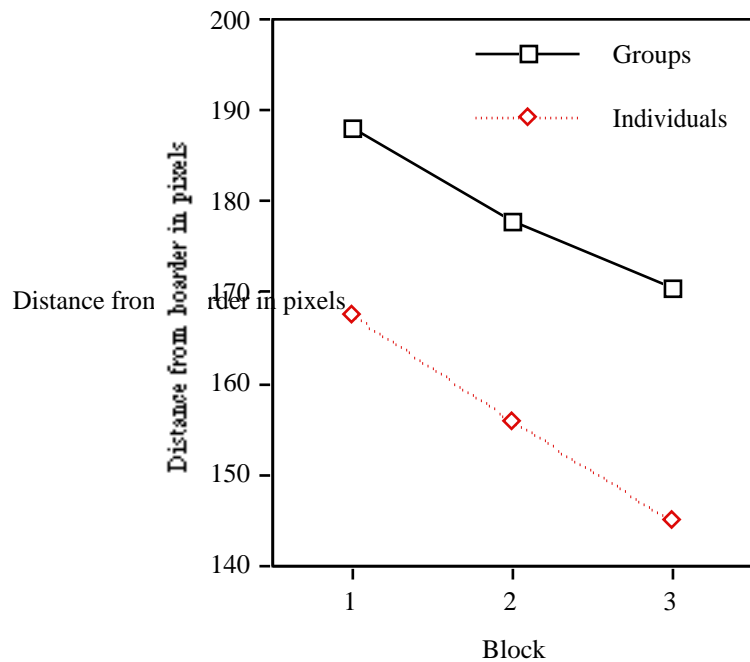


Figure 5: Decision point in individuals and groups across consecutive blocks.

Figure 5 shows the result of the analysis of decision points, i.e., the distance of the tracker from the border at the time of the first anticipatory brake.

As they became more familiar with the task, individuals and groups moved the tracker closer to the border before they initiated the first anticipatory brake. This result indicates that resolving the action conflict took less time in later trials. As expected, in the Group condition the tracker was always further from the border, when the first anticipatory brake occurred. A 2 x 3 ANOVA with the factors Experimental Group (Individuals and Groups, between) and Block (1, 2, and 3) revealed a significant main effect for the Group factor,  $F(1, 28) = 4.6, p < .05$ , and the Block factor,  $F(2, 56) = 11.2, p < .001$ . There was no significant interaction.

### Location of turn-around points

Figure 6 illustrates the results of the analysis of turn-around points, i.e., the absolute distance between the border and the point at which the tracker stopped before changing its direction.

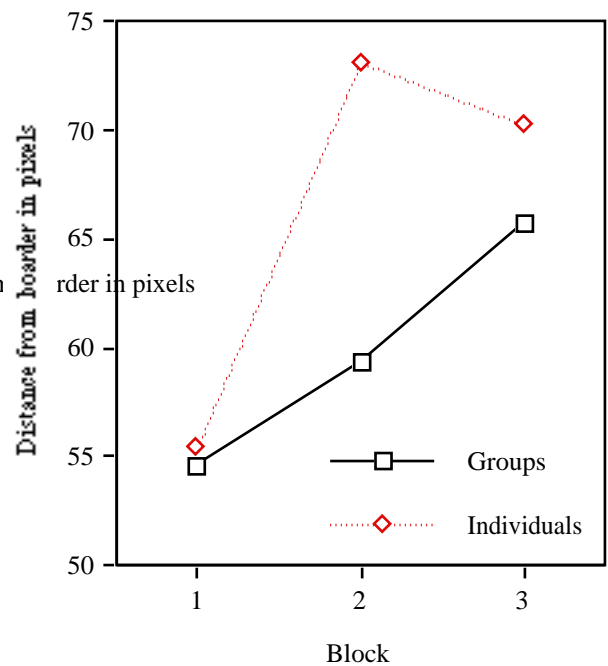


Figure 6: Turn-around point chosen by individuals and groups across consecutive blocks.

As expected, in later blocks, the turn-around point became further removed from the border in both experimental conditions. Individuals produced a sharper increase than groups from the first to the second block. A 2 x 3 ANOVA with the factors Experimental Group (Individuals and Groups, between) and Block (1, 2, and 3) revealed a significant main effect for the Block factor,  $F(2, 56) = 11.6, p < .001$ , and a marginally significant interaction,  $F(2, 56) = 2.62, p = .08$ . The difference between individuals and groups was highly significant during the second block,  $t = 4.21, p < .001$ . The main effect of experimental group was not significant.

## Homogeneity of turn-around points.

Figure 7 depicts the results of the analysis of the homogeneity of turn-around points.

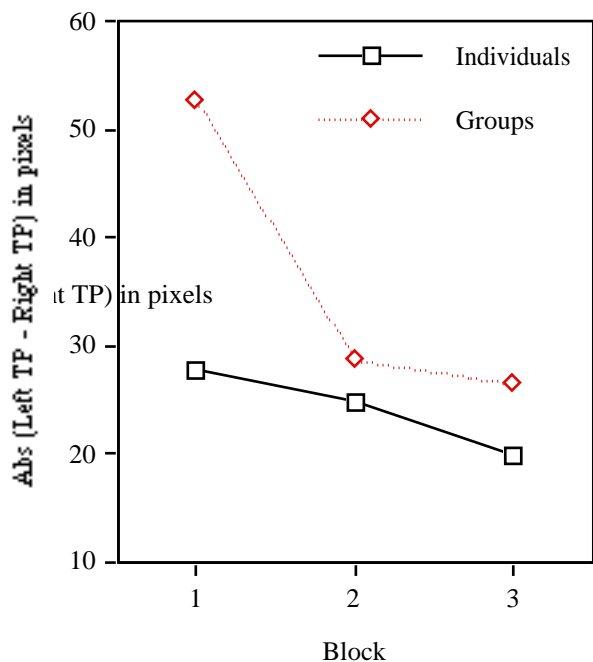


Figure 7. Homogeneity of turn-around points in individuals and groups across consecutive blocks.

Individuals turned the tracker at roughly the same point on both sides of the screen, i.e., there was only a small difference of about 30 pixels. The homogeneity of turn-around points increased only slightly across consecutive blocks. In contrast, persons in a group picked heterogeneous turn-around points, initially. In later trials, the selected turn-around points which were almost as homogeneous as those chosen by individuals. A 2 x 3 ANOVA with the factors Experimental Group (Individuals and Groups, between) and Block (1, 2, and 3) revealed a significant main effect for the Block factor,  $F(2, 56) = 8.25, p < .001$ , and a significant interaction,  $F(2, 56) = 3.30, p < .05$ . There was no significant main effect of experimental condition.

## Discussion

Individuals as well as groups are able to learn to coordinate conflicting actions with respect to a common goal, in real time, but groups clearly perform worse initially. The results illustrate robustly the different constraints that groups must deal with as they attempt to coordinate conflicting actions. To be sure, both groups and individuals improve by employing the advantageous anticipatory strategy. This is reflected in the fact that both gave rise to increases in anticipatory braking, as well as increases in the distance of the turn-around point from the border. Within groups however, this anticipatory strategy had to be worked out via the environment. Thus, it seems that group members take into account the potentially interfering actions of their partner by starting to brake at a further distance from the border. In addition, they seem to "agree" on a certain point in space at

which to turn the tracker, as is evidenced by increased homogeneity of the turn-around point. As soon as such an agreement has been reached, both the homogeneity of the turn-around points and the degree of error become almost indistinguishable from that produced by individuals.

The additional constraints on action coordination that arise within groups, as opposed to within an individual, are due to the fact that embodied cognitive systems have to make use of the environment to coordinate conflicting actions. This need to "lean" on the environment in group action, may constitute a selective pressure responsible for the phylogenetic emergence of cognitive systems capable of integrating the anticipated effects of another system's actions, into the planning of their own. This capability, in turn, may have afforded the emergence of the ability to produce environmental effects whose intended outcome was not solely entailed in the effect itself, but rather, in the impact that effect was anticipated to have upon the planning abilities of other cognitive systems. In short, the group need to collaborate through the environment may have driven the embodiment and environmental projection of symbol systems. This is consistent with Clarks (1996) assertion that the essence of language is joint action.

## Acknowledgements

We thank Rüdiger Flach for helpful comments, and Irmgard Hagen, Eva Seigerschmidt, and Patric Bach for their help in collecting the data.

## References

- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc.
- Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral & Brain Sciences*, 20(4), 723-767.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral & Brain Sciences*, 22(4), 577-660.
- Clark, A. (1997). The dynamical challenge. *Cognitive Science*, 21(4), 461-481.
- Clark, H. H. (1996). *Using language*. Cambridge, England UK: Cambridge University Press.
- Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh*. New York: Basic Books.
- Mead, G. H. (1934). *Mind, self and society: from the standpoint of a social behaviorist (Ed. with intro by C. W. Morris.)*. Chicago: University Press.
- Port, R. F., & van Gelder, T. (Eds.). (1995). *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA, USA: Mit Press.
- Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology*, 9(2), 129-154.
- Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA, USA: Mit Press.
- Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA, USA: Mit Press.
- Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.



# Modeling infant learning via symbolic structural alignment

**Sven E. Kuehne** ([skuehne@ils.nwu.edu](mailto:skuehne@ils.nwu.edu))

Department of Computer Science, Northwestern University  
1890 Maple Avenue, Evanston, IL 60201 USA

**Dedre Gentner** ([gentner@nwu.edu](mailto:gentner@nwu.edu))

Department of Psychology, Northwestern University  
2029 Sheridan Rd., Evanston, IL 60201 USA

**Kenneth D. Forbus** ([forbus@ils.nwu.edu](mailto:forbus@ils.nwu.edu))

Department of Computer Science, Northwestern University  
1890 Maple Avenue, Evanston, IL 60201 USA

## Abstract

Understanding the mechanisms of learning is one of the central questions of Cognitive Science. Recently Marcus et al. showed that seven-month-old infants can learn to recognize regularities in simple language-like stimuli. Marcus proposed that these results could not be modeled via existing connectionist systems, and that such learning requires infants to be constructing rules containing algebraic variables. This paper proposes a third possibility: that such learning can be explained via structural alignment processes operating over structured representations. We demonstrate the plausibility of this approach by describing a simulation, built out of previously tested models of symbolic similarity processing, that models the Marcus data. Unlike existing connectionist simulations, our model learns within the span of stimuli presented to the infants and does not require supervision. It can handle input with and without noise. Contrary to Marcus' proposal, our model does not require the introduction of variables. It incrementally abstracts structural regularities, which do not need to be fully abstract rules for the phenomenon to appear. Our model also proposes a processing explanation for why infants attend longer to the novel stimuli. We describe our model and the simulation results and discuss the role of structural alignment in the development of abstract patterns and rules.

## Introduction

Understanding the mechanisms of learning is one of the central questions of cognitive science. Recent studies (Gomez & Gerken, 1999; Marcus, Vijayan, Rao & Vishton, 1999) have shown that showed that infants as young as seven months can process simple language-like stimuli and build generalizations sufficient to distinguish familiar from unfamiliar patterns in novel test stimuli. In Marcus et al's study, the stimuli were simple 'sentences,' each consisting of three nonsense consonant-vowel 'words' (e.g., 'ba', 'go', 'ka'). All habituation stimuli had a shared grammar, either ABA or ABB. In ABA-type stimuli the first and the third word are the same: e.g., 'pa-ti-pa.' In ABB-type stimuli the second and the third word are identical: e.g., 'le-di-di'. The infants were habituated on 16 such sentences, with three repetitions for each sentence. The infants were then tested on a different

set of sentences that consisted of entirely new words. Half of the test stimuli followed the same grammar as in the habituation phase; the other half followed the non-trained grammar. Marcus et al. found that the infants dishabituated significantly more often to sentences in the non-trained pattern than to sentences in the trained pattern.

Based on these findings Marcus et al. proposed that infants had learned abstract algebraic rules. They noted that these results cannot be accounted for solely by statistical mechanisms that track transitional probabilities. They further argue that their results challenge connectionist models of human learning that use similar information, on two grounds: (1) the infants learn in many fewer trials than are typically needed by connectionist learning systems; (2) more importantly, the infants learn without feedback. In particular, Marcus et al. demonstrated that a simple recurrent network with the same input stimuli could not model this learning task.

In response, several connectionist models have attempted to simulate these findings. Unfortunately, all of them to date include extra assumptions that make them a relatively poor fit for the Marcus et al experiment. For example, Elman (1999; Seidenberg & Elman, 1999) use massive pre-training (50,000 trials) to teach the network the individual stimuli. More importantly, they turn the infants' unsupervised learning task into a supervised learning task by providing the network with external training signals. Other models tailored to capture the data of the study seem unlikely to be applicable to other similar cognitive tasks (Altmann & Dienes, 1999). Using a localist temporal binding scheme, Shastri and Chang (1999) model the infant results without pretraining and without supervision, but still require an order of magnitude more exposure to the stimuli than the infants received.

We propose a third alternative. There is evidence that structural alignment processes operating over symbolic structured representations participate in a number of cognitive processes, including analogy and similarity (Gentner, 1983), categorization (Markman & Gentner, 1993), detection of symmetry and regularity (Ferguson, 1994), and learn-

ing and transfer (Gentner & Medina, 1998). Although these representations and processes are symbolic, they do not need to be rule-like, nor need they involve variables. Instead, we view the notion of correspondence in structural alignment as an interesting cognitive precursor to the notion of variable binding<sup>1</sup>. Correspondences between structured representations can support the projection of inferences, as the analogy literature shows, and therefore a symbolic system can draw inferences about novel situations even without having constructed rules. Moreover, as discussed below, comparison can be used to construct conservative generalizations. Across a series of items with common structure such a process of progressive abstraction can eventually lead to abstract rule-like knowledge. The attainment of rules, in those cases where it occurs, is the result of a gradual process. As we will show, symbolic descriptions can be used with structural alignment to model learning that is initially conservative, but which occurs fast enough to be psychologically realistic.

We first describe our simulation model of the Marcus et al task, which uses a simple combination of preexisting simulation modules, i.e., SME, MAGI, and SEQL. All of these modules have been independently tested against psychological data and independently motivated in prior modeling work. With the exception of domain-specific encoding procedures, no new processing components were created for this task. We then describe the results of our simulation of the Marcus et al data, showing that our simulation can learn the concepts within the number of trials that the infants had, without supervision and without pre-learning. We also show that the simulation can exhibit the same results with noisy input data. Finally, we discuss some of the implications of the symbolic similarity approach for models of cognitive processing.

### Modeling infant learning via structural alignment

A psychological model of the infants' learning must include the kind of input, the way the infants are assumed to encode the individual sentences, and the processes by which they generalize across the sentences. The architecture of our simulation is shown in Figure 1. We first describe our assumptions concerning the infants' processing capacities. Then we describe each component in turn.

**Processing Assumptions:** We assume that infants can represent the temporal order within the sentences (Saffran, Aslin & Newport, 1996). We further assume that the infants notice and encode identities within the sentences: for example, the fact that the last two elements match in an ABB sentence. This assumption is consistent with evidence that human infants, as well as with studies of nonhuman primates (Oden *et al*, in press), can detect identities. We also assume that infants can detect similarities between sequentially presented stimuli, consistent with studies of infant habituation, which demonstrate that infants respond to sequential sameness (e.g., Baillargeon, 1994).

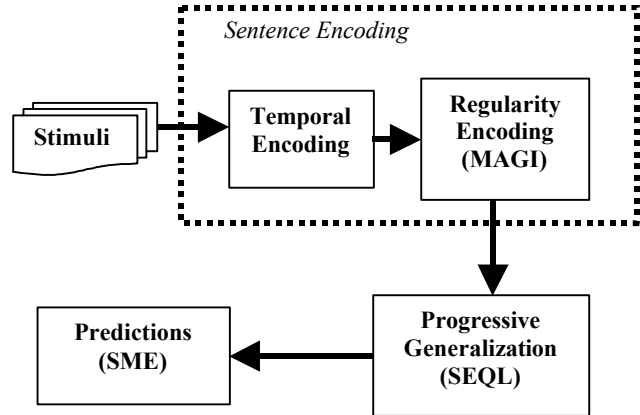


Figure 1: Simulation Architecture

**Input stimuli:** To make our simulation comparable with others, we use a representation similar to that of Elman (1999), namely, Plunkett & Marchman's (1993) distinctive feature notation. Each word has twelve phonetic features, which can be either present or absent. The presence or absence of each feature for each word is encoded by symbolic assertions. If feature  $n$  is present for word  $w$ , the assertion  $(Rn\ w)$  is included in the stimulus, and if absent, the assertion  $(Sn\ w)$  is included. Thus the acoustic features of each word are encoded as twelve attribute statements.

We modeled the Marcus et al experiment both without noise (Experiment 1) and with noise (Experiment 2). Marcus et al. used a speech synthesizer to control the pronunciation of the stimuli, but while this reduces variability, it cannot eliminate the possibility that the infant might encode something incorrectly.

**Temporal encoding:** We assume that the infant encodes the temporal sequence of the words in a sentence in two ways. First, each incoming word has an attribute associated with it, corresponding to the order in which it appears (i.e., FIRST, SECOND, or THIRD). We further assume that the infant encodes temporal relationships between the words in a sentence; to code this, an AFTER relation is added between pairs of words in the same sentence indicating their relative temporal ordering. The particular labels used in this encoding step are irrelevant – there are no rules in the system that operate on these specific predicates – the point is simply that infants are encoding the temporal order of words within sentences.

**Regularity Encoding:** We assume that the infants notice and encode identities within the sentences: for example, the fact that the last two elements match in an ABB sentence. Thus the simulation must incorporate a process that detects when words are the same. We use the MAGI model of symmetry and regularity detection (Ferguson, 1994) to automatically compute these relationships. MAGI treats symmetry as a kind of self-similarity, using a modified version of structure-mapping's constraints to guide the self-alignment process. MAGI has been successfully used with inputs ranging from stories to mathematical equations to visual stimuli,

<sup>1</sup> That structure-mapping algorithm neither subsumes, nor is subsumed by, traditional pattern matching such as unification is shown in Falkenhainer, Forbus, & Gentner (1988).

and it has done well at modeling certain aspects of visual symmetry, including making new predictions (Ferguson *et al* 1996). Here MAGI is used on the collection of words in a sentence. For any pair of words  $w1$  and  $w2$  that MAGI finds sufficiently similar, this module asserts (SIM  $w1$   $w2$ ), and a DIFF statement for every other pair of words in the sentence. (If MAGI does not find any pairs similar, DIFF statements are asserted for every pair of words.) This module also asserts (GROUP  $w1$   $w2$ ) for pairs of similar words, to mark that they form a substructure in the stimulus, and adds DIFF statements between groups and words not in the group. This use of MAGI is an example of what Ferguson (1994, in preparation) calls *analogical encoding*.

## SEQL

Once each sentence is encoded, we assume infants can detect the similarities between sequential pairs of sentences. The detection of structurally parallel patterns across a sequence of examples is modeled by SEQL (Skorstad, Gentner & Medin, 1988; Kuehne, Forbus, Gentner & Quinn, 2000), a model of the process of category learning from examples. SEQL constructs category descriptions via incremental abstraction. That is, the representation of a category is a structured description that has been generated by successive comparison with incoming exemplars. If the new exemplar and the category are sufficiently similar, the category description is modified to be their intersection -- i.e., the commonalities computed via structural alignment by a generalization algorithm. If the new exemplar is not sufficiently similar, it is stored separately and may later be used as the seed of a new category.

The structural alignment process is implemented via SME, (Falkenhainer *et al* 1988; Forbus *et al* 1994) a cognitive simulation of analogical matching. Here the base description is a category description, and the target description is the new exemplar. The structural alignments that SME computes are used in three ways by SEQL. First, the numerical structural evaluation score it computes<sup>2</sup> is used as a similarity metric, a numerical measure for deciding whether or not two descriptions are sufficiently similar. Second, the candidate inferences it computes serve as a model for category-based induction (c.f. Blok & Gentner, 2000; Forbus, Gentner, Everett, & Wu, 1997). Third, the correspondences in the best mapping SME produces serves as the basis for SEQL's generalization algorithm.

SEQL maintains a set of generalizations and a set of singular exemplars. When a new exemplar comes in, it is compared against existing generalizations to see if it can be assimilated into one of them. Otherwise, it is compared with the stored exemplars to see if a new generalization can be formed. If it is insufficiently similar to both the generalizations and the stored exemplars, it is stored as an exemplar itself.

SEQL begins with no generalizations; it simply stores its first exemplar. If the next exemplar is sufficiently close to the first, their overlap is stored as the first generalization. A

<sup>2</sup> Although SME can compute multiple mappings, we use the structural evaluation score of the best mapping, normalized by the size of the base description.

generalization consists of the overlap between the two input descriptions: that is, the shared structure found by alignment. Thus generalizations are structured descriptions of the same type as the input descriptions, although containing fewer specific features. If a new exemplar is sufficiently similar to a generalization (as determined comparing the structural evaluation score to a set threshold), then (a) the generalization is updated by retaining only the overlapping description that forms the alignment between the generalization and the exemplar; and (b) candidate inferences are projected from the generalization to the exemplar. Non-overlapping aspects of a description (e.g., phonetic features or relations that aren't shared) are thus "worn away" with each new assimilated description. (The threshold that determines when descriptions are sufficiently similar to be assimilated helps prevent descriptions from diminishing into vacuity.)

Returning now to the infant studies, we assume that babies are carrying out an ongoing process of comparing and aligning the incoming exemplars with an evolving generalization. We further assume that the relational candidate inferences from the general pattern to a new exemplar represent expectations on part of the infant.<sup>3</sup> When these expectations are violated by an incoming stimulus that does not fit the generalized pattern (e.g., an ABB test sentence after the ABA generalization has been formed), we assume the infant requires extra time to process the inconsistent stimulus.

## Simulation Experiments

In both experiments, we followed the procedure of Marcus *et al*. Each stimulus was a simple three-word sentence, encoded as described earlier. There were two sets of training stimuli, one following the ABA pattern and one following the ABB pattern. The training stimuli were (ABA) de-di-de, de-je-de, de-li-de, de-we-de, ji-di-ji, ji-je-ji, ji-li-ji, ji-we-ji, le-di-le, le-je-le, le-li-le, le-we-le, wi-di-wi, wi-je-wi, wi-li-wi, wi-we-wi and (ABB) de-di-di, de-je-je, de-li-li, de-we-we, ji-di-di, ji-je-je, ji-li-li, ji-we-we, le-di-di, le-je-je, le-li-li, le-we-we, wi-di-di, wi-je-je, wi-li-li, wi-we-we. The test stimuli in both experiments were four descriptions representing two novel ABA-type (ba-po-ba, ko-ga-ko) and two novel ABB-type sentences (ba-po-po, ko-ga-ga). The threshold value for SEQL was set to 0.85 in both experiments.

### Experiment 1

This experiment is most comparable to previous simulation models of the phenomena, in that we assume noise-free encoding of the stimuli. A simulation run consists of exposing SEQL to all of the stimuli from a particular training set (either ABA or ABB) once and then seeing the response given the four test sentences. To avoid possible biasing due to sequence effects (See Kuehne *et al.*, 2000), 20 simulation runs were made for each training set using different random

<sup>3</sup> SME can also produce attribute-level candidate inferences, and does so on these stimuli. We assume that, since these inferences concern directly perceivable features, testing them takes very little time.

orders. Identical match score and relational candidate inferences were produced for all sequences with a given stimulus set. In each case, SEQL produced a single generalization during the learning phase. For the test phase we used encodings of the corresponding stimuli used with infants, as noted above. Tables 1a and 1b show the results of this series for two generalizations paired against the four test sentences.

**Table 1a: ABA training stimuli**

Test Stimulus	Match Score	Candidate Inferences
<b>Ba-po-ba</b>	<b>0.658</b>	<b>None</b>
<b>Ko-ga-ko</b>	<b>0.689</b>	<b>None</b>
Ba-po-po	0.486	(DIFF po1 ba1) (DIFF po1 po2) (SIM ba1 po2)
Ko-ga-ga	0.455	(DIFF ga1 ko1) (DIFF ga1 ga2) (SIM ko1 ga1)

**Table 1b: ABB training stimuli**

Test Stimulus	Match Score	Candidate Inferences
Ba-po-ba	0.328	(SIM po1 ba2) (DIFF ba1 (GROUP po1 ba2))
Ko-ga-ko	0.350	(SIM ga1 ko2) (DIFF ko1 (GROUP ga1 ko2))
<b>Ba-po-po</b>	<b>0.776</b>	<b>None</b>
<b>Ko-ga-ga</b>	<b>0.753</b>	<b>None</b>

The in-grammar (bold) and out-of-grammar (plain text) matches show clear differences in their match scores. In-grammar matches are above 0.64 and do not generate relational candidate inferences. Out-of-grammar matches have match scores below 0.5, and lead to relational candidate inferences. Thus out-of-grammar test sentences lead to longer looking behavior, as predicted.

## Experiment 2

As noted earlier, we believe that noise-free stimulus encodings are unrealistic. Consequently, we used the same procedure as Experiment 1, but this time introducing noise into the representations for the training and test stimuli. For each sentence, one of the words was randomly picked, and one of its attributes (also chosen at random) was dropped or flipped, with the rest of its description being unchanged. Such changes can be significant: for example, flipping a single phonetic feature turns the word ‘de’ into the word ‘di’. Again, 20 simulation runs were made for each training set using different random orders. Naturally the match scores and, to a lesser degree, the generated candidate inferences, did vary across the individual runs. Tables 2a and 2b show the results. The scores were averaged over all 20 runs.

Although the noise affected the details of the computations, the overall pattern of results remains the same. The in-grammar (bold) match scores are far higher than the out-of-grammar (plain text) scores; and the out-of-grammar

stimuli produce relational candidate inferences while the in-grammar stimuli do not.

**Table 2a: ABA training stimuli**

Test Stimulus	Average Match Score	Candidate Inferences Min, Average, Max
<b>ba-po-ba</b>	<b>0.647</b>	<b>0, 0, 0</b>
<b>ko-ga-ko</b>	<b>0.682</b>	<b>0, 0, 0</b>
ba-po-po	0.435	2, 2.45, 3
ko-ga-ga	0.395	2, 2.55, 3

**Table 2b: ABB training stimuli**

Test Stimulus	Match Score	Candidate Inferences Min, Average, Max
ba-po-ba	0.339	2, 2, 2
ko-ga-ko	0.352	2, 2.05, 3
<b>ba-po-po</b>	<b>0.805</b>	<b>0, 0, 0</b>
<b>ko-ga-ga</b>	<b>0.783</b>	<b>0, 0, 0</b>

## Comparison with other models

The results of Marcus et al. (1999) have sparked an active debate focused on two issues: (1) Can current connectionist models (e.g., simple recurrent networks) model these results? (2) Do infants generate abstract rules that include variables?

Regarding the adequacy of simple recurrent networks, Marcus et al. state “Such networks can simulate knowledge of grammatical rules only by being consequently trained on all items to which they apply; consequently, such mechanisms cannot account for how humans generalize rules to new items that do not overlap with the items that appeared in the training.” Elman’s (1999) response describes his use of a simple recurrent network to model this task. Elman’s model requires tens of thousands of training trials on the individual syllables, and treats the problem as a supervised learning task, unlike the task facing the infants. By contrast, our simulation handles the learning task unsupervised, and produces human-like results with only exposure to stimuli equivalent to that given to the infants. Moreover, our model also continues to work with noisy data, something not true of any other published model of this phenomenon that we know of.

The learning in our model is due to the “wearing away” of non-identical phonetic attributes through subsequent comparisons. Although SEQL’s learning proceeds faster than connectionist models, it is still slower than systems that generate abstractions immediately (e.g., explanation-based learning (DeJong & Mooney, 1986)). In SEQL’s progressive alignment algorithm, the entities in the generalizations lose their concrete attributes across multiple comparisons, leaving the relational pattern of each grammar as the dominant force in the generalization only after a reasonable num-

ber of varied examples are seen.<sup>4</sup> There is considerable evidence for this kind of conservative learning (Forbus & Gentner, 1986; Medin & Ross, 1989).

Turning to the second issue, whether infants have variables and generate abstract rules, Marcus et al (1999) claims “[I]nfants extract abstract algebra-like rules that represents relationships between placeholders (variables), such as ‘the first item X is the same as the third item Y,’ or more generally that ‘item I is the same as item J.’” But our simulation does not introduce variables, in the sense commonly used in mathematics or logic. The generalizations constructed by SEQL do indeed include relational patterns that survive repeated comparisons because they are shared across the ingrammar exemplars. Furthermore, the entities (words) in the generalizations have many fewer features than the original words, as a result of the wearing away of features in successive comparisons. One could consider these patterns as a form of psychological rule, as proposed by Gentner and Medina (1998), with the proviso that the elements in the rule are not fully abstract variables, although they might asymptotically approach pure variables.

## Discussion

This paper proposes a third kind of explanation for the infant learning phenomena of Marcus et al (1999): incremental abstraction of symbolic descriptions via structural alignment. We believe our explanation is currently the best one for three reasons. First, it models the infant data with fewer extra concessions than previously published models (i.e., no pre-training, no supervision, and noisy data). Second, the processes we postulate are cognitively general; they apply to a large set of phenomena. Third, the abstraction processes we propose are consistent with research demonstrating that human learning is initially conservative (Brooks, 1987; Forbus & Gentner, 1986; Medin & Ross, 1989). Interestingly, there is ongoing research in developing symbolic connectionist models consistent with these processes (e.g., Holyoak & Hummel, 1997).

Many issues remain to be explored. For example, although our system does not introduce variables in its generalization process, there is a sense in which the entities in the generalization are on their way to becoming variables. Gentner and Medina (1998) have proposed that the process of progressive alignment can lead to rules. They further suggested that the application of rules to instances can be accomplished using the same general processes of structural alignment and projection that are used in analogy. The difference is that the base domain is an abstraction, the entities are ‘dummies’ with no features to either help or impede the match with the specific entities in the exemplar. Another issue concerns the incorporation of statistical notions in SEQL. Although SEQL is to a certain degree noise-resistant,

---

<sup>4</sup> SEQL learns with only one exposure to the 16 learning sentences, whereas Marcus’s infants received three exposures for each sentence. It is possible that the infants would have learned with only one pass; however it is also possible that the infants were less consistent in detecting the similarities than our simulation with its current parameters.

we suspect that to model large-scale learning, it will need to keep track of more statistical information than it does currently, so that properties wear away more slowly.

We note that it is common to conflate symbolic processing with rule-based behavior, and parallel processing with connectionist models. The model described here is symbolic, but it need not involve variables or rules. Further, it involves extensive parallel processing (most of SME and MAGI’s computations are parallel). Given the complexity of the phenomena, such confluations seem unwise.

The debates stirred by the Marcus et al. results bear on a critical issue in human learning and development: namely, what knowledge or mechanisms must be assumed to account for the rapid and powerful achievements demonstrated by infants in both cognition and language. Our results suggest that the general learning mechanism of structure-mapping theory may go a long way in accounting for these accomplishments.

## Acknowledgments

We thank Ron Ferguson, Ken Kurtz and Tom Mostek for valuable help and discussions. This research was supported by the Cognitive Science Division of the Office of Naval Research.

## References

- Altmann, G.T.M. and Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks, *Science* 284, 875.
- Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9(4), 321-324.
- Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science*, 3(5), 133-140.
- Blok, S. V., & Gentner, D. (2000). Reasoning from shared structure. *Proceedings of the 22<sup>nd</sup> Meeting of the Cognitive Science Society*.
- Brooks, L. R. (1987). Decentralized control of categorization: The role of prior processing episodes. In U. Neisser (Ed.), *Concepts and conceptual development: The ecological and intellectual factors in categorization* (pp. 141-174). Cambridge: Cambridge University Press.
- Christiansen, M.H. and Curtin, S.L. (1999). The power of statistical learning: No need for algebraic rules, in Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society, Erlbaum, Mahway, NJ.
- Christiansen, M.H. and Curtin, S.L. (1999). Transfer of learning: rule acquisition or statistical learning? *Trends in Cognitive Science* 3, 289-290
- DeJong, G.F. and Mooney, R.J. (1986). Explanation-based learning: An alternative view. *Machine Learning* 1(2), pp. 145-176
- Elman, J. (1999). Generalization, rules, and neural networks: A simulation of Marcus et. al, (1999). Ms., University of California, San Diego.
- Falkenhainer, B., Forbus, K., and Gentner, D. (1986). The Structure-Mapping Engine. In: *Proceedings of AAAI 86*, Philadelphia, PA, August.

- Falkenhainer, B., Forbus, K.D. and Gentner, D. (1989). The Structure Mapping Engine: an algorithm and examples. *Artificial Intelligence*, 41: 1-63
- Ferguson, R.W. (1994). MAGI: A model of analogical encoding using symmetry and regularity. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Ferguson, R.W., Aminoff, A. and Gentner, D. (1996). Modeling qualitative differences in symmetry judgments. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Forbus, K. D., & Gentner, D. (1986). Learning physical domains: Toward a theoretical framework. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2, pp. 311-348). Los Altos, CA: Kaufmann.
- Forbus, K. D., Ferguson, R. W., and Gentner, D. (1994). Incremental Structure-mapping. In: *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cognitive Science*, 7: 155-170.
- Gentner, D. and Markman, A.B. (1997). Structure-mapping in analogy and similarity. *American Psychologist*, 52, 45-56.
- Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, 65, 263-297.
- Goldstone, R.L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition* 52(2), 125-157.
- Goldstone, R.L., Medin, D.L., and Gentner, D. (1991). Relational similarity and the non-independence of features in similarity judgements. *Cognitive Psychology*, 23, 22-264.
- Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition* 70,109-135.
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3), 427-466.
- Kuehne, S.E., Forbus, K.D., Gentner, D. and Quinn, B. (2000). SEQL- Category learning as incremental abstraction using structure mapping, *Proceedings of the Twenty-second meeting of the Cognitive Science Society*.
- Marcus, G.F., Vijayan, S., Bandi Rao, S. and Vishton, P.M. (1999). Rule-learning in seven-month-old infants. *Science*, Vol. 283, 77-80
- Marcus, G.F. (1999). Do infants learn grammar with algebra or statistics?, Response to Seidenberg & Elman, Negishi, and Eimas. *Science* 284, 436-437
- Marcus, G.F. (1999). Simple recurrent networks and rule-learning: <http://psych.nyu.edu/~gary/science/es.html>.
- Markman, A.B. and Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, 25, 431-467.
- McClelland, J.L. and Plaut, D.C. (1999). Does generalization in infant learning implicate abstract algebraic rules?, *Trends in Cognitive Science* 3, 166-168
- Medin, D.L., Goldstone, R., and Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100(2), 254-278.
- Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem-solving, and induction. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 5, pp. 189-223). Hillsdale, NJ: Erlbaum.
- Oden, D. L., Thompson, R. K. R., and Premack, D. (in press). Can an ape reason analogically? Comprehension and production of analogical problems by Sarah, a chimpanzee (Pan troglodytes). In D. Gentner, K. J. Holyoak, & B. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science*. Cambridge, MA: MIT.
- Plunkett, K. and Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, 48, 21-69.
- Saffran, J., Aslin, R. and Newport, E. (1996). Statistical learning by 8-month-old infants, *Science*, 274, 1926-1928
- Seidenberg, M.S. and Elman, J. (1999), Do infants learn grammar with algebra or statistics?, *Letter, Science* 284, 434-436
- Seidenberg, M.S. and Elman, J. (1999). Networks are not hidden rules, *Trends in Cognitive Science* 3, 288-289
- Skorstad, J., Gentner, D. and Medin, D. (1988). Abstraction processes during concept learning: a structural view. In: *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Montreal: Lawrence Erlbaum Associates.

# Learning-Based Constraints on Schemata

Peter C.R. Lane (pcl@psychology.nottingham.ac.uk)  
Fernand Gobet (frg@psychology.nottingham.ac.uk)  
Peter C-H. Cheng (pcc@psychology.nottingham.ac.uk)  
ESRC Centre for Research in Development, Instruction and Training,  
School of Psychology, University of Nottingham,  
University Park, NOTTINGHAM NG7 2RD, UK

## Abstract

Schemata are frequently used in cognitive science as a descriptive framework for explaining the units of knowledge. However, the specific properties which comprise a schema are not consistent across authors. In this paper we attempt to ground the concept of a schema based on constraints arising from issues of learning. To do this, we consider the different forms of schemata used in computational models of learning. We propose a framework for comparing forms of schemata which is based on the underlying representation used by each model, and the mechanisms used for learning and retrieving information from its memory. Based on these three characteristics, we compare examples from three classes of model, identified by their underlying representations, specifically: neural network, production-rule and symbolic network models.

## Introduction

One of the unifying themes in cognitive science is the use of *schemata* for explaining the units of knowledge within humans. However, the specific properties which comprise a schema usually vary between authors. Early work in the AI and cognitive traditions (e.g. Rumelhart, 1980) set the scene for the use of schemata in computational models of learning. It is now appropriate, with a number of successful models in the literature, to see what forms of schemata arise within a learning-based system. This question is especially interesting because computational models do not simply implement basic concepts such as schemata with an added learning mechanism. Instead, each computational model is based on some core representational structure and primitive learning mechanisms, from which structures such as schemata may be inferred.

The aim of this paper is to consider examples from a number of computational models and simply extract those elements which most relate to schemata. The difficulty here is that the models have not been tested on identical tasks, and so the comparison must be at a more qualitative level. Hence, we begin with some informal definitions of schemata to define our analytical framework.

## Learning and Using Schemata

Brewer (1999) defines schemata as “the psychological constructs that are postulated to account for the molar forms of human generic knowledge.” The idea is that knowledge of visual scenes or discourse structure may be considered in terms of basic units. For instance, house-scenes typically consist of rooms, each room containing certain basic properties, such as walls or furniture. The schema for a room will contain *slots* for the properties, and, in the absence of spe-

cific information, these slots will be filled with *default* values. So a room will, by default, be considered to have four walls, a ceiling, a door, lighting, probably a window, and so forth.

Less committal is the definition by Rumelhart (1980; italics in original): “A schema theory is basically a theory about knowledge. It is a theory about how knowledge is represented and about how that representation facilitates the *use* of the knowledge in particular ways.” Rumelhart therefore focuses on the form of the schema theory (representation and reuse), whereas Brewer (1999) defines the form of the schema (a molar form of knowledge). Rumelhart’s definition is also echoed in that of Sweller (1988), whose concern is with modelling problem-solving behaviour. According to Sweller (1988), a schema is simply a “structure which allows problem solvers to recognize a problem state as belonging to a particular category of problem states that normally require particular moves. ... certain problem states can be grouped, at least in part, by their similarity and the similarity of the moves that can be made from those states.” Each of these definitions stresses the functionality of the knowledge in the schema. Also worth noting is that the schema is a form of retrieval structure, identifying elements from earlier experience which can be reused in the current situation.

Our interest in this paper is in describing computational models of learning, and for this purpose, as will become evident later, a fairly loose definition of schemata is required to provide the basis of comparison between different models. Hence, we will use the following definition:

*A schema is a cognitive structure for representing and retrieving classes of typical situations for which a similar response is required of the learner.*

Our comparison looks at the variation in schema-form based on the different assumptions underlying each model. The greatest assumption made is the basic *representation* used by the model for storing learnt information in its memory. This representation may be highly structured, localised or distributed. The type of the representation affects the processes which the model can use to *learn*, where learning is the process of converting what has been experienced into an internal representation. In this context, some representations provide better support for incremental real-time learning, whereas others are better for complex rule induction. The type of representation also affects the *retrieval* of information from the model’s memory for use in novel situations. Some systems assume that every item of memory is compared to determine the closest match to the current situation, whereas others maintain a hierarchy for indexing

their memory and consequently only search a subset of the total memory.

These three characteristics, for representing, learning and retrieving a schema, provide a framework for analysing how different computational models address the questions of learning and using schemata. We use this framework in the next three sections, where we compare examples from three classes of model. The classes are distinguished by their underlying representations: neural network, production-rule and symbolic network models. The examples are selected to be representative (without attempting to be comprehensive).

### Neural network models

The ability of a PDP (Parallel Distributed Processing) model (otherwise known as a neural network) to learn schemata was addressed at an early stage by Rumelhart, Smolensky, McClelland and Hinton (1986), who described how such properties can arise within a class of PDP models. However, they did not address the question of learning. A better demonstration of these ideas within the context of a learning system is the Sentence Gestalt (SG) model of St. John and McClelland (1990). We also consider the CLARION system of Sun, Merrill and Peterson (in press), which is a hybrid model of skill learning.

#### Sentence comprehension

The aim of the SG model (St. John & McClelland, 1990) is to capture the process by which people fill out semantic information whilst reading a sentence. For example, given the sentence 'Bobby pounded the board together with nails', the inference "with a hammer" is made automatically. We can explain such behaviour by hypothesising that people recall (subconsciously) some schema for the sentence from which default information (the hammer) can be inferred. The SG model attempts to account for such phenomena. It consists of a two-stage recurrent neural network. The first stage learns a distributed representation for the sentence, called the sentence gestalt, from a temporal sequence of constituents. Each constituent is either a simple noun phrase, a prepositional phrase or a verb. The second stage acts as a probe for information contained in the sentence gestalt. Each probe is a role/filler pair, and the sentence gestalt is probed by presenting either a role or a filler, from which the network is to supply the complete pair. Requested information need not refer directly to words in the sentence. For example, after seeing 'Mary ate the spaghetti', the model should return the filler "fork" for the role "instrument".

The experiments performed by St. John and McClelland demonstrate that the SG model successfully assigns constituents to thematic roles based on syntactic and semantic constraints. Further, the model can disambiguate meanings and instantiate vague terms as appropriate to their context and the training data previously seen by the model. This behaviour fulfills the requirements for schemata as discussed previously: the model classifies sentences into various groups, and these groups can have variable or default information associated with them.

We can now consider the schemata used in SG against the three basic characteristics of our framework:

#### Representing a schema

All knowledge contained within a neural network is held implicitly across the weights within the network. Once activation is presented on the input, every weight and node within the network interact to generate an output. In this situation, specific schemata are not really *represented* within the network, in the sense of identifiable units, but instead *emerge* as a consequence of the specific set of inputs. Hence, the schemata used by the model cannot be extracted for use as explicit rules, but instead must be inferred from their effects on the network's output.

#### Learning a schema

Given the nature of distributed representations, it is not possible to learn about just one schema, because of the unpredictable effect on other information held in the weights. Indeed, the process by which the SG model (and most similar neural network models) is trained involves continuous passes of the entire training dataset whilst the weights in the network are gradually altered to approximate the mapping between the input data and its target output. This process means that the network captures generalisations true of the entire dataset, making it robust in novel situations.

#### Retrieving a schema

Again, the nature of the distributed representation within the model implies that the whole network is activated when obtaining a response to a novel input. Hence every piece of acquired information (every weight value) is used in generating a response. This process additionally ensures a robust response in novel but similar situations, because the retrieval process is based on the *similarity* between the novel input and the model's previous experience. For instance, if a large number of examples are presented to the network, and the responses analysed, it will be seen that those examples which are most similar tend to generate similar responses. Conversely, if a novel input is partly similar to one type of example in the training data, and partly similar to another type, the computed response will fall somewhere between that for the two items of training data. Note that the similarity in input to the network is heavily dependent on the form of encoding used for representing each item of data to the network on numeric input units.

#### Bottom-up skill learning

CLARION (Sun, Merrill & Peterson, in press) is a hybrid model for bottom-up skill learning. It is designed to model the process by which low-level perceptual-motor skills are converted into explicit rules, and also capture the interaction between these two levels of knowledge whilst carrying out a complex task. CLARION assumes that declarative knowledge is represented explicitly within a rule-based system, whereas procedural knowledge is represented implicitly within a neural network. CLARION has been tested in a perceptual-motor task involving navigation through a minefield, in which the model must learn to react to particular visual patterns of mines with appropriate navigation instructions to avoid the mines and reach a target. The dual use of knowledge is reflected in subjects' responses: mostly they



react instinctively, but after some experience in the domain some explicit planning is reported. CLARION's use of two knowledge levels is intended to capture this shift towards more explicit knowledge.

The novelty in CLARION is that the rules can either be pre-programmed (i.e. taught in the standard top-down manner) or learnt based on the low-level knowledge in the neural network. Specifically, if the neural network suggests an action which satisfies its criterion for success, then the current sensory state is turned into the condition part of a new production in the rule set, with its action part being the currently suggested action. Further learning processes on the rules update statistics and may refine and alter rules for efficiency. CLARION therefore contains two independent learning mechanisms, but the two can also work together with an interesting transfer of bottom-up (procedural) knowledge into the explicit rule-set. As with SG, schemata are evident in the similarity-based generalisations made by the model.

#### *Representing a schema*

CLARION uses a two-level representational structure: a rule-based system and a feed-forward neural network. As with SG, schemata are seen to emerge through the interaction of many elements in the model. Hence, the network and the rules can generalise robustly to novel situations based on partial similarity. The purpose of the rule-based system is to 'fix' the generalisations learnt by the neural network and prevent later experience 'blurring' them. These rules may in themselves represent broader classes of situation, because some of the attributes can have variables instead of specific values, rather akin to slots on a more generic template.

#### *Learning a schema*

The procedural knowledge in CLARION is learnt in a similar manner to the SG model described above, using a modified form of backpropagation: an additional reinforcement term is included in the training error because the correctness of a specific action is only known at the completion of the task. The rule-based declarative knowledge includes mechanisms for constructing new rules, or expanding or shrinking the conditions of existing rules. The mechanism for constructing a new rule is merely to include, for a successful action, the situation and action as the condition and action parts of a new rule. Expanding or shrinking a rule's conditions amounts to increasing or decreasing the likelihood of the rule matching future inputs by altering the range of possible values in one of its attributes. Before making any such changes to a rule's conditions, an information gain for each rule is computed to determine whether a modified version would do better than the current rule.

#### *Retrieving a schema*

As with SG, the whole of CLARION's memory is probed simultaneously to determine all information relating to the current situation. The possible actions suggested by the separate procedural and declarative levels are then chosen through a weighted competition, reflecting the degree of emphasis CLARION is placing on each type of knowledge. Note that in both levels CLARION relies on a similarity-

based metric to generalise to novel situations. This is natural in the chosen domain, where all inputs are visual scenes; the rules basically contain a localist representation of information similar to that in the neural network.

### **Summary**

The form of schemata possible in these neural network models is determined partly by their learning mechanisms and partly by their retrieval mechanisms. The basic neural network is capable of learning complex mappings from the input to output data, and inherent mechanisms within the neural network are used to retrieve information most similar to the current situation. In CLARION, situations may be learnt explicitly with specific rules consisting of core and variable information.

### **Production-rule models**

Production rules have been a popular representation for a number of computational models, two notable examples being Soar (Laird, Newell & Rosenbloom, 1987) and ACT-R (Anderson & Lebiere, 1998). However, such models are also difficult to discuss in our framework, as their inherent power makes them suitable for application in a wide range of domains and settings, as well as for testing various theories of learning: there are few architectural constraints which have a significant bearing on the forms of knowledge learnt. Here, we describe the generic learning and retrieval mechanisms in Soar.

#### **Soar: chunking of productions**

The Soar system integrates perceptual-motor behaviour with basic capabilities for learning and problem solving. All knowledge within Soar is held in the form of productions, with a working memory holding specific attributes and their values. Soar operates in a cycle, attempting to satisfy some goal within its working memory. This cycle takes the contents of working memory and matches it to productions in its knowledge-base. These matching productions place new goals or other elements into working memory (this is known as the *elaboration* phase, which proceeds until all eligible productions have fired, *quiescence*), and then a decision is made as to which of the new goals to pursue next.

#### *Representing a schema*

All behaviour within Soar is goal oriented, in the sense that the system is always trying to satisfy some goal or another. Each goal contains three slots: the current problem space, state and operator. The specific representations for information in these slots can vary across applications. A particular schema may not be represented specifically in a production, but instead, in a specific context, a number of similar rules will be matched, suggesting interrelated subgoals, and so yield the effect of a schema.

#### *Learning a schema*

Learning within Soar is based on a chunking process that creates new rules. Each rule recreates the results of subgoals in relevantly similar future situations (Laird, Rosenbloom &

Newell, 1986). Chunking relies on an analysis of the dependencies within the solution to a given subgoal to create new rules. A new rule is created for each independent result, with a condition relating to the dependency analysis of the subgoal, and an action relating to the specified subgoal. This chunking mechanism is a universal learning mechanism, similar to explanation-based learning (see Rosenbloom & Laird, 1986). The interesting facet of learning within Soar is its ability to focus on those aspects of the situation used for problem solving, and to use only these relevant aspects in chunking. This focus ensures that the chunks learnt by Soar will generalise to novel situations. In addition, Soar has a process of variabilisation, in which information is made as general as possible before it is stored as a chunk in a production.

#### *Retrieving a schema*

The retrieval mechanisms within Soar operate only in its elaboration phase, in which “all directly available knowledge relevant to the current situation is brought to bear” (Laird, Rosenbloom & Newell, 1987). In this phase, every production in its memory whose condition directly matches something in the working memory is activated, and its suggested subgoals and other information are added to memory. Matching productions against working memory is based on the similarity of the attributes and their values.

#### **Summary**

Just as with neural networks, no specific structure corresponding to a schema exists in Soar. However, the basic learning mechanism within Soar, chunking, does limit the form and content of learnt productions. Firstly, productions are retrieved based on their similarity to items in working memory. The features placed within a production are taken from the set of dependent relations in the attainment of a goal. In addition, some variabilisation can occur on the features.

### **Symbolic network models**

This section considers a pair of models which construct symbolic networks of symbol-level information within a hierarchy. Each of these lays some claims to universality of application, but have currently only demonstrated good results in one or two areas. The first is the CHREST model, which learns about chess patterns, and the second is EUREKA, which learns about physics problems.

#### **CHREST: storing chunks into templates**

The CHREST (Chunk Hierarchy and REtrieval STRucture) model of expertise (Gobet & Simon, in press) is a recent development of EPAM (Elementary Perceiver and Memoriser) (Feigenbaum & Simon, 1984). The learning processes in EPAM include mechanisms for constructing a discrimination network and incorporating information into it; the learnt information is known as *chunks*. CHREST includes extra mechanisms for learning *templates* (Gobet & Simon, in press); it is this template which is of interest to us here, as it possesses schema-like properties.

A template is created in the following manner. During training, CHREST (just like EPAM) builds a discrimination network of chunks of information. Specific to CHREST is the ability to create lateral links (Gobet, 1996): in this case, *similarity links*. These similarity links can be used whilst searching the network to suggest chunks not directly linked by the tests in the network. However, the novel aspect of this is that a node can reorganise information in similar chunks (satisfying an overlap criterion) into a template. This template contains a core pattern, based on the original chunk, and a set of slots, for the information which varied across the associated chunks.

#### *Representing a schema*

CHREST represents all information as chunks within nodes in a discrimination network: a chunk is a familiarised pattern. Nodes are linked by test links, which require some features to be matched on traversal. Some of the nodes in the network contain *templates*, where a template contains a core chunk and a number of slots. However, CHUMP (Gobet & Jansen, 1994) and CHREST+ (Lane, Cheng & Gobet, 2000) additionally allow nodes in the network to be associated with information about possible moves or problem solutions, allowing CHREST to learn to solve problems.

#### *Learning a schema*

The discrimination network within CHREST is learnt through four learning mechanisms. Beginning from the root node, CHREST sorts a novel pattern through the network until no further test links can be applied. At the node reached, two things can occur. First, the pattern may match the chunk, in which case more information can be added to the chunk from the pattern (familiarisation). Second, the pattern may mismatch the chunk, in which case a further test link and node are created based on the mismatching features (discrimination). The third learning mechanism constructs *similarity links* between two nodes when their chunks have at least 3 identical items. Finally, for a node with at least 5 similarity links satisfying an overlap criterion, the chunk may be replaced by a template. This template uses the existing chunk as its core, and the varying information across the other nodes as its slots.

#### *Retrieving a schema*

Retrieving knowledge within CHREST is achieved simply by following the test links from the root node, applying the tests to the target pattern until no further test applies. The chunk at the node reached is the retrieved schema.

#### **EUREKA: restructuring knowledge**

EUREKA (Elio & Sharf, 1990) demonstrates how an effective organisation for large amounts of domain-specific knowledge can support efficient recognition and application of relevant knowledge to the problem at hand. Secondly, the model demonstrates how the qualitative shift from novice to expert levels of knowledge and organisation can arise within a learning framework. EUREKA uses a discrimination network, rather like the CHREST model described above, but instead of simple chunks, the nodes in EUREKA's network

hold Memory Organization Packets (MOPs) (Schank, 1980). Each MOP represents a complex knowledge structure holding generalised knowledge extracted from a group of individual experiences. Differences between experiences are encoded in the tests between the links in the discrimination network, and so similar previous experiences are retrieved based on the features in the network which match the current experience.

EUREKA has been applied to physics problems, and is initialised with a set of MOPs containing basic knowledge about physics concepts, equations and inference rules. However, this knowledge does not contain any information about their usefulness or relevance in any particular type of problem. When EUREKA is given its first physics problems, it must use its basic knowledge in conjunction with a means-ends problem-solving strategy to construct a solution. Having done this, EUREKA then places the entire problem and its solution (features, inferences and solution steps) into a P-MOP (Problem MOP). This P-MOP is then stored in the P-MOP network, where some reorganisation of the network may occur. When solving later problems, EUREKA can use information in a P-MOP in preference to its means-ends analysis, which can lead to a shift in EUREKA's problem-solving strategy towards a greater use of important abstract physics concepts, such as force or energy, usually not present in the problem statement. Also, the use of a P-MOP instead of means-ends analysis means the model begins to solve problems working forwards from the given information instead of backwards from the target, in accordance with observed differences between novice and expert problem solvers (cf. Koedinger & Anderson, 1990; Larkin, McDermott, Simon & Simon, 1980).

#### *Representing a schema*

EUREKA stores information in a network of P-MOPs. At the root of the network is a P-MOP representing a "generic physics problem". Each P-MOP contains several elements: firstly, a set of *norms* represent the features which a problem must satisfy for this P-MOP to apply; secondly, a set of *indices* (links) to other P-MOPs, with the index specifying the feature(s) which distinguish between them; thirdly, the P-MOP includes a general inference rule; fourthly, the P-MOP includes a specific solution method for carrying out the inference rule; and fifthly, the P-MOP includes a count for the number of problem-solving experiences which it organises (i.e. has matched in the past). The P-MOP representation is a clear example of an explicit schema, with the norms indicating the class of similar problems to which its inference rules will apply.

#### *Learning a schema*

EUREKA's learning mechanisms operate through a process of *reorganisation*. Once a problem has been solved, everything about the problem and its solution is collected into a problem-solving experience. This experience is then compared with the existing P-MOP retrieved from the network. If any of the norms differ between the P-MOP and the experience, these are removed from the P-MOP and used as indices to new organisation beneath this P-MOP; any inference rules referring to these differing norms are also removed

from the P-MOP and included in the new organisation. This process has the side-effect that partial solution methods may reside on P-MOPs. A further reorganisation can occur in cases where a descendant P-MOP covers most of the problem-solving experiences of its parent P-MOP; in such situations the organisation of the network is not efficient, and one of the discriminating features might be better seen as a commonality.

These two learning mechanisms can lead the network to focus on abstract features in the following way. A property such as a force may not be represented within the problem statement, however, it will be referred to in the problem solution. As problem-solving experiences are gathered, a number will be seen to include force within their solution, and so this feature will become a norm within the P-MOP. From there, the feature may be used to discriminate between different P-MOPs, because it has been derived as a feature of a number of problem-solving experiences.

#### *Retrieving a schema*

Each P-MOP in EUREKA's memory is a separate schema, and each is indexed through the P-MOP network. Any of the features in the initial problem representation can serve as indices into the P-MOP network. Whenever the feature appears as a difference in the P-MOP, the corresponding index is traversed. If a number of indices may be traversed, then EUREKA prefers the index leading to the P-MOP that organises the most problem-solving experiences. Hence, EUREKA is directed preferentially to patterns that recur most often. During the traversal, EUREKA will apply the inference rules of any P-MOPs that match the current situation; this process will alter the current situation (the set of equations and unknowns) and so affect the further traversal of the P-MOP network. Note that EUREKA's bias towards P-MOPs which organise larger numbers of problem-solving experiences means that P-MOPs arising from reorganisation of the network will be preferred during problem solving. It is this bias which ensures EUREKA will preferentially use P-MOPs emphasising the presence of forces or abstract entities: as discussed above, such P-MOPs are formed from the aggregate of several more concrete P-MOPs, and so organise a larger number of problem-solving experiences.

### **Summary**

The symbolic network models are closer to the spirit of traditional schemata theories. In particular, there is a close correspondence between the information in a P-MOP or the pairing of problem and solution nodes within CHREST, and the schemata discussed in Koedinger and Anderson (1990). Both models can use information to partially match a current situation. However, different learning mechanisms encode different kinds of information in their nodes; CHREST restricting itself to perceptual similarity, with EUREKA inferring more abstract quantities for use in discrimination.

### **Conclusion**

This paper has taken an inductive approach to the question of how to learn schemata by applying an analytical framework to a number of computational models, and describing

the ways in which these models represent, learn and retrieve schemata. Our aim has been to uncover, from existing models, the origins of constraints on the possible forms of schemata. From our analysis we can see some similarities across all the models. Firstly, all use a distributed form of representation, in the sense that schemata for novel situations will usually arise from a number of partial matches, although the symbolic network models possess more explicit schema-like structures. Secondly, all use a similarity-based form of retrieval, differing in the features which may be used for discrimination. In particular, EUREKA allows abstract features (not perceptually obvious) to become significant.

However, the differences in behaviour of the various models are largely down to their specific learning mechanisms. As stated in the introduction, the motivation for these models has not been to learn schemata, as such, but instead to learn effectively in general situations. We therefore conclude that, for the purposes of developing a more meaningful definition of schemata, we should begin by analysing the available range of learning mechanisms in models such as those referred to here. These learning mechanisms should be explored in their cognitive implications. For instance, the use of seriality or resource bounds, the malleability of learnt features and how wide-ranging any changes to previous knowledge may be. Most of these properties will come directly from the learning mechanisms, whereas others will be imposed by the interaction of the learning mechanisms with the other properties of the system, such as its use of perceptual-motor stimuli. Once these properties have been understood, the use of schemata for describing the units of knowledge within humans will become grounded in the processes by which that knowledge has been learnt.

## References

- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought* (Lawrence Erlbaum).
- Brewer, W. F. (1999). Schemata. In R. A. Wilson & F. C. Keil (Eds.) *MIT Encyclopedia of the Cognitive Sciences*, pp. 729-730.
- Elio, R. & Scharf, P. B. (1990). Modeling novice-to-expert shifts in problem-solving strategy and knowledge organization. *Cognitive Science*, 14, 579-639.
- Feigenbaum, E. A., & Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science*, 8, 305-336.
- Gobet, F. (1996). Discrimination nets, production systems and semantic networks: Elements of a unified framework. *Proceedings of the Second International Conference of the Learning Sciences* (pp. 398-403). Evanston, III: Northwestern University.
- Gobet, F. & Jansen, P. (1994). Towards a chess program based on a model of human memory. In H. J. van den Herik, I. S. Herschberg, & J. W. Uiterwijk (Eds.), *Advances in Computer Chess 7*. Maastricht: University of Limburg Press.
- Gobet, F. & Simon, H. A. (in press). Five seconds or sixty? Presentation time in expert memory. *Cognitive Science*.
- Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science*, 14, 511-550.
- Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, 33, 1-64.
- Laird, J. E., Rosenbloom, P. S., & Newell, A. (1986). Chunking in Soar: The anatomy of a general learning mechanism. *Machine Learning*, 1, 11-46.
- Lane, P. C. R., Cheng, P. C.-H., & Gobet, F. (2000). CHREST+: Investigating how humans learn to solve problems using diagrams. *AISB Quarterly*, 103, 24-30.
- Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Models of competence in solving physics problems. *Cognitive Science*, 4, 317-345.
- Rosenbloom, P. S., & Laird, J. E. (1986). Mapping explanation-based generalization onto Soar. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 561-567). Philadelphia, PA: MIT Press.
- Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R.J. Spiro, B.C. Bruce and W.F. Brewer (Eds.) *Theoretical Issues in Reading Comprehension* (Lawrence, Erlbaum), pp. 33-58.
- Rumelhart, D. E., Smolensky, P., McClelland, J. L. & Hinton, G. E. (1984). Schemata and sequential thought processes in PDP models. In D. E. Rumelhart & J. L. McClelland (Eds.) *Parallel Distributed Processing, Vol. II*. MIT Press, Cambridge, MA.
- St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46, 217-257.
- Shank, R. C. (1980). Language and memory. *Cognitive Science*, 4, 243-284.
- Sun, R., Merrill, E. & Peterson, T. (in press). From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive Science*.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning, *Cognitive Science*, 12, 257-285.

# Retrospective Effects in Human Causality Judgment

M.E. Le Pelley (mel22@hermes.cam.ac.uk)

D.L. Cutler (dlc29@hermes.cam.ac.uk)

I.P.L. McLaren (iplm2@cus.cam.ac.uk)

Department of Experimental Psychology; Downing Site  
Cambridge CB2 3EB, England

## Abstract

The phenomenon of retrospective revaluation has been a challenge to many associative learning theories as it involves a change in the associative strength of a cue on trials on which that cue is absent. The present experiment combines several retrospective learning contingencies in a single, within-subjects experiment, allowing for valid comparisons between contingencies. One of the most popular models of retrospective revaluation, Dickinson & Burke's (1996) modification of Wagner's (1981) SOP theory, fails to explain the full pattern of results. A connectionist model that explains retrospective revaluation in terms of changes in retrievability in memory, rather than as new learning about absent cues, is shown to provide a better account of the results.

## Introduction

Perhaps the biggest challenge to traditional theories of associative learning in recent years has come from studies of retrospective effects in cue competition. Such effects have overturned the central tenet of many of the most influential learning theories (e.g. Rescorla & Wagner, 1972; Wagner, 1981) – that only cues present on a given trial may engage the learning process.

Consider a typical retrospective revaluation study, as shown in Table 1. Stage 1 involves training of the cue compounds AB and CD to predict some outcome. In Stage 2 one of the cues (the competing cue) from each compound is selected for either further training (in what is known as the backward blocking condition) or extinction (unovershadowing). The typical result of such studies is that, following stage 2 training, when the cues that have not received any further training in stage 2 (the target cues) are tested, D is now rated as a better predictor of the outcome than B. Thus the perceived predictive validity of a cue can be altered after initial compound training with that cue, either by training the other cue of the compound pair as a valid predictor of the outcome (as in backward blocking) or by extinguishing it (as in unovershadowing). The inference from this is that the associative strength of the target cue representation (B or D above) to outcome representation association can change on trials in which that cue is not presented (A+ and C-).

Retrospective revaluation has now been reliably demonstrated in a number of experiments with humans, using causal judgments of a cue→outcome relationship as indicators of the strength of the association between their representations (e.g. Chapman, 1991; Dickinson & Burke,

1996; Shanks, 1985). We will consider in some detail here one of the more popular theories of associative learning; Wagner's (1981) SOP model, and Dickinson & Burke's (1998) modification allowing it to explain retrospective effects.

SOP proposes that stimuli are represented by nodes in an associative memory that are composed of a number of elements. These elements can be in one of three states at any instant; one inactive state (I) and two active states (A1 and A2). Presentation of a stimulus excites the elements representing that stimulus into A1. These elements then decay back to I via A2. Exciting a node via an associative connection, however, causes a transition from I directly to A2. Changes in the associative connection between two nodes depend on temporal overlap of the states of their elements. Whenever the elements of two nodes are in A1, there is an increment in the excitatory strength between them. When the elements for one node are in A1 and those of another are in A2, there is an increment in the strength of an inhibitory connection. Critically, SOP states that only cue elements in A1 will engage the learning process (i.e. learning will only accrue to cues that are physically present on a trial). Hence, as there can be no learning about absent cues, SOP is unable to explain the results of retrospective revaluation studies.

Dickinson & Burke (1996) proposed a modification to SOP to allow it to explain retrospective revaluation (Table 2). They suggested that CS elements in A2 could engage learning, with an increment in excitatory strength whenever there was an overlap in activation states (be this in A1 or in A2) and an increment in inhibitory strength whenever elements were in different states. Thus they specified the sign with which learning occurs to be a symmetrical function of elemental activation states.

Consider now the contingencies shown in Table 1. During the first stage both target and competing cues, and the US, are presented, and so all will have elements in A1. Hence target and competing cue elements will form excitatory connections to US elements, and within-compound associations will form between target and competing cues. In the unovershadowing contingency cue C is now presented, and will retrieve D elements into A2 via the within-compound link. The US will also have elements in

Table 1: A typical retrospective revaluation design.

Condition	Stage 1	Stage 2	Test
Backward Blocking	AB+	A+	B?
Unovershadowing	CD+	C-	D?

Table 2: Modified SOP.

		US Element	
		A1	A2
CS Element	A1	<b>E</b>	<b>I</b>
	A2	<b>I</b>	<b>E</b>

E : Excitatory connection strengthened

I : Inhibitory connection strengthened

A2 (retrieved via the C→US connection). As both D and the US have elements in the A2 state, modified SOP predicts an increment in the excitatory strength between them. Hence D's rating is predicted to increase as a result of the C- trials, even though it is absent.

The case for the backward blocking contingency is less clear. Presentation of A in stage 2 will retrieve B elements into A2. The outcome is presented, so some of its elements will be in A1, but it is also predicted by virtue of the A→US connection, so it will also have elements in A2 (these elements cannot go straight from A2 to A1 when the US is presented; they must pass through the I state first). Thus any inhibitory A2-A1 learning between B and the US will be offset to some extent by excitatory A2-A2 learning. Whether the model predicts a net increase or decrease in the rating of B depends on which of the processes engaged by congruent and incongruent elemental states is stronger.

The overall result, though, is that after training D will be rated higher than B. According to modified SOP, then, the driving force behind retrospective revaluation is unovershadowing; backward blocking has a smaller role. This is supported by Larkin, Aitken & Dickinson (1998), who tried to measure the effects of unovershadowing and backward blocking separately by comparing each to a control contingency, EF+ X+ (for which neither target cue nor competing cue is trained in Stage 2). As predicted by modified SOP, they found evidence for a significant effect of unovershadowing, but the evidence for backward blocking was weaker and fell short of significance.

Backward blocking and unovershadowing are not the only retrospective effects that have been found in human causal learning. It has long been known that following A+, AB- training, B will typically become established as an inhibitor of the US, able to counteract the excitatory potential of A. Chapman (1991) reversed this procedure, to give an AB-, A+ design. This procedure was sufficient to establish B as an inhibitor of the US (i.e. it received a lower rating on test than C or D from a CD-, X- control contingency). The inhibitory properties of B must have been assumed in retrospect, as A was only established as a good predictor of the US following AB- trials.

Note that the phenomenon of backward-conditioned inhibition is in line with the predictions of modified SOP. During the first stage a within-compound association is learnt between A and B. A then retrieves B elements into A2 in stage 2. The outcome is presented, and so has elements in A1. The resulting A2-A1 activity will result in formation of an inhibitory link between B and the US.

Thus modified SOP is well equipped to deal with some of the major findings of retrospective learning studies with humans. In the present experiment we use these retrospective effects as a benchmark from which to provide a more critical assessment of the mechanism for retrospective revaluation proposed by modified SOP.

The design of the experiment is shown in Table 3. We used an allergy prediction paradigm, as employed by Dickinson & Burke (1996) and Larkin et al. (1998). Participants play the role of a food allergist trying to judge the likelihood that various foods will cause an allergic reaction in a hypothetical patient (Mr. X). The foods, then, constitute the cues, and the allergic reaction is the outcome. Following training, subjects rated how strongly each of the foods predicted the occurrence of an allergic reaction.

Table 3: Design of the experiment.

Condition	Pre-exp	Cond 1	Cond 2
B. Block		AB+	A+
Unover		CD+	C-
L,A&D Control		EF+	G+
PR Control		HI+	H+/H-
BCI		JK-	J+
BCI Control 1		LM-	L-
BCI Control 2		NO-	P-
BB Pre-exp	QR		Q+
BB Pre-exp Control 1	ST		S-
BB Pre-exp Control 2	UV		
Fillers		WX-	

These ratings were taken as a measure of the associative strength of a connection from cue to outcome.

We also follow Dickinson & Burke (1996) and Larkin et al. (1998) in using a large number of cues. This creates a large memory load, hopefully preventing subjects from basing their ratings on inferences made from explicit episodic memories of the various trial types. Instead subjects should have to rely on associative processes to provide an "automatic" measure of the causal efficacy of each cue.

The first two rows of Table 3 (B. Block and Unover) are the contingencies of a standard retrospective revaluation experiment, as shown in Table 1. Retrospective revaluation is demonstrated if D is rated higher than B on test.

The "L,A&D" contingency is a control of the kind used by Larkin et al. (1998). Following compound training in Cond 1, neither cue is presented in Cond 2, and so no revaluation will occur. Thus backward blocking and unovershadowing can be assessed independently relative to this control. Backward blocking would be evidenced by a lower rating of B than E or F; unovershadowing by a higher rating of D than E or F.

The "PR Control" is a second control that might allow us to dissect out the effects of backward blocking and unovershadowing. Following compound training in Cond 1, the competing cue receives partial reinforcement. Thus there are an equal number of H+ and H- trials. Suppose that unovershadowing is much stronger than backward blocking. On each H- trial in stage 2 there would be an unovershadowing effect, with I's association to the US becoming stronger. On each H+ trial there would be little effect, as backward blocking is weak. The contingency becomes, in effect, HI+ H-, i.e. unovershadowing, and so we expect I's rating to be similar to D (from an actual CD+ C- contingency). The opposite would be true if backward blocking were stronger than unovershadowing. In general, the PR Control target cue will receive a rating closer to the target cue of the retrospective revaluation contingency having the stronger effect.

The next three rows show a backward-conditioned inhibition contingency and two controls respectively. As described earlier, backward-conditioned inhibition will be demonstrated if K is rated lower on test than M (from Control 1) and N or O (from Control 2).

"BB Pre-exp" is short for backward blocking pre-exposure. This involves compound pre-exposure during the first stage (cf. compound training for backward blocking), followed by excitatory training of the competing cue in Cond 2. Modified SOP predicts that unovershadowing will have a larger effect than backward blocking in retrospective revaluation, because in a backward blocking con-

tingency the US is strongly predicted in the second stage, such that it has elements in A2. The excitatory A2-A2 learning then offsets the effect of inhibitory A2-A1 learning. As a result of using compound pre-exposure in the "BB Pre-exp" contingency, though, the US will not be expected on Cond 2 trials. Hence when it is presented all of its elements should be free to enter A1. R will be retrieved into A2 by Q via the Q-R association developed during pre-exposure. The resulting A2-A1 overlap should produce strong inhibitory conditioning. Modified SOP thus makes the clear prediction that R will be rated lower than T (from Control 1) and U or V (from Control 2).

The Filler trial was used so that there were an equal number of positive and negative trials during Cond 1.

## Method

**Participants** Twenty-four Cambridge University students (14 female, 10 male; age 19-23) took part in the experiment.

**Apparatus** The experiment was run on a Power PC Macintosh with a 14" monitor.

The foods used were: Oranges, Tomato, Cheese, Lobster, Rice, Peaches, Banana, Grapes, Yoghurt, Melon, Broccoli, Aubergine, Eggs, Potatoes, Carrots, Lentils, Sardines, Gammon, Dates, Mushrooms, Raspberries, Jam, Onion, Steak. These foods were randomly assigned to the letters A to X in the experimental design for each subject.

**Procedure** At the start of the experiment each subject was given a sheet of instructions presenting the "allergy prediction" cover story for the experiment. They were told that in the first block they would be looking over records of foods eaten at the clinic by Mr. X, but would not be told whether or not allergic reactions occurred, while in the second and third blocks they would be asked to make predictions based on the foods eaten. They were also told that at the end of the experi-

ment they would be asked to rate each of the foods according to how strongly it predicted allergic reactions.

On each pre-exposure trial, the words "Meal [meal number] contains the following foods:" followed by the two foods appeared on the screen. Subjects were then cued to enter the initial two letters of each of the foods. This was to ensure that they paid attention to the pairings of foods when no allergy prediction was required. There were three trial types in this stage: the order of trials was randomized over each set of three with the constraint that there were no immediate repetitions across sets. Participants saw each pair of foods eight times in this stage. The order of presentation on the screen (first/second) within each compound pair was randomized.

The same message appeared on the screen on Cond 1 and Cond 2 trials. However, now the subjects were asked to predict whether or not eating the foods would cause Mr. X to have an allergic reaction, using the "x" and "." keys (counterbalanced). The screen then cleared, and immediate feedback was provided. On positive trials the message "ALLERGIC REACTION!" appeared on the screen; on negative trials the message "No Reaction" appeared. If an incorrect prediction was made, the computer beeped. There was an explicit break between Cond 1 and Cond 2, when subjects were told that they would now see a new set of meals, some of which contained foods they had seen earlier and some of which didn't. There were eight trial types in Cond 1, and nine in Cond 2. The order of trials was randomized over each set of eight or nine. Participants saw each meal eight times in Cond 1 and Cond 2. Four of the eight H trials in Cond 2 were positive, and the other four were negative, in random order.

In the final rating stage subjects were asked to rate their opinions of the effect of eating each of the foods on a scale from -10 to +10. They were to use +10 if the food was very likely to cause an allergic reaction in Mr. X, -10 if eating the food was very likely to prevent the occurrence of allergic reactions which other foods were capable of causing, and 0 if eating the food had no effect on Mr. X (i.e. it neither caused nor prevented allergic reactions).

All of the foods seen in training were then presented in random order for rating. For clarification, participants also had access to a card on which the instructions on how to use the rating scale were printed. Once a food had been rated it disappeared from the screen and the next appeared, so that participants could not revise their opinions upon seeing later foods.

## Results

Figure 1 illustrates the percentage of trials on which subjects thought an allergic reaction would be caused by the food(s) shown in each of the 8 trial sets of Cond 1 and 2. Subjects' responses were clearly appropriate to the relevant underlying contingencies by the end of each stage, with all of the positive trial types eliciting more "Allergic Reaction" responses, negative trial types receiving more "No Reaction" responses, and the H+/H- trials receiving about 50% positive and negative responses.

Of more interest are the ratings of the causal efficacy of each of the foods. The mean rating given to each of the 24 foods is shown in Figure 2. A one-way, repeated measures ANOVA was carried out on these ratings as a preliminary to assessing the effects of interest by means of planned comparisons. There was a significant main effect of food [ $F(23,529) = 22.46, p < 0.001$ ]. Retrospective reevaluation was seen, in that the target cue of the backward blocking contingency (B) was rated significantly lower than that of the unovershadowing contingency (D) [ $F(1,23) = 7.24, p < 0.01$ ]. Hence it appears that our experimental paradigm is sensitive to changes in the perceived causal efficacy of cues on trials on which those cues are not present.

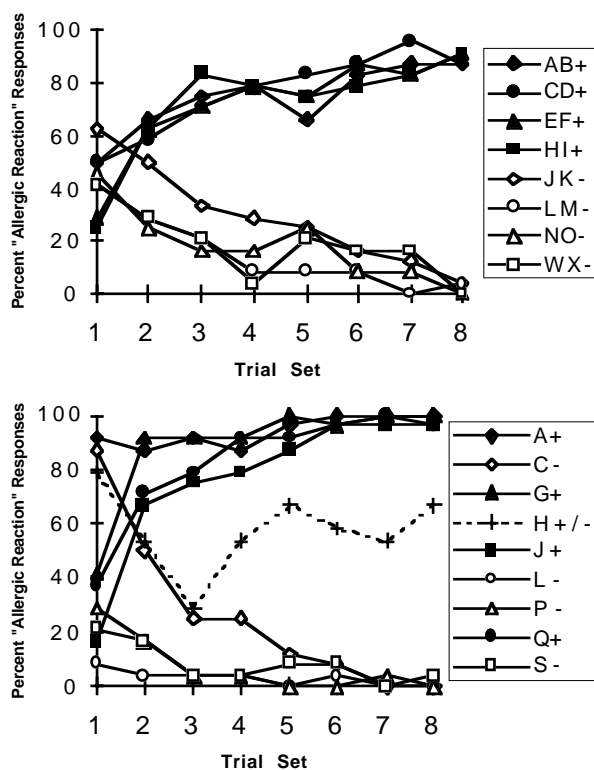


Figure 1. Acquisition of discriminations in (A) Stage 1 and (B) Stage 2.

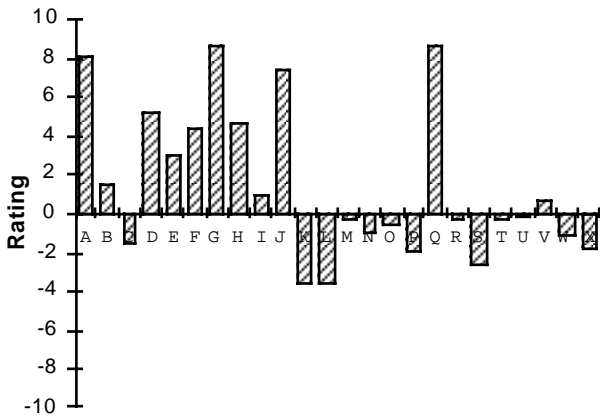


Figure 2. Mean ratings given to the 24 foods.

The result for the L,A&D Control contingency is given by the average of cues E and F, which are equivalent. This does not differ significantly from B or D [ $F(1,23)=2.78$  and  $1.23$  respectively,  $ps > 0.05$ ]. Given this failure to reach significance, our results neither confirm nor contradict those of Larkin et al. (1998).

We now turn to the results of the PR Control contingency. The rating of the target cue from this contingency (I) is very similar to that of the backward blocking contingency, but quite different from that of the unovershadowing contingency. This is supported statistically: B and I do not differ significantly [ $F < 1$ ], whereas the difference between D and I is highly significant [ $F(1,23) = 7.44$ ,  $p < 0.01$ ]. It was stated earlier that the rating of the target cue of the PR Control contingency should be more similar to the target cue of whichever retrospective revaluation process (backward blocking or unovershadowing) is stronger. Hence, given that the rating of I is more similar to B than D, the PR Control contingency indicates that backward blocking is a stronger process than unovershadowing.

We also have evidence for backward-conditioned inhibition in this experiment. Cue K is rated lower than its equivalents in the two control contingencies (M, and the average of N and O, none of which differs significantly from each of the others). These differences are significant [ $F(1,23) = 7.45$ ,  $p < 0.01$  and  $F(1,23) = 6.98$ ,  $p < 0.01$  respectively]. There is no evidence for any retrospective learning in the BB Pre-exp contingency, however. The target cue of this contingency is R. The two controls here are T and the average of U and V (none of which differ from one another). In the former case, the means are identical; in the latter the difference is not significant [ $F < 1$ ].

## Discussion

Looking at the results above, we can see that modified SOP is successful in explaining some aspects of this experiment (the occurrence of retrospective revaluation and backward-conditioned inhibition). However, it also has important failures.

The fact that the PR Control indicates a stronger role of backward blocking than unovershadowing in this experiment is a great problem for modified SOP. It implies that the simplistic approach taken to the associative processes occurring in these contingencies is insufficient to provide a full account of human behaviour with respect to learning about absent cues, as the model predicts that unovershadowing will be more influential than backward blocking.

Note that this result also argues against subjects' using a rational, Bayesian approach to the contingencies seen. According to this idea, subjects would integrate the information experienced in the two stages to derive the most likely cause of the US. For example, A- trials following AB+ trials (unovershadowing) indicate that it *must* have been B that caused the US on the AB+ trials. Hence B's rating will increase as a result of A- trials. Less information is given by A+ trials following AB+, though: B *could* still be a cause of the US on AB+ trials. Hence this rational approach predicts that unovershadowing will be stronger than backward blocking, whereas the results of the PR Control indicate the opposite.

In addition, modified SOP predicts a large difference between the target cue of the BB Pre-exp group and its controls, but no difference is seen. The BB Pre-exp contingency should be the situation in which the inhibitory A2-A1 process, proposed to underlie backward blocking, should be most prominent, and yet no effect was seen. This is particularly noticeable as the BCI contingency, which on the surface appears very similar, did show an effect. The failure to find an effect in one of the two contingencies is hard to reconcile with modified SOP.

In summary, then, it seems that modified SOP is able to explain the existence of retrospective revaluation and associated phenomena, but that the mechanics of the explanation offered do not agree with our empirical findings. We now offer an alternative class of model which seems better able to cover the known facts with respect to human studies of retrospective revaluation.

## APECS: A model of associative learning

It is possible to explain the results of the above experiments using a version of McLaren's (1993) APECS model. The mechanics for learning in APECS are similar to standard backpropagation (Rumelhart, Hinton, & Williams, 1986), but differ in that once the weights appropriate to a mapping have developed, the learning represented in those weights is protected. This is achieved by reducing the learning rate parameters for the hidden unit carrying the mapping. The effect is to "freeze" the weights to and from a certain configural unit at the value they hold immediately following experience of that configuration. Crucially, this freezing of weights to and from a hidden unit occurs only if that hidden unit has a negative error value, i.e. *if it is part of a mapping that predicts an incorrect outcome for the current input*. This reduces the interference arising as a result of subsequent experience of similar (but not identical) input patterns. Indeed APECS was originally designed as a solution to the problem of catastrophic interference in learning (McCloskey & Cohen, 1989).

Specifically, APECS has different learning rate parameters for input-hidden and bias-hidden connections. The former are frozen to prevent interference; the latter remain high. Hence extinction (suppression of inappropriate responses) is achieved by an increase in the negative bias on the hidden unit carrying the inappropriate mapping, rather than by reduction of weights (which would cause the original mapping to be lost from the network). Given appropriate input cues, the negative bias on the hidden unit can be overcome and the original mapping retrieved.

In addition, in our instantiation of APECS each different pattern of stimulation is represented by its own hidden unit, similar to Pearce (1987).



Consider what happens in the network on AB+ training. It will learn a mapping from A and B input units to the US output unit, mediated by a hidden unit that can be thought of as representing the configuration of A and B ( $AB_{\text{hidden}}$ ). On each AB+ trial the excitatory connections to and from  $AB_{\text{hidden}}$  will grow stronger. Now consider the gap between AB+ trials, when no inputs are presented. According to the logistic activation function employed with APECS, when no inputs are presented the hidden units will have an activation of 0.5 (see Rumelhart et al., 1986). This activation will feed along the  $AB_{\text{hidden}} \rightarrow \text{US}$  connection learnt on the preceding trial, and activate the US unit. This is obviously inappropriate when no inputs are presented. The US unit will take on a negative error, which is propagated back to  $AB_{\text{hidden}}$ . As explained earlier, a negative error means that the weights to and from the hidden unit are frozen. In order to suppress the expression of the US on gaps between the AB+ trials, the  $AB_{\text{hidden}}$  unit will therefore develop a negative bias.

In a backward blocking contingency we now train on A+ trials. As explained above, in this instantiation of APECS, each different input  $\rightarrow$  output mapping is assigned a new hidden unit. Hence a new unit is recruited to carry the A+ mapping. As the previous excitatory connection from A to the US (via the AB configural unit) is still useful in reducing the output error (i.e. there is a positive output error on A+ trials, which is propagated back to the  $AB_{\text{hidden}}$  unit), the learning rate for the  $A \rightarrow AB_{\text{hidden}} \rightarrow \text{US}$  connections will also remain high. Given the negative bias on the  $AB_{\text{hidden}}$  unit, and the fact that training was with A and B in stage 1, A alone will not succeed in fully activating the US at the start of A+ training. Hence the connections from A to both hidden layer units, and from the hidden layer units to the US node, will strengthen.

Now, when no stimuli are applied (during inter-trial interval) the hidden units will deliver some positive activation to the output unit. Thus the  $A_{\text{hidden}}$  unit will assume a negative bias. More importantly, the negative bias on the  $AB_{\text{hidden}}$  unit will also become increasingly negative to counter the extra positive activation feeding to the US.

What now happens when B is presented on test? B will provide  $AB_{\text{hidden}}$  with positive activation, but this may not effectively counter the unit's large negative bias, and hence the US will receive little activation. This, of course, assumes that  $AB_{\text{hidden}}$ 's gain in negative bias outweighs the strengthened  $AB_{\text{hidden}} \rightarrow \text{US}$  connection. This is guaranteed, as the two bias changes must together counter the increased  $A \rightarrow \text{hidden}$  and  $\text{hidden} \rightarrow \text{US}$  weights of *both* routes to the US, whereas only the increased  $AB_{\text{hidden}} \rightarrow \text{US}$  connection will facilitate B's ability to retrieve the US.

Hence APECS is able to explain backward blocking: as a result of A+ trials following AB+ training, B becomes less able to retrieve the US. Note that, unlike modified SOP, this is not as a result of new learning about B (the  $B \rightarrow \text{hidden}$  connection is unchanged on A+ trials), but rather as a result of changes in the retrievability of a previously-learned association.

Consider now the A- trials of the second stage of an unovershadowing contingency. Once again a new hidden unit is recruited to carry this mapping. In this case, however, the  $AB_{\text{hidden}}$  unit is not reused: it carries an inappropriate excitatory mapping and so will have a negative error. Hence its weights are frozen, and it takes on an increased negative bias. In addition, an inhibitory mapping from A to US will develop via the  $A_{\text{hidden}}$  unit in

order to counter the positive activation flowing to the US via the original mapping.

Between A- trials, when no inputs are presented, the US will receive excess negative input as a result of this new, un-suppressed inhibitory mapping. The network counters this problem in two ways. One is for the  $A_{\text{hidden}}$  unit to develop a negative bias. The other is for the negative bias on the  $AB_{\text{hidden}}$  unit to reduce, allowing through more positive activation.

The upshot of the decrease in negative bias on the  $AB_{\text{hidden}}$  unit is that presentation on B will now excite the US more effectively than following initial AB+ training: this is the standard unovershadowing effect.

Thus the features of APECS that prevent it from suffering from catastrophic interference also allow it to explain retrospective revaluation. Indeed, a backward blocking contingency can be seen as an interference design, where two different pathways (via  $A_{\text{hidden}}$  and  $AB_{\text{hidden}}$ ) compete to activate the same outcome. Hence A+ trials interfere with memory of the AB+ mapping, causing this pathway to be suppressed. In unovershadowing the situation is reversed: the two pathways have opposite outcomes (AB+ and A-), and so on A- trials the AB+ pathway need not be suppressed, and can even become stronger to counter the influence of the new negative pathway.

On performing initial simulations of retrospective revaluation using APECS it was found that unovershadowing consistently showed a larger effect than backward blocking. This sits well with the results of Larkin et al.'s (1998) study. But how then could this model explain the PR Control, which indicates that backward blocking has the greater effect?

The answer lies in the nature of the AB+ A+/A- design used. It was mentioned earlier that each new input  $\rightarrow$  output mapping recruits a new hidden unit. Hence the occurrence of A+ and A- trials in stage 2 will lead to the recruitment of two new hidden units, one carrying an excitatory mapping, the other an inhibitory mapping. Thus there are now two excitatory pathways to the US (via the  $AB_{\text{hidden}}$  and  $A_{\text{hidden}}$  units) as opposed to only one inhibitory pathway (via the  $A_{\text{hidden}}$ ). This means that any influence of the inhibitory pathway on each excitatory pathway (i.e. unovershadowing) will be relatively slight, as the effect is shared between, and countered by, both excitatory pathways. The effect of one excitatory pathway on the other (backward blocking), though, is relatively unaffected, as it is still a one versus one situation. Hence backward blocking is relatively preserved in this contingency, whereas unovershadowing is greatly reduced.

Figure 3 shows simulation results for the retrospective revaluation contingencies of our experiment, along with the empirical results. The simulation results are the average of 24 simulations run with APECS, each representing one subject, with exactly the same trial order as experienced by the real subjects. Each trial involved 1000 learning cycles. A hidden unit is defined as being "active" when it receives positive activation from the input layer. Thus if cue A is presented to the network, any hidden unit representing a configuration that includes cue A will be active. Activity extends into the period immediately following each trial, when no inputs are presented (again for 1000 learning cycles). The learning rate parameters for input-hidden and hidden-output units are both 0.85 when a hidden unit is active and has a positive error, and 0.001 when it is not. The parameter for bias-hidden changes is 0.25

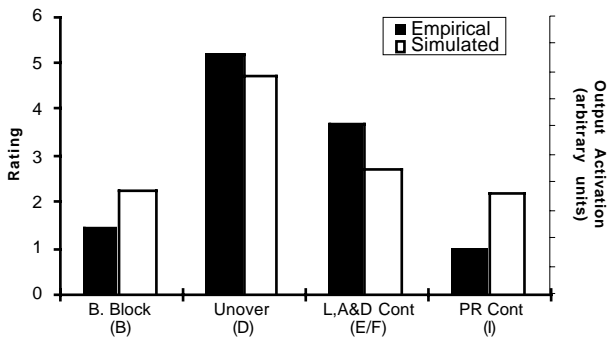


Figure 3. Empirical and simulated data for the retrospective revaluation contingencies.

when a hidden unit is active, 0.001 when it is not. Thus we make the reasonable assumption that changes due to learning take place faster than changes in memory, i.e. learning represents rapid acquisition, and memory represents a more gradual decline in retrievability.

What predictions does APECS make for other contingencies used in our experiment? Easiest to understand is the BB Pre-exp contingency. On the pre-exposure trials, there is no error on the output unit (whether or not the outcome occurs is not known). Given that it is output error that drives learning in error-correcting networks, this lack of error means that there is no drive to form associations. Hence the cues involved in these trials remain unconnected to the US following pre-exposure. As these cues have no connections to the output, their associative status cannot change on any subsequent trials on which they are not presented, and so no differences will be seen amongst these groups (as observed empirically).

The situation is slightly more complex for the BCI contingency. Typically a context in which outcomes occur will become a weak excitator of the outcome itself. Thus cues presented on negative trials in this context will become weak inhibitors in order to overcome this excitation (demonstrated by the negative ratings given to cues W and X). Hence J and K will develop an inhibitory link to the US via a  $JK_{\text{hidden}}$  unit. This unit will take on a slight negative bias to prevent its expression when no inputs are presented. The network is now presented with J+ trials. A new hidden unit will be recruited to carry this excitatory mapping. There will be a slight increase in the negative bias on the  $JK_{\text{hidden}}$  unit, but given that the inhibitory influence of this pathway is slight, the drive to suppress it will also be slight. On the gap between J+ trials, the US will receive excess positive input. One way to decrease this is for the  $J_{\text{hidden}}$  unit to develop a negative bias to suppress the excitatory mapping just learnt. A second way to reduce the activation of the US is to decrease the suppression of the  $JK_{\text{hidden}}$  unit, allowing more activity to flow through the inhibitory pathway. This release in suppression between trials (driven by the strong excitatory pathway) more than compensates for the increase in suppression on J+ trials (driven by the weak inhibitory pathway), and so overall the suppression of the  $JK_{\text{hidden}}$  unit (which carries the inhibitory mapping) decreases over J+ trials. Hence the ability of K to retrieve the US decreases as a result of J+ training following JK- trials: backward-conditioned inhibition is seen.

This is again confirmed by simulation. Figure 4 shows the results for the relevant contingencies from the simulation of this experiment described above.

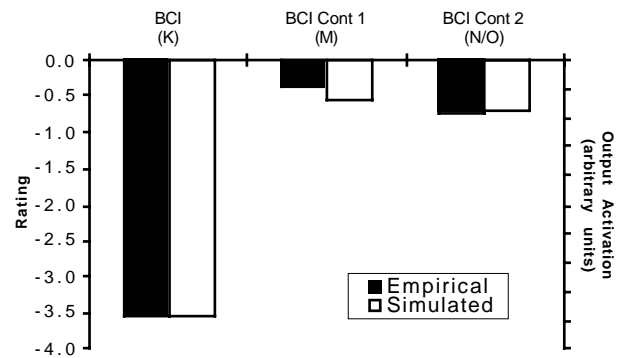


Figure 4. Empirical and simulated data for the backward-conditioned inhibition contingencies.

In conclusion then, it seems that modified SOP is able to explain certain retrospective effects in human causality learning on a coarse scale, but that the explanation offered for these effects (novel learning about absent cues following retrieval via within-compound associations) does not stand up to closer scrutiny. A memory-based explanation, with retrospective effects manifest as changes in retrievability rather than new learning about absent cues, shows better agreement with empirical data, and may prove a more fruitful approach for future investigation.

## References

- Chapman, G. B. (1991). Trial-order affects cue interaction in contingency judgement. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 17, 837-854.
- Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology*, 49B, 60-80.
- Larkin, M. J. W., Aitken, M. R. F., & Dickinson, A. (1998). Retrospective revaluation of causal judgements under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24, 1331-1352.
- McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, 24, 109-166.
- McLaren, I. P. L. (1993). APECS: A solution to the sequential learning problem. *Proceedings of the XVth Annual Convention of the Cognitive Science Society* (pp. 717-722). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, 94, (61-73).
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McClelland & the PDP Research Group (Eds.), *Parallel Distributed Processing* (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press.
- Shanks, D. R. (1985). Forward and backward blocking in human contingency judgements. *Quarterly Journal of Experimental Psychology*, 37B, 1-21.
- Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behaviour. In N.E. Spear & R.R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 5-47). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

# A Constructivist Model of Robot Perception and Performance

Joseph A. Lewis and George F. Luger

jalewis@cs.unm.edu, luger@cs.unm.edu

Department of Computer Science; FEC 323 University of New Mexico

Albuquerque, NM 87131

## Abstract

We present a new architecture for robot control rooted in notions from Brooks' subsumption architecture and extended to include an internal representation which matures as it experiences the world. Our architecture is based on the Copycat program of Mitchell and Hofstadter, a model of fluid representation whose details we discuss. We show how our architecture develops a representation of its environment through a continuing interaction with it. The architecture is founded on a dynamical systems interpretation of representation and demonstrates the importance of the use of "embodiment". It reflects a constructivist epistemology, with the robot designed to utilize its environment in its exploration.

## Introduction

We present an architecture for robot control based on the constructivist insight that representation occurs as a product of the active interpretation of perception-based experience. This architecture supports the control program for a robot whose task is to move about, explore, and map its world. The robot generates a representation of its environment by converting sequences of sensory data into perceived "objects". We believe that our approach will allow the robot to behave more robustly than does the use of the more traditional "preinterpreted" (McGonigle 1998) representations of its world.

In this paper, we describe the details of the model and then show its capacity to construct interactively a representation of surfaces and gaps (discovering the "objects") in its environment. The preliminary results demonstrate the use of this emergent architecture to solve simple robotics problems and to generate emergent structures that represent persistent features of the changing data from the environment. We also discuss work currently underway to allow the robot's behavior to be improved by the emergent representations.

Our work builds on research from several disciplines. These include: behavior-based robotics (Brooks and Stein 1994), the "dynamical nature" of representation and intelligence (Steels 1995, 1996), and the philosophical insights of Maturana and Varela (1980) and Clark (1997), on the self-organizing nature of living systems and their "coupling" with their environments. Further support for our approach comes from Holland's (1986, 1998) ideas on emergence in the context of classifier systems, and work on "fluid representations" in software architectures, for example Copycat, proposed by Mitchell and Hofstadter (Mitchell 1993). We continue the focus on "situating"

cognitive behavior in its environment originating with (Winograd and Flores 1986).

Traditional cognitive science and artificial intelligence have focused on building the (supposedly static) structures involved in representational processes. The peculiar fluid quality of actual structures that support complex problem solving in changing environments has resisted elucidation. More recently a shift of focus, generated in part from the study of complex adaptive systems, has driven research to attempt to characterize the dynamical processes underlying these representational structures. Architectures whose representations are implicit in behavior, supported by dynamical constraints and triggers from the environment, have begun to validate the constructivist claim that "refinement of an interpretive framework is usually driven by the tension between the pattern of interpretation and the demands of successful interaction." (Luger 1994). These models also provide suitable tests for the assertion that representations only have meaning in the context of embedding experiences.

Our control architecture implicitly defines intelligence with the four characteristics of evolving complex adaptive systems proposed by Steels (1996). The first of these criteria is *self-maintenance* (we prefer the term *autopoiesis* from Maturana and Varela (1980) who also describe a "mutual maintenance" relationship among system components). The remaining criteria for describing intelligence are *adaptivity*, *information preservation*, and, in response to the demands of a complex environment, a *spontaneous increase in complexity*.

We also follow Steels (1996) suggestions that there are two ways that intelligent systems can achieve these four criteria. The first is through the use of a general purpose dynamical architecture. The second is through the capture of the emergent properties of interactive behavior, enabling the formation of concepts about and representations of the environment. We feel that the emergence of structures evolved through "coupling with" an environment is a defining feature of intelligence, and call this *behaviorally coupled representation*. Furthermore, this "embodiment" is so critical to the study of intelligence that at least at the present state of our understanding, building and testing robots is an insightful necessity.

## A New Architecture for Robot Control

Most early approaches to robotics subscribe to an implicit *sense-model-plan-act* framework (Brooks 1991b). In the

1980s, concern arose about the performance and complexity entailed by this framework when applied to adaptive autonomous agents functioning in actual environments. This concern motivated a shift in thinking about the design of robotic systems as well as conjectures about the organization and use of intelligence itself.

The subsumption architecture (Brooks 1991a) marked the beginning of *behavior-based robotics*. Behavior-based robotics emphasizes the integration of semi-independent layers that produce behaviors directly from input rather than each contributing to a stage of the sense-model-plan-act framework. The focus is on interaction with the environment as a trigger for behavior rather than use of explicit representation. The ability to react to dynamic features of an unpredictable environment and to generate robust behavior despite sensor uncertainty is a signature of this behavior-based approach. Testing physically constructed robots interacting with complex worlds bears much weight in this new paradigm of robotics research. The behavior-based approach is a useful framework for organizing our understanding of intelligence (Brooks 1991b).

Brooks was right to criticize AI for the use of representational schemes with fixed and predetermined interpretations. As a result of moving away from the use of explicit representations, however, too little emphasis has been placed on the "appropriate" role of representation in intelligent problem solving. We want to pair Brooks' insights with a flexible representation that evolves with its interactions within an environment. A new dynamical model of representation, focusing on the role of emergent structure in behaviorally coupled systems, will accompany our new framework for robotics. McGonigle, referring to the polarity between representational stances, claims "we have the concept of a co-evolving agent and environment leading to a mutual specification..." (McGonigle 1998). To explore this new notion of representation, we must develop models that are both dynamical and embodied. Then we must seek mechanisms in those models for the emergence of structures coupled through system behavior to the environment.

Maturana identifies a hallmark of living systems which he calls *structural coupling* (Maturana and Varela 1980). Structural coupling means that the environment triggers changes in the internal structures of a system; but the nature of those changes is dictated by the dynamics of the system rather than being specified by the environment. An "embodied" model is one which participates in the dynamics of its world and which undergoes changes in its internal processes triggered by events in the environment. Representation for a robot control system can be achieved by providing a sufficiently rich dynamical system inside the robot to enable structural coupling to take place between the robot control architecture and the environment.

In spite of admonitions against representation, the use of partial world models may actually increase the ability of dynamical systems to meet the real-time demands of their environments. Clark discusses this in connection with

Kawato's work on proprioception (Clark 1997). Partial models devoted to the improvement of specific behavior are called *niche models* (Clark 1997). Representations can be partial because they derive their meaning from the context of interactions within an environment.

## "Fluid Representation" and Copycat

Copycat (Mitchell 1993) is one of the first computer programs to attempt to capture the dynamical processes from which symbolic or representation-based behavior can emerge. Copycat solves analogy problems such as, if "abc" becomes "abd" what does "ijk" become? Such seemingly simple analogies involve evolving, context-dependent processes of integration and differentiation that are at the core of intelligent problem solving.

In addition to its novel mechanisms for parallelism and flexible adaptation, one of Copycat's most important components is the *slipnet*. The slipnet is a semantic network organized with spreading activation and multiple kinds of links among its nodes, some of which can change in length. The processes which evolve representational structure impact the topology of the slipnet, making the program's own behavior part of the adaptive control. For example, if several interacting processes have successfully built structures about *opposite* relationships among the input, the node for *opposite* in the slipnet becomes more active. Furthermore, *opposite* links become shorter and more likely to be traversed, and further processes to explore *opposite* are generated. Figure 1 shows the lengths of links between two nodes, *successor* and *predecessor*, as 85. This value shrinks as the label node for those links, namely *opposite*, gets an increase in activation (shown inside the ovals), making substitutions of one for the other more likely. In addition to spreading activation, this is the method by which slipnet evolves its meanings in response to events in its environment.

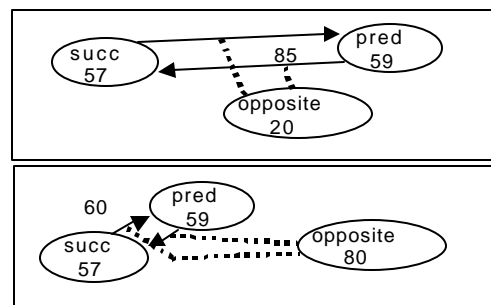


Figure 1: The Evolving Slipnet

Interacting with the slipnet in Copycat are the *coderack* and the *workspace*. The interactions of these three components of Copycat are mediated by the system's *temperature*, which measures the cohesion of the workspace structures. The workspace is a global arena for creating structures that the other components of the

system can inspect. In this sense it is much like a Blackboard (Luger and Stubblefield 1998) or the message area in Holland's (1986) classifier system. Copycat's coderack is a priority biased probabilistic queue containing codelets. Codelets are small pieces of executable code designed to interact with the objects in the workspace, exploring different facets of the problem space and attempting to further some small part of the evolving solution. The codelets are very much like the individual classifiers in Holland's (1986) original system.

Copycat is a unique hybrid between serial and parallel execution, between goal-driven and data-driven search, and in particular between the symbolic and connectionist paradigms. The Copycat architecture models the fluid representation of concepts and their adaptive application to the active construction of features from perceived data.

One limitation of the Copycat program is that it has only one point of interaction with its environment (the initial exposure to the letter-string analogy problem). There are no means for continuing interaction with the external environment, only an ongoing maturation of the internal structures of the program guided by its own context-sensitive semantic network.

A second limitation of Copycat is the program's restricted domain. The domain structure in Copycat, which facilitates exploration of fluid concepts in high-level perceptual processes, also restricts the interpretations available to the program of its developing representation. For example, the relationships possible between structures in Copycat, like *predecessor*, *successor*, and *opposite*, are derived from abstract ordering relationships in the alphabet. We have extended the program to include richer semantic relationships whose application can continue to evolve throughout the program's interaction with its environment. Related issues, for example, the ability to interactively discern new rules and interpretations from observed behavior, are addressed in the Metacat project (Marshall 1999). By using the ideas from Copycat and Metacat in our own embodied world of the robot, we have begun to address these limitations.

## The Madcat Architecture

The Madcat project explores how an architecture similar to Copycat can be used to detect abstract features of sensory data obtained from an ongoing dialog with the environment. With its three mutually self-maintaining components, the slipnet, workspace and the coderack, the Copycat architecture is an autopoietic system and a starting point for a general model of embodied intelligence. Copycat exhibits the characteristics of an evolving complex adaptive system relying on a subsymbolic dynamical system whose structural coupling supports its representation of a domain. In Madcat the emergence of representational structures is coupled to the environment through system behavior.

The Madcat project extends the Copycat architecture to the control system for a robot, producing a control

architecture capable of ongoing interaction with a dynamic environment. The Madcat robot is a Nomad Super Scout II capable of translational and rotational motion with 6 bump sensors, 16 sonar sensors, and a color vision camera (not incorporated into the current model; see Further Research). This collaboration between Copycat and the Nomad robot produced the project name *Madcat*. The ultimate goal of our research is to construct a robot architecture that, from its emergent exploratory behavior, can build a flexible representation of its environment that improves its real-time performance.

In our research we look for behaviors that can be made more effective by niche models (Clark 1997). We build the individual components of the architecture and their rules to interact with data from the sensors and relationships among that data. The resulting emergent structures are correlated with the events in the environment, such as the passing of a corner. The internal "representations" of these events interact with the control system to produce behavior that is based on that "representation".

For example, we overcome certain sensor limitations in the robot using this emergent representation scheme. The maturation of the representation through interaction with the environment is what makes this feasible for a robot whose motion creates constant change in its sensory data. This evolving representation in the behavior-based framework is an important feature of this model.

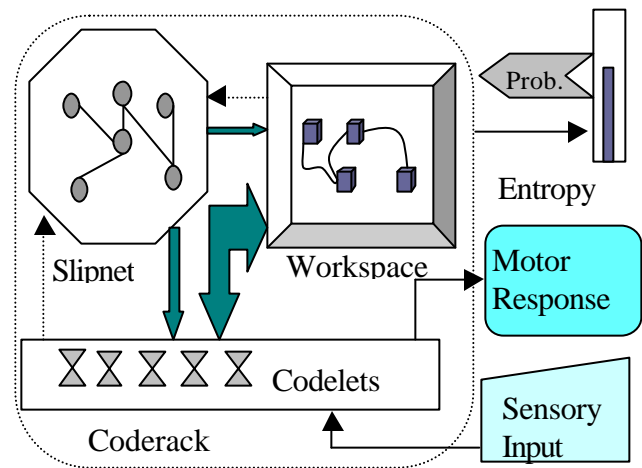


Figure 2: The Madcat Architecture

Figure 2 shows the components of the architecture and their relationships. Simple reflex-like behaviors, such as obstacle-avoidance and wall-following, are achieved by instantiations of four basic rules for a given set of readings (called a *snapshot*). These rules are expressed in codelets with high priorities. The coderack is a stochastic priority queue where the choice of the next codelet is made probabilistically with a bias toward the higher urgency codelets. This provides the flexibility to discover alternate possibilities. For further discussion of the importance of randomness in the coderack and elsewhere see (Mitchell

1993). A codelet is just a C++ object containing only one method that is executed when that codelet object is selected from the coderack. The method may initiate robot movement as in reflex behavior, or it may take a new snapshot and launch further codelets to build emergent structures and generate behavior from them.

The workspace serves as the locus of structure-building activity of the codelets from the coderack. Activity in the workspace biases codelet choices in the coderack. The slipnet contains nodes and links that dictate the data to which the codelets respond and the kinds of structures they build. The slipnet topology changes in response to activity in the workspace but its nodes and links remain fixed. The *entropy* reflects how well emergent structures fit into the data the robot encounters and affects the biases of the system. A high entropy inclines the system toward random behavior and perception of different patterns in the data. With low entropy the system gravitates toward the established structures.

The control functions for the robot are made available as C functions that can be linked into developed software. The Madcat architecture itself is implemented in C++. Besides the C-based interface of the robot, the choice of C++ was dictated by the need for real-time behavior. We are building in Java an interface to the architecture that will be used as a development and testing tool.

## The Behavior of Madcat

The first goal of the Madcat architecture was to demonstrate that certain basic competencies, roughly those of Brooks (1991a), could be implemented using this emergent architecture. The chosen behaviors are obstacle avoidance, wandering, and wall-following. Obstacle avoidance is defined as the behavior of moving to avoid a collision. Wandering is defined as the behavior of choosing a random direction of motion when no other particular movement is required. We define wall-following as the behavior of moving approximately parallel to the nearest surface, without necessarily moving nearer to that surface to do so.

In the behavior-based approach of Brooks (1991a) these behaviors would be supported by individual interacting layers, each capable of a particular behavior. In an emergent architecture, such as Madcat, a few simple rules interacting among all the data readings give rise to the appropriate behavior. Instead of layers, an emergent architecture relies on competition between peer behaviors to generate coherent global behavior.

There are four basic rules for responding to the data readings. These have been determined empirically by considering immediate needs of particular elements, as is done in cellular automata for instance. Genetic algorithms, reinforcement learning, or other methods might also be used. For readings that come from the sonar sensors above either wheel the robot should move forward to follow the surfaces which reflected the signals from those sensors. For readings that come from sonar sensors clockwise from

either wheel but not beyond the forward or rear sensors the robot should rotate clockwise to become parallel with the surfaces that reflected the signals from those sensors. An analogous rule holds for readings from those sonar sensors counterclockwise from the wheels. If the robot senses contact from one of the six regions of the bump sensor, then it should back up a small amount and turn away from the region to avoid further contact. When each of these rules is given a priority proportional to the proximity of the readings, the desired three behaviors emerge as a result of the moment by moment interactions of the rules, readings, and features of the environment.

Wall-following can be seen in Figure 3 where the robot moves counterclockwise, turning corners to remain on a course parallel to the nearest wall. Obstacle-avoidance is also demonstrated, as the robot turns in response to surfaces detected in its path. Wandering is subordinate to these first two behaviors and so only appears at the end of the path in the upper left corner.

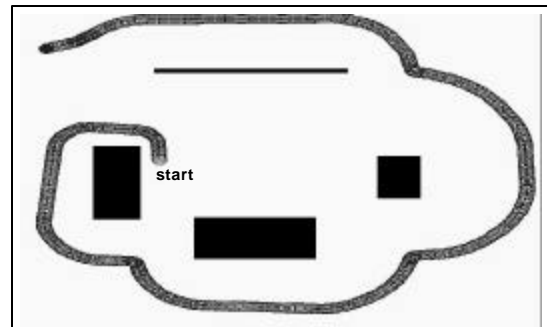


Figure 3: Obstacle Avoidance, Wandering, and Wall-Following

The second goal of the Madcat architecture is to generate emergent structures correlated with environmental features. These support more effective real-time behavior. For example, the direction choice for wandering can be made more useful if the system has a rough model of what it has already encountered. Random directions can be chosen from among those not yet explored. As another example, consider that the sonar sensors produce the same measurement for all readings below 6 inches, preventing the distinction of a corner from a continuation of a nearby wall. If the system contains structures representing a wall located directly ahead, it may use this information to turn away from the wall with which it would otherwise collide.

At the top of Figure 4 the robot passes a convex corner. *Single Surface Element (SSE)* structures, corresponding to each of the sonar readings taken while traversing this path, are built in the workspace. Bonds can be built between these SSEs according to the relationships in the data. For instance, *Adjacent Equivalence Bonds (AEB)* may be built between SSEs from adjacent sonar sensors if their values are within a certain percentage of each other. *Candidate Surface Bonds (CSB)* tend to be built linking a sequence of AEBs, which might possibly constitute a surface. Bonds built from a single snapshot are only tentative. As the data from successive snapshots continue to bear certain

relationships, bonds based on those become strengthened. The *Maximum Difference Bond (MDB)* identifies the apexes in curved surfaces. These only occur after many snapshots have produced well-established structures. Figure 4 illustrates this process. There is no attempt to maintain a direct spatiotemporal correlation between internal structures and external features; rather the relative importance of the structures dictates the ones to which the robot's behavior responds.

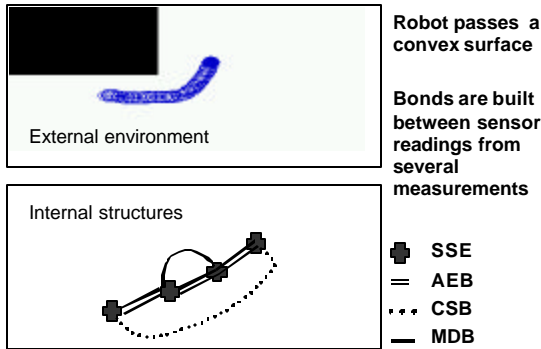


Figure 4: Emergent Structures Form in Response to Environmental Features

Figure 5 shows the robot approaching a wall to which its sensors are blind. The wall to its left is closer than six inches, below which distance the sonar system is unable to make any distinctions. This makes the approaching wall look like a continuation of the wall to the left. However, during the approach, structures will form which reflect the sonar readings of the forward wall. If a CSB is built in time, the robot will notice it when scanning its internal surfaces for discrepancies with the environment. At that point it can choose to turn and avoid the wall based on its internal niche model of the world. This will demonstrate the use of emergent representation to improve real-time behavior. We expect many similar improvements to be possible based on the emergent representation.

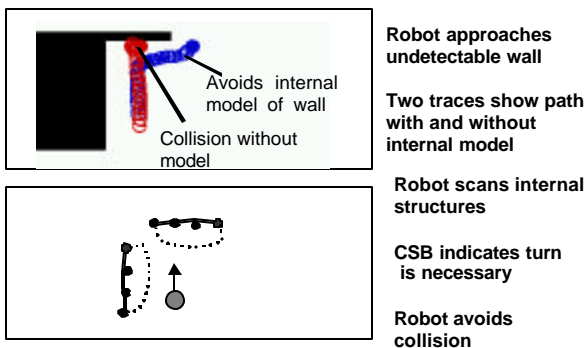


Figure 5: Emergent Structures Aid in Navigation

The role of the slipnet is to provide context-dependence to the competing behaviors in Madcat. For example, consider the creation of an AEB, proposed by some codelet. The comparison of values between adjacent SSEs uses information from the slipnet concerning relative

distances of objects in the current environment to discern how precisely the comparison should be made. When the objects detected are at a greater distance from the robot both trigonometric considerations and reliability of the sensors dictate that a greater difference in readings may still correspond to a single surface. Alternatively, when the robot is near its targets, a small difference between surfaces by adjacent sensors more likely indicates distinct surfaces. As another example, the SSEs between which the AEB will be built are themselves chosen probabilistically with a bias coming from the slipnet's indications of which objects have greatest relevance at that moment. Indeed every time a codelet must choose an object on which to perform an operations (e.g. build a structure around it) the bias for the probabilistic choice is made based on the activation level of the nodes in the slipnet associated with the object and the action of the codelet.

Occasionally, the parallel nature of the architecture will give rise to the proposed construction of an object that conflicts in some way with an existing object (e.g., duplication, overlap, and opposition). As in the Copycat architecture, the choice of whether to veto the construction or destroy the conflicting object and continue is made probabilistically with a bias that comes from information in the slipnet about which kinds of objects are currently more useful to build. This information comes from the context to which the slipnet has been exposed in the preceding moments of the robot's behavior. Indeed at times the priorities implicit in the current arrangement of the slipnet will bias the probabilistic codelet executions so that the system explores otherwise unnoticeable options.

The entropy measure, like the temperature in Copycat, is used as a feedback mechanism for the entire architecture. When entropy of the workspace is calculated, values are obtained from the workspace objects that indicate their relative importance and degree of incorporation into larger structures. The calculation of these values includes the level of activation of the node in the slipnet corresponding to that type of object. So an object whose node in the slipnet has high activation is likely to have greater importance and higher expectation for structure-inclusion. Thus, even the self-organizing feedback in the system is mediated by the context-driven relevance of the concepts in the system. Information in the slipnet about relative priorities of certain kinds of structures and actions can be used to select or restrict entire classes of behavior.

The slipnet captures this context information through its interactions with the workspace and the codelets. When a codelet successfully builds a structure in the workspace, the slipnet node which originated that codelet gets a boost of activation. That activation spreads to neighboring nodes in the slipnet as a function of the length of the link between them. Thus, related nodes also get some additional activation. As the activation of a node goes up, so does its chances of emitting codelets designed to explore the possibility of building structures in the workspace based on the concept represented by that node. Activation decays in the slipnet so that over time, if no new

objects of a given type are being built, then codelets stop being produced to look for them. Of course there is a certain low probability for generating any type of codelet so the system never stops discovering new possibilities. The mechanisms of the slipnet capture the priorities indicated from the context of recent interaction of the environment and drive the decisions in the entire system.

### Further Research

There are two specific areas of further development. The first is to use the internal models of environmental features to augment visual decomposition algorithms used with the color vision camera. The worm algorithm (McGonigle 1998) is commonly used, but it is easily misled. The presence of sonar edges in the internal model can help to corroborate edges found by a variant of the worm algorithm. This kind of synthesis is important in the intelligence of living organisms. We would like to build models with this capacity.

The second extension to our research is related to the idea that events in the environment enable certain behavior sets and disable others. We would like to model the sudden shift of priorities and behaviors in a system in response to events in the environment. Certain colors detected by the camera act as triggers for the system. When these occur, changes in the links in the slipnet and the priorities of codelets occur which override the bias to explore and complete internal models in favor of seeking out a resource or avoiding danger.

### Conclusion

We offer both a definition and an instantiation of intelligent problem solving in robotics based in evolving complex adaptive systems. We refine the behavior-based approach to robotics by requiring that representation, redefined as the emergence of structures coupled to the environment through behavior, be given greater focus. We believe that the four issues of embodiment, emergence, symbolic behavior, and representation will be very important in the challenging task of understanding intelligent activity in changing problem domains.

We have demonstrated the feasibility of an emergent architecture in solving simple robotics problems. We have demonstrated that emergent structures in an embodied architecture can be behaviorally correlated to features of the environment, producing niche models useful for generating adaptive behavior. Work is underway using this architecture for improved visual decomposition algorithms and environmentally triggered behavior shifts.

### Acknowledgments

This research has been supported at the University of New Mexico by the NSF CISE Research Infrastructural award CDA-9503064 and by the NASA PURSUE Program (PAIR)

Grant No. NCC5-350. The contributions of Andy Claiborne, Matthew Fricke, Tim Mitchell, Deborah Pearlman, Monica Rogati, and Len Lopes have been invaluable.

### References

- Brooks, R. (1991a). *Intelligence Without Representation*. Reprinted in Luger, G. (ed). 1995. *Computation & Intelligence*. 343-364. Cambridge: MIT Press.
- Brooks, R. (1991b). New Approaches to Robotics. *Science* 253:1227-1232.
- Brooks, R. and Stein, L. (1994). Building Brains for Bodies. In *Autonomous Robots 1*: 7-25. Boston: Kluwer Academic Publishers.
- Clark, A. (1997). *Being There*. Cambridge: Bradford Books/MIT Press.
- Holland, J. (1998). *Emergence: From Chaos to Order*. Reading, MA: Perseus Books.
- Holland, J. (1986). Escaping Brittleness: The Possibilities of General-Purpose Learning Algorithms Applied to Parallel Rule-Based Systems. Reprinted in Luger, G. (ed). 1995. *Computation & Intelligence*. 275-304. Cambridge: MIT Press.
- Luger, G. (1994). *Cognitive Science: The Science of Intelligent Systems*. San Diego: Academic Press.
- Luger, G. and Stubblefield, W. (1998). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. London Addison Wesley.
- Marshall, J. (1999). *Metacat: A Self-Watching Cognitive Architecture for Analogy-Making and High-Level Perception*. PhD Dissertation, Indiana University. Bloomington, IN.
- Maturana, H. and Varela, F. (1980). *Autopoiesis and Cognition*. Dordrecht, Holland:D. Reidel.
- McGonigle, B. (1998). Autonomy in the making: Getting robots to control themselves. In *International Symposium on Autonomous Agents*. Lanzarote: Oxford University Press.
- Mitchell, M. (1993). *Analogy-Making as Perception*. Cambridge: Bradford Books/MIT Press.
- Steels, L. (1996). The origins of intelligence. In *Proceedings of the Carlo Erba Foundation Meeting on Artificial Life*. Berlin: Springer-Verlag.
- Steels, L. (1995). Intelligence - Dynamics and Representations. In *The Biology and Technology of Intelligent Autonomous Agents*. Berlin: Springer-Verlag.
- Winograd, T. and Flores, F. (1986). *Understanding Computers and Cognition*. Norwood, N.J.:Ablex.



# Point-Light Displays Illuminate the Abstract Nature of Children's Motion Verb Representations

Jing Liu (jingliu@udel.edu)

Roberta M. Golinkoff (roberta@udel.edu)

Kim Piper (kimpiper@udel.edu)

School of Education, University of Delaware, Newark, DE 19716

He Len Chung (hlchung@astro.ocis.temple.edu)

Kathy Hirsh-Pasek (khirshpa@nimbus.ocis.temple.edu)

Christopher H. Ramey (rameychb@aol.com)

Department of Psychology, Temple University, 1801 North Broad Street, Philadelphia, PA 19122

Bennett I. Bertenthal (bbertent@dura.spc.uchicago.edu)

Department of Psychology, University of Chicago, 5801 South Ellis Ave., Chicago, IL 60637

## Abstract

The present studies explored children's representations of motion verbs first, in an elicited production study using point-light displays (lights attached to the joints of the human body) and then in a language comprehension task using the Intermodal Preferential Looking Paradigm (IPLP). Results indicated that children indeed ascribe meanings to the portrayals of actions in point-light displays. When children could not spontaneously produce the actual verb for an action, they used either a more specific or a more general familiar verb that was considered appropriate by adults. These findings suggest that even by the age of 3, children's representations of the actions that verbs label are amazingly abstract. This is the first set of studies to probe the nature of children's verb representations under circumstances where the portrayal of the action is stripped of an apparent agent, a location, instruments, or in some cases, the objects ordinarily required in transitive actions (e.g., a shovel in shoveling). Using point-light displays provides the field with a methodological tool for exploring the components of verb representation in both children and adults and for investigating children's verb acquisition.

## Introduction

The purpose of this paper is to take a closer look at verb representations and to examine the question of how 3-year-olds, who already know many verbs, extend familiar verbs to novel events depicted in point-light displays. To extend a verb to a point-light action, children must activate their action representations and map the patterns of light sequences to their verb representations. The use of point-light stimuli provides a stringent test of children's motion verb representations due to the removal of context.

Considering that the semantic structure of verbs provides a kind of conceptual frame for constructing larger linguistic units such as phrases and sentences, verb learning is of central importance for young language learners. Since motion words are among children's first words (e.g., Bloom, 1993; Smith & Sachs, 1990), two essential questions are

whether children can successfully form word-action mappings and how they extend an action verb.

Research has found that infants are keenly aware of movement and can use movement to individuate objects and actions (Sharon & Wynn, 1997; Spelke, Katz, Purcell, Ehrlich, 1994). There is also evidence that 18-month-olds can distinguish the causal actions of push and pull and map novel words to these actions after limited exposure to the word-action pairs (Casasola & Cohen, in press). However, it is not clear what perceptual cues children use to form word-action mappings and how they decontextualize the use of verbs to successfully extend verb labels to the actions they witness. For example, how do children understand that the word "jump" refers to a category of jumping motions that include different kinds of jumps made by the same actor (e.g., Elmo jumping off tables and chairs), and the same action performed by different actors (e.g., Elmo or Lala jumping off the chair)?

One line of research suggests that children use lexical principles to narrow down the possible meanings of words (Golinkoff, Mervis, & Hirsh-Pasek, 1994). According to Golinkoff, Mervis, Hirsh-Pasek, Frawley, and Parillo (1995), lexical principles guide children to learn not only object nouns but also action verbs. For example, the "principle of extendibility" posits that nouns not only label the original exemplar but a category of objects of "like kind." When transferred to the acquisition of action verbs, this principle states that an action verb, like a label learned for an object, can be extended to more than just the original action. Yet, what is the basis for verb extension?

Golinkoff et al. (1995) suggested that shape, or the overall motion configuration of the action, may provide a basis for children's verb extension. The shape of an action can be primarily affected by the path of motion (horizontal vs. vertical such as walking versus jumping), the involvement of arms and/or legs, and the type of instruments, if any, involved in the action. With respect to nouns, shape provides important information about the function and categorical membership of an object and serves as an

important basis for object noun extension (Landau, Smith, & Jones, 1998). It is quite likely that, when no other information is accessible, the shape of an action will provide defining information about the event, as “many verbs of motion have ... a typical appearance, a physiognomy” (Marconi, p. 159). This typical appearance of action may represent what Pinker (1989) labeled the “shape” of an event, and provide the basis of children’s verb extension. In fact, certain semantic factors such as MANNER, PATH, and CONVEYANCE, seem to be embedded in the “shape” of an event (Talmy, 1985). By attending to the overall shape or configuration of an action, children may ignore the context details of the individual event and construe the semantic invariant that a verb encodes. This would result in a more abstract and flexible representation of the action to which a particular verb can be mapped. For example, while the particular individual or object is not necessarily a part of the meaning of the verb FALL, the so-called shape of downward trajectory does represent the core, typical appearance of the action “falling.”

The claim that shape serves as an important basis for verb extension is in accord with Gibson’s (1966) view that in perceiving events we detect the “invariants” that persist from one event to another of the same type. The overall configuration of an action may be an invariant of the event. This claim is equally in line with Mandler (1992) who argues that events are stored as “image schemas” or “dynamic analog representations” abstracted from children’s interpretations of the spatial relations between objects. These meaningful image schemas help reduce the infinitely varying perceptual displays into a limited number of meaningful concepts that can be described by words.

Here we argue that children may use the “shape” of an action that loosely represents the invariant cues from one action to another as a basis for verb extension. It should be noted, however, that in using terms such as “invariant,” we are not assuming that the representations themselves are fixed and rigid. On the contrary, we posit that the shape of an action is a prototypical representation that is flexible enough to include actions that are similar to, but do not exactly match the original exemplar in terms of, say, the agent and location.

Point-light displays were first used to study adults' event perception and biomechanical motion by attaching small lights to the head and main joints of an individual’s body and filming the person performing different actions against a dark background (Johansson, 1973). Because point-light displays are deprived of detailed contextual information such as the agent and location of an event, the information about an action given in such displays is expressed in the overall shapes of the light sequences. Previous studies using these moving light displays with infants and adults demonstrate the significance of prior knowledge in the perception of biomechanical point-light images (Bertenthal, 1993). For example, while 5-month-old infants can discriminate between a point-light walker shown in an

upright versus upside-down orientation, 3-month-olds do not demonstrate this sensitivity (Bertenthal, 1993). Thus, whether children can identify the actions depicted in point-light displays based on their previous experience provides a strong test of the hypothesis that toddlers use abstract, shape-based event representations to extend familiar motion verb labels. Research has found that 3-year-olds use many motion verbs, we therefore first investigated 3-year-olds' ability to identify point-light depictions of human actions.

## Experiment 1

This study explored 3-year-olds' ability to spontaneously produce a label for an action depicted in point-light displays. Since children do not encounter point-light displays in everyday life, to successfully complete this task, children must perceive the patterns of lights as meaningful action sequences, activate their verb representations for these actions, and ascribe meanings to these point-light sequences.

### Method

#### Participants

Thirty-eight children participated in the study. The final sample had 29 children, 16 boys and 13 girls, mean age = 3 years 7 months. Nine children were excluded from the data because 3 failed to produce a description for more than 3 of the actions and 6 either refused to talk to did not finish the study.

**Stimulus Videotapes** Biomechanical displays of motion verbs were created by videotaping a person in action with light-emitting diodes (LEDs) affixed to the major joints of the body (ankles, knees, hips, wrists, elbows, shoulders). These displays consisted only of a collection of white dots moving across a black screen. Figure 1 provides an example of a person walking, translated into frozen still images of point lights.

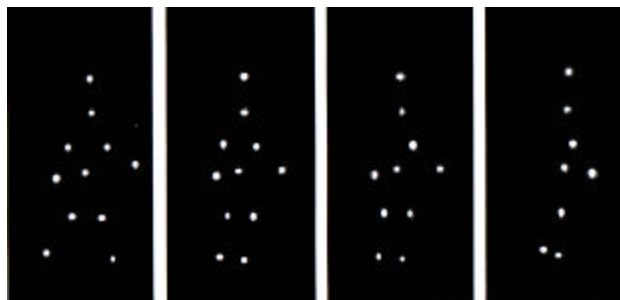


Figure 1: Canonical Point-light Walker

Each of the 8 actions was performed continuously for 3 seconds and repeated 4 times so that each action was

displayed for 12 seconds, followed by 5-second blank tape (Table 1). The actions were randomly ordered into two different sequences. To familiarize subjects with point-light displays, a point-light depicted cat walking from right to left on the screen was presented first.

### Procedure

The child sat on the parent's lap in front of a 32" TV. The parent was told not to say anything. Then the experimenter told the child they were about to see a fun videotape and would be asked to tell what they saw. After the cat display, the experimenter paused the VCR and prompted the child three times for a label for the action. If the child only produced an object label, such as "a cat" or "a doggie," the experimenter would probe again, "What was the cat/doggie doing?" If the child produced an unrelated answer such as "dots" or "snow," the experimenter would label the cat for the child and encourage the child to label the action. If the child did not produce an action label after being probed twice, the experimenter would produce the label, "Was the cat walking?" The same procedure was then repeated for each of the human actions except that the experimenter did NOT produce any description for the human action. If the child failed to produce a description for a human action after being probed twice, the experimenter went on to the next action. The experimenter made neutral comments as a response.

Table 1: Descriptions of the Actions in Point-light Displays

- \*Walking: Person walking.
- \*Dancing: Person twisting in place.
- Shoveling: Person bending at the waist and standing back up (without moving legs) as if shoveling snow.
- Picking flowers: Person bending at the waist and standing back up (without moving legs) as if picking up flowers from the ground and then putting them back in a handheld basket.
- \*Rolling: Person performing a somersault.
- \*Running: Person running.
- \*Skipping: Person skipping.
- \*Hopping: Person hopping on one foot.

\* Person moves diagonally from top left corner to bottom right corner across the screen.

**Data coding** Children's productions were divided into two groups: relevant vs. irrelevant. Words or phrases that did not indicate any motion were considered irrelevant, such as "have no feet." The rest were considered relevant. A group of college students (n = 15) were asked to rank the appropriateness of each relevant response on a 1-7 Likert scale after seeing each action in the point-light displays in the same way as children did.

Children produced a total of 110 different descriptions for the 8 actions. After eliminating 7 irrelevant answers and

combining 25 relevant answers by ignoring verb objects (i.e., "picking up snowballs" and "picking up stuff" were considered the same), there were 78 responses left. Each action had a range of 4-15 responses from children. These were listed on the ranking sheet. To examine the consistency of adults' ranking, 77 responses were used as foils for each other (the response "moving" was excluded as too general). The foils were distributed in ranking sheet for each action such that if adults were consistent in their ranking, these foils would receive on average lower rankings than the non-foils. The proportion of actual responses and foils for each action was 1:1, resulting in a total number of 154 responses for the 8 actions on the ranking sheet.

Table 2. Number of Children Who Gave Highly Appropriate Responses (standard deviation in parentheses)

Action	# of Children <sup>1</sup> /Total <sup>2</sup>	Mean Rating (>=5) <sup>3</sup>	Mean Rating (all responses)
rolling	10 / 25	6.33 (.70)	3.87 (2.00)
dancing	15 / 28	6.91 (.57)	3.17 (2.14)
picking flowers	10 / 25	6.41 (.68)	4.31 (2.22)
running	18 / 27	6.74 (.07)	4.65 (2.56)
walking	26 / 29	6.97 (.28)	4.77 (2.63)
hopping	16 / 27	6.78 (.70)	4.10 (2.42)
skipping	6 / 25	6.74 (.42)	3.46 (1.72)
shoveling	3 / 19	5.80 (.58)	2.69 (2.13)

1. Number of children who gave highly appropriate responses.
2. Total number of children who gave responses. Note that some children did not give responses for some actions.
3. Average rating for the highly appropriately responses.

### Results

Overall, most children produced labels for all actions after one prompt. Adults' rating for children's actual responses (M = 3.80, SD = 2.10) were significantly lower than for the foils (M = 1.45, SD = .84),  $F(1, 155) = 84, p < .001$ . Based on the adults' ratings, on average children's responses for most of the actions were considered appropriate (Table 2). Furthermore, 58% of children's responses were considered very appropriate (M = 6.73, SD = .58. All ratings greater than or equal to 5 were considered very appropriate). There was a significant correlation between the number of children who produced highly appropriate responses and adults' appropriateness rating of those responses,  $r = .36, p = .001$ , suggesting that the more appropriate a response was, the more children gave that response. However, few appropriate answers were given for the actions SKIPPING

and SHOVELING. Ten children were unable to produce a label for SHOVELING and 4 failed to produce a label for SKIPPING. The majority of others described SHOVELING as “dancing” or “exercising;” and SKIPPING as “running,” “jumping” and “walking.”

## Discussion

Findings of this study suggest that 3-year-olds can indeed perceive the abstract images depicted in point-light displays as meaningful actions based on the patterns of light sequences. Overall, 77% of the children produced responses for all actions. Given that these children had never been exposed to point-light displays, they must access abstract verb representations in order to identify these contextually bare actions. However, omission of objects in point-light displays might have increased the difficulty of identifying actions involving instruments, such as SHOVELING. In addition, some children's descriptions of SKIPPING overlapped with their description of WALKING and RUNNING, the overall shape of which are similar but for the differences in the movement of arms and legs. This interestingly indicates children's reliance on the overall configurations of these actions. Perhaps this problem, too, was caused by the nature of spontaneous production, a demanding task. Additional data of action labeling from 5 adults (100% correct) suggest that correct verbs can be produced for these abstract point-light displays. If children indeed possess abstract, shape-based verb representations and point-light displays can capture the properties of these actions, children should be able to map a familiar verb to these actions in a less demanding task.

## Experiment 2

The purpose of this study was to determine whether 3-year-old children could correctly extend familiar verbs to actions depicted in point-light displays in a comprehension task.

## Method

### Participants

Children's comprehension of at least 7 out of the 8 action verbs was confirmed with parents on phone calls. Out of the final sample of 32 subjects (44 were tested), 15 did not understand “skipping” and 3 did not understand “shoveling.” The age selected for testing was determined empirically from these phone. The final sample had 19 boys and 13 girls, age range from 2;11 to 3;2, mean age = 3;1.

**Stimulus videotapes** Two separate tapes were created, each containing half of the actions, paired in length and number of frames. The pairs of actions were created such that the lights appearing in the displays were balanced for size, number, brightness, and movement to ensure that each pair of actions was equally salient to children (Table 3). A female speaker recorded the linguistic stimuli in infant-

directed speech for all trials, as well as between the trials, on one track of the videotape.

### Procedure

Children were tested in the Intermodal Preferential Looking Paradigm (Hirsh-Pasek & Golinkoff, 1996). The child sat on the parent's lap in front of two 19" TV monitors. The videotapes were played in complete synchrony, accompanied by the linguistic stimuli which emanated from the center of the two monitors.

**Familiarization trial** The study began with a brief, 6-second trial during which a point-light display of a cat walking across the screen appeared simultaneously on both monitors for 6 seconds. The cat and its action (i.e., “See the cat walking!”) were labeled to give children some familiarity with interpreting the contents of these motion-specified stimuli.

**Salience trials** Two salience trials followed during which a pair of actions appeared simultaneously on both screens, one action on the left monitor and another on the right. Salience trials had three purposes: First, they showed children that contrasting events could appear on both screens at the same time. Second, they were used to calculate stimulus salience. Finally, they provided exposure to the names of the actions without telling children which screen either action was on, e.g., “Hey, one is walking and one is dancing!” Thus, children were directed to watch both screens.

**Test trials** Two test trials followed to see if the child could distinguish between the displays and successfully choose the action that matched the linguistic stimulus. Now the child was exhorted to watch the screen containing the labeled verb, e.g., “Look at dancing! See dancing?” The target action appeared on the same screen side for both test trials in a block, the same side as the two preceding salience trials.

**Intertrial intervals** Each trial was separated by a 3-second intertrial interval during which both screens went blank. A red light mounted centrally between the two televisions lit up during this time to attract children's attention to the center, off the screens. This practice ensured that children would not just remain on one screen for long periods of time, but would have to choose which screen to look at for each trial. The appropriate linguistic stimulus for the trial to follow was first heard during the intertrial interval, so that prior to each test trial, the child was directed to find the matching screen (e.g., “Can you find dancing?”).

**Apparatus and Data Coding** All equipment – except for the two 19" color monitors – was shielded from the child's view. The videotapes were shown on 3/4" video decks. A 1 KHz tone was recorded for the duration of each trial on the

second, inaudible channel of the videotape and was “read” by a specially designed tone decoder which functioned in two ways: 1) it turned the centrally-mounted red light on during the intertrial intervals, and off during the trials; and 2) it signalled the beginning and end of each trial to the computer (a PC computer).

**Dependent and Counterbalanced Variables** The dependent variable was the mean visual fixation time to the named action (the match) versus to the foil (the non-match) during each pair of the test trials. For each test trial, visual fixation time was collected starting during the intertrial interval from the point at which a child watched the center light for .3 seconds or more. Coding began during the intertrial interval because if a child failed to reach the .3 second intertrial interval criterion on a trial, that trial was not included in the data analysis (this occurred on only 5 trials, or 2% of the time). When a child missed a trial, his or her overall visual fixation mean to the match and non-match across the remaining test trials was substituted in that cell. Thus, each child contributed 8 data points to the analysis: the mean visual fixation time to the match and to the non-match for each of 4 pairs of test trials.

Four factors relating to order of stimulus presentation were counterbalanced across subjects: 1) the number of matches on a screen side; 2) the order of the matches; 3) the order of the two actions mentioned during the salience trial; and 4) the member of a verb pair labeled as the match.

## Results

Comparison of mean visual fixation times during the salience trials in the three-way mixed ANOVA (sex (2) X verb pair (4) X match versus non-match (2)) suggested that there were no a priori preferences for one action or another in any pair (Table 3). However, a significant difference was found between the mean visual fixation time to the match ( $M = 3.36$ ,  $SD = 1.85$ ) vs. the nonmatch ( $M = 2.29$ ,  $SD = 1.44$ ) during the test trials,  $F(1,3) = 48.81$ ,  $p < .01$ . This effect was carried by the vast majority of children. Out of 32 subjects, 29 or 91% had mean visual fixation times in favor of the match for all of the four verb pairs.

## Discussion

When the motion-specified point-light images were presented in the IPLP (Hirsh-Pasek & Golinkoff, 1996), children who were 6 months younger than those in Experiment 1 demonstrated extension of all familiar verbs by watching the screen that matched the requested verb more than the nonmatch screen. Apparently, children could attend to the differences between the actions depicted in point-light displays in a comprehension task. Children could even map verbs to actions with which they were not familiar, such as SKIPPING. Perhaps their recognition of the unfamiliar verb was due to their successful mapping of a familiar verb to a familiar action presented as a comparison

in the IPLP. These results provided evidence that the combination of point-light displays and IPLP could be a powerful and sensitive tool to investigate children’s motion verb representations.

Table 3 Mean visual fixation times (in seconds) to the four stimulus pairs during salience and test trials (M = Match; NM = Non-match)

Verb pairs*	Saliency Trials	Test Trials**
	M (SD)	M (SD)
	NM (SD)	NM (SD)
Walking vs. Dancing	2.74 (1.46)	3.67 (2.29)
	2.82 (1.55)	2.22 (1.20)
Picking flowers vs. Shoveling	2.49 (1.15)	3.31 (1.69)
	2.84 (1.29)	2.37 (1.50)
Running vs. Rolling	2.39 (1.05)	3.40 (1.87)
	2.71 (1.18)	2.17 (1.37)
Skipping vs. Hopping	2.73 (1.49)	3.16 (1.56)
	2.53 (1.46)	2.05 (1.71)

\*The verb requested in each pair (i.e. match) was counterbalanced.

\*\* The match was watched more than the non-match for all test trials,  $p < .05$ .

## General Discussion

The purpose of these studies was to determine whether 3-year-olds could correctly extend familiar verbs to actions depicted in point-light displays. Experiment 1 found a majority of 3-year-olds could accurately identify the actions depicted in point-light displays. When children could not produce an accurate label for the point-light actions, they often used familiar verbs for actions that had similar overall shape to the target action. More than half of children’s responses (58%) were considered very appropriate by adults. Experiment 2 found that children who had just turned 3 could recognize all of the actions in the Intermodal Preferential Looking Paradigm (Hirsh-Pasek & Golinkoff, 1996), suggesting that the overall shape of an action could be a reliable basis for motion verb extension.

This is the first set of studies to suggest that young children can identify dynamic, complex events depicted in point-light displays and extend familiar verbs to the actions embedded in these events. Previous research focused on infants’ discrimination of familiar and unfamiliar biomechanical images (Bertenthal, 1993), and adults’ recognition of the familiarity and gender of a point-light walker (Kozlowski & Cutting, 1977). No work, however, had assessed the utility of these displays with young

children, let alone combining the stimuli with a language task. Point-light displays, which permit the use of dynamic events and contain so little contextual information, could serve as a critical tool to probe children's verb representations. Since the only available information in such abstract images was the "shape" of events, or an action's overall motion configuration, children's success in identifying these actions suggests that shape may be an important component of children's motion verb representations that guides motion verb extension.

We are not implying that children rely only on shape for motion verb extension. However, without disputing children's use of other complex verb learning processes (e.g., syntactic analysis), we underscore the flexibility of children's motion verb representations and the demonstration of their productive reliance on shape as one cue for the extension and categorization of motion verbs.

It is also important to note not all verb types can be illustrated through point-light displays. Verbs such as "see" or "think," for example, cannot be easily depicted through movie sequences, while motion verbs are ideal. Nonetheless, since motion verbs are generally among the first verbs acquired, future research employing these images may help uncover how early verb categories develop. Given that these contextually deprived displays can make contact with motion representations held by infants as young as 5 months, and that the children in this study could map a verb correctly to one of two choices, it appears that these stimuli could be used with older infants and children as well. Investigators could examine at what point young language learners are able to attach verb labels to these abstract images, and more specifically, when "shape" becomes a meaningful cue for extending novel verbs and forming motion verb categories. These types of research efforts can bring the critical study of verbs to the foreground, and help advance the understanding of fundamental verb learning processes that have for too long been neglected.

## References

- Bertenthal, B. I. (1993). Perception of biomechanical motions: Intrinsic image and knowledge-based constraints. In C. Granrud (Ed.), Carnegie symposium on cognition: Visual perception and cognition in infancy (pp. 175-214). Hillsdale, NJ: Earlbaum.
- Bloom, L. (1993). The transition from infancy to language: Acquiring the power of expression. New York: Cambridge University Press.
- Casasola, M. & Cohen, L. B. (in press). Infants' association of linguistic labels with causal actions. Developmental Psychology.
- Gibson, J.J. (1966). The senses considered as perceptual systems. Boston, MA: Houghton-Mifflin.
- Gibson, J.J. (1979). The ecological approach to visual perception. Boston, MA: Houghton-Mifflin.
- Golinkoff, R. M., Mervis, C. B., Hirsh-Pasek, K. (1994). Early object labels: The case for lexical principles. Journal of Child Language, 21, 125-155.
- Golinkoff, R. M., Hirsh-Pasek, K., Mervis, C. B., Frawley, W. B., & Parillo, M. (1995). Lexical principles can be extended to the acquisition of verbs. In M. Tomasello & W. E. Merriman (Eds.), Beyond the names for things: Young children's acquisition of verbs (pp. 185-221). Hillsdale, NJ: Lawrence Earlbaum Associates, Inc.
- Golinkoff, R. M., Jacquet, R., Hirsh-Pasek, K., Nandakumar, R. (1996). Lexical principles underlie verb learning. Child Development, 67, 3101-3110.
- Hirsh-Pasek, K., & Golinkoff, R. M. (1996). The origins of grammar: Evidence from early language comprehension. Cambridge, MA: MIT Press.
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. Perception and Psychophysics, 14, 201-211.
- Kowzowski, L. T., & Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point-light display. Perception and Psychophysics, 21, 575-580.
- Landau, B., Smith, L. B., & Jones, S. (1998). Object perception and object naming in early development. Trends in Cognitive Sciences, 2, 19-24.
- Mandler, J.M. (1992). The foundations of conceptual thought in infancy. Cognitive Development, 7, 273-285.
- Marconi, D. (1997). Lexical competence. Cambridge, MIT Press.
- Pinker, S. (1989). Learnability and cognition: The acquisition of argument structure. Cambridge, MA: MIT Press.
- Sharon, T. & Wynn, K. (1997). Individuation of actions from continuous motion. Psychological Science, 9, 357-362.
- Smith, C.A., & Sachs, J. (1990). Cognition and the verb lexicon in early lexical development. Applied Psycholinguistics, 11, 409-424.
- Soken, N.H. & Pick, A.D. (1992). Intermodal perception of happy and angry expressive behaviors by seven-month-old infants. Child Development, 63, 787-795.
- Spelke, E.S., Katz, G., Purcell, S.E., & Ehrlich, S.M. (1994). Early knowledge of object motion: Continuity and inertia. Cognition, 51, 131-176.
- Talmy, L. (1985). Lexicalization patterns: Semantic structure and lexical forms. In T. Shopen (Ed.), Language typology and syntactic description: Vol. III. Grammatical categories and the lexicon. New York: Cambridge University Press.

# Learning at Different Levels of Abstraction

Bradley C. Love  
Department of Psychology  
The University of Texas at Austin  
Austin, TX 78712 USA  
*love@psy.utexas.edu*

## Abstract

Previous category learning research and the SUSTAIN (Supervised and Unsupervised STRatified Adaptive Incremental Network) model of category learning suggest that preferred category level (in a hierarchy of categories) shifts towards lower-level (i.e., more specific) categories when stimuli are perceived to be more distinctive. This shift is in accord with work in expertise. In their domain of expertise, experts excel (relative to novices) at classifying stimuli at lower category levels, but their advantage is attenuated with higher-level categories. The work described here directly tests (within a single study) this predicted interaction between category level and stimulus distinctiveness using well controlled artificial stimuli. The results are consistent with prior work utilizing natural stimuli. The results are also informative for evaluating whether attention is dimension-wide (i.e., all items are represented in a common multi-dimensional space of the same extent) or cluster specific (i.e., different conceptual clusters can stress different stimulus dimensions so that different aspects of different stimuli are stressed). The results suggest that attention is not dimension-wide. Instead, attention can stress different aspects of different stimuli. The implications of these findings for models of category learning are discussed.

## Introduction

Humans frequently utilize and acquire category knowledge at multiple levels of abstraction. For example, the same object can be classified as a vehicle, as a car, or as a 1978 Lincoln Continental. Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976) argue that objects are most easily classified at the intermediate category level which most effectively partitions the world into informative clusters. However, Tanaka and Taylor (1991) have found that different groups of people tend to prefer different levels of abstraction with experts preferring lower-level categories (i.e., narrower or finer grained categories) compared to novices who prefer higher-level categories (i.e., broader or more abstract categories).

One domain in which all adult humans are experts is the domain of face perception. Medin, Dewey, and Murphy (1983) found that people are faster to associate unique names to photographs of nine female faces than they are to categorize the photographs into two categories. The logical structure of the two categories is shown in Table 1. One possible explanation for the relative ease of identification learning is that the stimuli used in Medin et al. (1983) were rich and distinct, varying along many dimensions not listed in Table 1, such as the shape of the face, the type of nose, etc.. This *idiosyncratic* information makes each stimulus item more dis-

Table 1: The logical structure of the two categories used in the category learning condition. The four dimensions were hair color, smile type, hair length, and shirt color.

Category A	Category B
1112	1122
1212	2112
1211	2221
1121	2222
2111	

tinct. Experts may be more sensitive to idiosyncratic information than novices. In the absence of idiosyncratic information, common wisdom holds that identification learning should be harder than category learning. In other words, the ease of category acquisition interacts with the nature of the stimuli such that learning at lower levels of abstraction (with identification learning being the lowest level of abstraction) becomes easier relative to learning at higher levels of abstraction as stimuli become more distinct.

Results from the category learning literature support this conclusion. Shepard, Hovland, and Jenkins (1961) trained subjects on the six category learning problems listed in Table 2 and found that Type I was the easiest to master, followed by Type II, followed by Types III-V, followed by Type VI. The Type IV problem has a family resemblance structure that resembles the category structure used in Medin et al. (1983). In the Type IV problem, each category consists of an underlying prototype (111 for category “A” and 222 for category “B”) and any item that matches a prototype on two out of three dimensions is a member of the corresponding category. The more difficult to master Type VI problem, while not an identification learning problem, requires subjects to memorize all eight stimulus items because no regularities exist across any pair of dimensions. This data, along with Medin et al.’s (1983) data, suggest that preferred category level and stimulus distinctiveness interact with learning being facilitated more at lower category levels with distinctive stimuli.

One category learning model can capture this interaction. SUSTAIN (Supervised and Unsupervised STRatified Adaptive Incremental Network) has successfully fit Shepard et al.’s (1961) and Medin et al.’s (1983) data using the same set of

Table 2: The logical structure of the six classification problems tested in Shepard et al. (1961). The three binary valued dimensions correspond to size (small or large), shape (triangle or square), and color (light or dark)

Input	I	II	III	IV	V	VI
111	A	A	B	A	B	B
112	A	A	B	A	B	A
121	A	B	B	A	B	A
122	A	B	A	B	A	B
211	B	B	A	A	A	A
212	B	B	B	B	A	B
221	B	A	A	B	A	B
222	B	A	A	B	B	A

parameters (Love & Medin, 1998a; Love & Medin, 1998b).<sup>1</sup> SUSTAIN is a connectionist model that clusters similar items together. When items are clustered together inappropriately (i.e., similar items from incompatible categories are placed in the same cluster), SUSTAIN adds a new cluster in memory to encode the misclassified item. For example, if SUSTAIN is applied to stimulus items and classifies them as members of the category mammals or the category birds it will develop one or more clusters (i.e., prototypes) for the bird category and one or more clusters for the mammal category. When SUSTAIN classifies a bat for the first time, the bat item will strongly activate a bird cluster because bats are similar to birds (both bats and birds are small, have wings, and fly). After incorrectly classifying the bat as a bird, SUSTAIN will create a new cluster to encode the misclassified bat item. The next time SUSTAIN classifies a bat, this new cluster will compete with the other clusters and will be the most strongly activated cluster (i.e., it will be more similar to the current stimulus than any other cluster), leading SUSTAIN to correctly classify the novel bat as a mammal and not as a bird. The new cluster would then become a bat prototype (a subcategory of mammal). The primary difference between SUSTAIN and exemplar models is that SUSTAIN can cluster examples together in memory (like a prototype model). Unlike a prototype model, SUSTAIN can form multiple clusters (i.e., prototypes) per category.

When applied to Shepard et al.’s data, SUSTAIN recruits fewer clusters for the simpler problems. For the simplest problem, the Type I problem, SUSTAIN only recruits two clusters (one for each category). For the most difficult problem, the Type VI problem, SUSTAIN resorts to recruiting eight clusters (one for each item; each stimulus is memorized). When applied to Medin et al.’s (1983) data, SUSTAIN recruits more clusters (nine clusters; one for each stimulus item) in identification learning condition than in the category learning condition (the modal solution involves seven clusters). It is important to note that abstraction does not occur in the identification learning condition (i.e., each cluster

responds to only one item), but does occur in the category learning condition. What is interesting about these data fits is that in one case memorizing more items (acquiring more fine grained clusters or subcategories) led to more efficient learning, while in the other case it led to less efficient learning. The critical difference between these two data sets is the distinctiveness of the stimuli.

Two factors conspire to cause SUSTAIN’s performance to interact with the nature of the stimuli. As the stimuli become more distinctive, clusters that respond to multiple items (i.e., prototypes) are not as strongly activated. In other words, the benefit of abstraction is diminished with distinctive stimuli. This occurs because distinctive items sharing a cluster are not very similar to each other (i.e., within cluster similarity is low). Notice that the diminished benefit of abstraction negatively impacts performance in the Medin et al.’s (1983) category learning condition, but does not affect identification learning. In identification learning, each item forms its own cluster (within cluster similarity is maximal). When SUSTAIN is altered so that it does not form abstractions in either condition, but instead recruits a subcategory unit for each item, SUSTAIN fails to predict the interaction or the identification learning advantage, suggesting that abstraction is critical for capturing this effect. Without abstraction, the inferred category structures (i.e., the clusters recruited) are identical for both conditions.

The second factor that leads SUSTAIN to predict that distinctiveness and category level should interact is that the effects of cluster competition are attenuated with distinctive stimuli. As items become more distinctive, the clusters that are recruited tend to be further separated in representational space (i.e., the clusters match on fewer dimensions and mismatch on more dimensions). In other words, the clusters become more orthogonal to one another. The more distinctive the clusters are, the less they will tend to compete with one another. For instance, when a distinctive stimulus is presented to SUSTAIN, it will tend to strongly activate the appropriate cluster and will only weakly activate the competing clusters. Reduced cluster competition with distinctive stimuli favors both identification and category learning, but differentially benefits identification learning (or more broadly, learning at lower levels of abstraction) because there are generally more clusters present (i.e., potential competitors) in identification learning. Simulations support this analysis. When SUSTAIN is modified so that clusters do not compete, SUSTAIN reaches criterion more often and overall accuracy is much higher in the category learning condition.

## Experiment

SUSTAIN’s ability to fit Medin et al.’s studies on item and category learning is notable because other models cannot predict the advantage for identification learning or the interaction between learning task and stimulus distinctiveness. More importantly, SUSTAIN offers a framework for understanding the results. At the same time, it seems impor-

<sup>1</sup>The data actually fit was from Nosofsky et al.’s (1994) replication of Shepard et al. (1961).



tant to place SUSTAIN's account of these findings on firmer ground. To begin with, one should be cautious about accepting SUSTAIN's characterization of Medin et al.'s (1983) results. SUSTAIN's successful fit of Medin et al.'s (1983) studies depended on the choice of input representation. The idiosyncratic information in each photograph was represented by adding a number of input dimensions. Each item had a unique value on each added dimension. This manipulation had the effect of making all the items less similar to each other and making between and within category similarity virtually the same in the category learning condition. This input representation led SUSTAIN to predict that identification learning should precede category learning with distinctive stimuli.

The general intuition that guided my choice of input representation seems justified. Unlike artificial stimuli, the photographs do vary along a number of dimensions. Still, replicating the results from Medin et al. (1983) under more controlled circumstances with artificial stimuli would bolster my claims. Also, it is possible that there may be something "special" about faces (c.f., Farah, 1992).

The stimuli used in the Experiment were schematic cars that varied on a few dimensions. Like the Medin et al.'s (1983) study, subjects were assigned to either an identification or a category learning task. To manipulate the distinctiveness of the stimuli, some subjects viewed stimuli that were uniquely colored, while other subjects viewed stimuli that shared a common color. In essence, this experiment augments Medin et al.'s (1983) design with two non-distinctive stimuli conditions. The key prediction SUSTAIN makes is that category level and distinctiveness should interact such that identification learning performance should improve more than category learning performance as the stimuli become more distinctive. The choice of stimuli in this experiment directly tests SUSTAIN's characterization of the Medin et al. (1983) results.

Unfortunately, whether or not the interaction crosses over (i.e., whether or not identification learning with distinctive stimuli turns out to be faster than category learning with distinctive stimuli) cannot be predicted by SUSTAIN because the size of the effect depends on the saliency of the color dimension (the distinctive dimension). A crossover interaction can be accommodated by SUSTAIN, but is problematic for many other models.

The Experiment's design also allows for a secondary prediction to be tested related to cluster encoding and attention. When clusters only respond strongly to one item (i.e., "exception" clusters), does the cluster focus on the distinctive item information? Conversely, when a cluster encodes a number of items (i.e., an "abstraction" or "rule" cluster), does the cluster focus on what is general to the items and suppress distinctive item information? If the answer to both of these questions is "yes", the results would strongly suggest that attention is not dimension-wide (e.g., Nosofsky, 1986; Kruschke, 1992), but is cluster specific (c.f., Aha & Goldstone, 1992). In other words, different clusters can attend to different stimulus di-

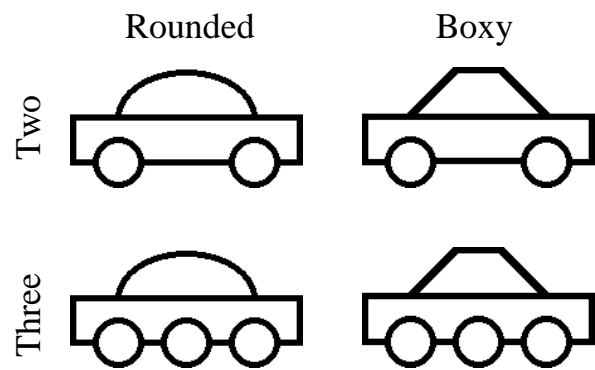


Figure 1: The stimuli varied in size (small or large), the number of wheels (two or three) and the shape of the cockpit (rounded or boxy shaped). Only one car size is shown in the figure. In the Experiment, eight different items were used.

mensions.

## Methods

**Subjects** Two hundred eighty-eight Northwestern University undergraduate students participated in the experiment for course credit or pay.

**Stimuli** Example stimuli are shown in Figure 1. In two of the four experimental conditions, each car was a different color. The eight colors were yellow, light blue, black, red, navy blue, pink, green, and grey. In the two other conditions, subjects viewed cars that were all the same color (either yellow, light blue, black, red, navy blue, pink, green, or grey).

**Design and Overview** The two variables (category level and distinctiveness) were crossed for a 2 X 2 between subjects factorial design. Subjects were randomly assigned to one of these four conditions.

The category level variable had two levels: identification learning and category learning. Subjects performing the identification learning task partitioned the eight items into eight categories (i.e., each stimulus formed its own category). Subjects performing the category learning task partitioned the eight items into two categories that had the same logical structure as Shepard et al.'s (1961) Type IV problem (see Table 2).

The distinctiveness variable had two levels: distinctive and non-distinctive. In the distinctive conditions, each item was a unique color. In the non-distinctive conditions, each item shared a common color. The learning phase ended when subjects completed consecutive error-free blocks of trials or after the completion of thirty-second block (each stimulus was presented in a random order once per block).

After completing the learning phase, sixty-six subjects completed two transfer blocks. Transfer trials were identical to learning trials with the exceptions that feedback was not provided and that each item was colored orange (a color not used during the learning phase).

## Procedure

Text was displayed in black on a white background. Trials began with a message displayed in the upper left corner of the screen alerting the subject to prepare for the next trial. After 1500 ms, this message was removed and the stimulus was displayed along with a message below it indicating that the subject should respond. Subjects were instructed to push the spacebar as soon as they decided on a response. After pressing the spacebar, subjects were prompted for their response. Subjects pressed either the “A” or “B” key in the category learning conditions. In the identification learning conditions, subjects used keys “A” through “H” to indicate their response. After responding, subjects received feedback. When subjects were correct, the message “Correct!” was displayed at the bottom of the screen. When subjects were incorrect, a message alerted the subject to the error and the correct response was displayed at the bottom of the screen. Following the subject’s response, the stimulus and all messages were displayed for 1500 ms. After another 1500 ms, the next trial began.

On transfer trials, text was displayed in black on a white background. Trials began with a message displayed in the upper left corner of the screen alerting the subject to prepare for the next trial. After 1500 ms, this message was removed and the stimulus was displayed along with a message below it indicating that the subject should respond. Subjects were instructed to push the spacebar as soon as they decided on a response. After pressing the spacebar, subjects were prompted for their response. Subjects pressed either the “A” or “B” key in the category learning conditions. In the identification learning conditions, subjects used keys “A” through “H” to indicate their response. Subjects did not receive feedback. Whether or not the subject responded correctly, the message “Thank You” was displayed. Following the subject’s response, the stimulus and all messages were displayed for 1500 ms. After 1500 ms, the next trial began. All possible factors were counterbalanced or randomly varied.

## Results

**Criterion** The mean of the number of blocks required by subjects in each condition is shown in Table 3. Table 3 also shows the mean of the reciprocal of the number of blocks required (this measure is less sensitive to outliers). A 2 X 2 (category level by distinctiveness) ANOVA was performed on both the untransformed scores and the transformed reciprocal scores. The transformed scores’ distributions were more similar to the normal distribution than were the distributions of the untransformed scores. Means are given only for the untransformed scores. Subjects required more blocks (14.5 vs. 12.9 blocks) in the category learning conditions than in the identification learning conditions (untransformed:  $F(1,218)=3.65$ ,  $MSe=140.5$ ,  $p=.06$ ; transformed:  $F(1,218)=4.40$ ,  $MSe=.00608$ ,  $p<.05$ ). Subjects required more blocks (17.0 vs. 10.5 blocks) in the non-distinctive conditions than in the distinctive conditions (untransformed:  $F(1,218)=60.95$ ,  $MSe=2348$ ,  $p<.001$ ; trans-

Table 3: The mean number of blocks required for each condition. In parenthesis, the mean of the reciprocals of the number of blocks required is shown.

	Identification Learning	Category Learning
Non-Distinctive	17.1 (.0661)	16.8 (.0753)
Distinctive	8.7 (.130)	12.2 (.100)

Table 4: The proportion correct for learning trials. In parenthesis, the proportion of subjects reaching the learning criterion is shown.

	Identification Learning	Category Learning
Non-Distinctive	.76 (.98)	.85 (.88)
Distinctive	.90 (1.00)	.90 (.96)

formed:  $F(1,218)=78.4$ ,  $MSe=.108$ ,  $p<.001$ ). The key prediction SUSTAIN makes is that category level and distinctiveness should interact such that identification learning performance should improve more than category learning performance as the stimuli become more distinctive. As predicted, distinctiveness and category level interacted such that distinctiveness sped up learning more (8.4 vs. 4.6 blocks) in the identification learning conditions than in the category learning conditions (untransformed:  $F(1,218)=5.08$ ,  $MSe=195.8$ ,  $p<.05$ ; transformed:  $F(1,218)=15.59$ ,  $MSe=.00216$ ,  $p<.001$ ).

A series of t-tests were conducted to probe individual cell differences. All differences were statistically significant at the .01 level except for the comparison (17.1 vs. 16.8 blocks) between the identification learning/non-distinctive condition and the category learning/non-distinctive condition (untransformed:  $t<1$ ; transformed:  $t(109)=1.54$ ,  $p=.13$ ).

Table 4 shows the proportion subjects that reached the learning criterion (the completion of consecutive error-free blocks) for each condition. Subjects reached the learning criterion more often in the identification learning conditions than in the category learning conditions ( $p<.05$  by a binomial test). Subjects also reached the learning criterion more often in the distinctive conditions than in the non-distinctive conditions ( $p<.05$  by a binomial test). Individual cell differences were not probed because of ceiling effects. Cell differences and interactions are explored in other analyses.

**Learning Trial Accuracy** In addition to analyzing the number of required learning blocks, the accuracy data were analyzed. We assume that after reaching the learning criterion subjects would respond correctly on the remaining trials if they maintained their motivational level. We scored the remaining post criterion blocks accordingly. The proportion correct for each condition is shown in Table 4.

A 2 X 2 (category level by distinctiveness) ANOVA was performed with the subjects’ accuracy rates serving as the dependent variable. Subjects were more accurate (.88 vs. .83) in the category learning conditions than in the identification learning conditions ( $F(1,218)=16.45$ ,  $MSe=.132$ ,  $p<.001$ ). Subjects were more accurate (.90 vs. .81) in the

Table 5: The proportion correct for transfer trials.

	Identification Learning	Category Learning
Non-Distinctive	.97	.93
Distinctive	.47	.81

distinctive conditions than in the non-distinctive conditions ( $F(1,218)=60.65$ ,  $MSe=.485$ ,  $p < .001$ ). The key prediction SUSTAIN makes is that category level and distinctiveness should interact such that identification learning performance should improve more than category learning performance as the stimuli become more distinctive. As predicted, distinctiveness and category level interacted such that distinctiveness led to a larger improvement in accuracy (.14 vs. .05) in the identification learning conditions than in the category learning conditions ( $F(1,218)=12.75$ ,  $MSe=.102$ ,  $p < .001$ ).

A series of t-tests were conducted to probe individual cell differences. All differences were statistically significant at the .01 level except for the comparison between the identification learning/distinctive condition and the category learning/distinctive condition ( $t < 1$ ).

**Transfer Trial Accuracy** Sixty-six subjects engaged in transfer trials after finishing the learning phase. The proportion correct for transfer trials for each condition is shown in Table 5. A 2 X 2 (category level by distinctiveness) ANOVA was performed with the subjects' accuracy rates serving as the dependent variable. Subjects were more accurate (.87 vs. .72) in the category learning conditions than in the identification conditions ( $F(1,62)=15.62$ ,  $MSe=.2838$ ,  $p < .001$ ). Subjects were more accurate (.95 vs. .64) in the distinctive conditions than in the non-distinctive conditions ( $F(1,62)=82.92$ ,  $MSe=1.506$ ,  $p < .001$ ). Distinctiveness and category level interacted such that identification learning led to a larger decrement in accuracy (.342 vs. -.0478) in the distinctive conditions than in the non-distinctive conditions ( $F(1,62)=34.34$ ,  $MSe=.6238$ ,  $p < .001$ ).

**Analyses of the Attentional Clustering Hypothesis** The tests presented here explore the possible existence of imperfect rule or abstraction clusters (focused on the non-distinctive dimensions) and exception clusters (focused on the distinctive dimension). Only the category learning data are relevant to the analyses presented here.

The Type IV problem used in the category learning conditions has essentially two types of items in each category: the prototypes (111 and 222) and all the other items. Within a category, each "other" item matches the prototype on two out of the three stimulus dimensions and mismatches on the third. Drawing this distinction between prototypes and other items proves useful in evaluating the attentional clustering hypothesis.

Although it is very difficult to analyze a subject's data and identify which items were exceptions and which items clustered with other items, SUSTAIN offers some direction. According to SUSTAIN, it is much more likely for an "other" item to be an exception than it is for a prototype item to be

Table 6: The proportion correct for Learning trials (to the left of the slash) and for transfer trials (to the right of the slash) by item type.

	Others	Prototypes
Non-Distinctive	.83 / .92	.92 / .96
Distinctive	.89 / .77	.94 / .93

an exception (prototypes are much more likely to cluster with other items). Drawing on this knowledge, 2 X 2 (item type by distinctiveness) ANOVAs can be performed to evaluate the hypothesis. If the hypothesis is correct, the advantage of the distinctive category learning condition over the non-distinctive category learning condition should largely be attributable to the greater ease in memorizing exceptions (i.e., non-prototype items) made possible by focusing on distinctive information. Thus, in the learning data, we should observe an interaction in which item type and distinctiveness interact such that non-prototype items benefit more from distinctiveness than prototype items. This prediction was confirmed (analysis below).

Conversely, for the transfer trials (where the distinctive information is obscured), the advantage of the non-distinctive condition should largely be attributable to better performance on the non-prototype items. In other words, the two factors should interact such that the distinctiveness hurts performance more for the non-prototype items than for the prototype items. This prediction was confirmed (analysis below). **Accuracy Data** A 2 X 2 (item type by distinctiveness) ANOVA was performed on the learning data from the category learning conditions with subjects' accuracy rates serving as the dependent variable (see Table 6). Subjects more accurately (.93 vs. .86) classified the prototypes than the other items ( $F(1,110)=139.18$ ,  $MSe=.264$ ,  $p < .001$ ). Subjects were more accurate (.92 vs. .87) in the distinctive conditions than in the non-distinctive conditions ( $F(1,110)=7.39$ ,  $MSe=.104$ ,  $p < .01$ ). One prediction of my proposed account of clustering and attentional focus is that item type and distinctiveness should interact such that learners in the distinctive conditions should see greater facilitation with non-prototype items than with prototype items in comparison to the non-distinctive conditions. As predicted, distinctiveness and item type interacted such that distinctiveness led to a larger improvement (.06 vs. .02) in the classification of non-prototype items in the distinctive conditions in comparison to the non-distinctive conditions ( $F(1,110)=7.33$ ,  $MSe=.0139$ ,  $p < .01$ ).

**Transfer Data** A 2 X 2 (item type by distinctiveness) ANOVA was performed on the transfer data from the category learning conditions with subjects' accuracy rates serving as the dependent variable (see Table 6). Subjects more accurately (.94 vs. .85) classified the prototypes than the other items ( $F(1,32)=15.78$ ,  $MSe=.155$ ,  $p < .001$ ). Subjects were more accurate (.94 vs. .85) in the non-distinctive con-

ditions than in the distinctive conditions and this result was marginally significant ( $F(1,32)=2.91$ ,  $MSE=.125$ ,  $p < .10$ ). One prediction of my proposed account of clustering and attentional focus is that item type and distinctiveness should interact such that transfer performance in the distinctive conditions should decline more for non-prototype items than for prototypes in comparison to item performance in the non-distinctive conditions. As predicted, distinctiveness and item type interacted such that distinctiveness led to a larger decline in performance (.15 vs. .03) in the classification of non-prototype items in the distinctive conditions in comparison to performance in the non-distinctive conditions ( $F(1,32)=5.49$ ,  $MSE=.054$ ,  $p < .05$ ).

## Discussion

SUSTAIN predicts that category level and distinctiveness should interact such that identification learning performance should improve more than category learning performance as the stimuli become more distinctive. SUSTAIN also predicts that identification learning can become easier than category learning as the stimuli become more distinctive. These predictions were borne out in an analysis of the number of learning blocks required by subjects and in another analysis of accuracy. This robust effects occurred by simply manipulating the color of the stimuli.

The experiment makes a bridge to the Medin et al.'s (1983) study in which subjects performed identification and category learning tasks with rich stimuli (photographs of human faces). SUSTAIN suggests that the identification advantage in the Medin et al. studies arises from the distinctiveness of the stimuli. The experiment tested this prediction directly and completed Medin et al.'s (1983) design with two non-distinctive conditions to complement the two distinctive conditions, allowing for natural comparisons to be made between the various conditions.

Another interesting aspect of the data was that subjects apparently emphasized (or attended) to different stimulus dimensions depending on which stimulus was being classified. In particular, when a stimulus was encoded by its own cluster, the distinctive aspects of the stimuli were emphasized. In contrast, when items shared a cluster, the cluster emphasized the non-distinctive aspects of the cluster members.<sup>2</sup> These results suggest that attention is not dimension-wide and is instead cluster specific. While SUSTAIN (in its current form) exhibits dimension-wide attention, it can be modified so that attention is cluster specific. The interpretability of SUSTAIN, which arises from its internal cluster representation of categories, suggested the analyses related to item type and attentional focus. Although the results of these analyses actually run counter to the specifics of the model, they speak favorably of SUSTAIN's general framework and its utility for directing empirical investigations.

<sup>2</sup>The idea that divergent examples are encoded in terms of how they differ from a default type is in the spirit of Kruschke's (1996) ADIT model of inverse base rate phenomena.

Overall, the results of the Experiment are troublesome for current models of category learning. Existing models have difficulty accounting for the interaction between category level and stimulus distinctiveness, the advantage of identification learning with distinctive stimuli, and attention not being dimension-wide. The first two findings are predicted by SUSTAIN and the third can be accommodated. SUSTAIN's ability to cluster together similar items from the same category (an ability that exemplar models lack) allows it to account for the results from the Experiment.

## Acknowledgments

I would like to thank Doug Medin for collaborating with me on this project. I would also like to thank Greg Ashby, John Kruschke, and James Tanaka for helpful comments on a related paper.

## References

- [1] AHA, D. W., AND GOLDSTONE, R. L. Concept learning and flexible weighting. In *n Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (1992), pp. 534–539.
- [2] FARAH, M. Is an object an object an object and object? Cognitive and neuropsychological investigations of domain-specificity in visual object recognition. *Current Directions in Psychological Science* 164-169 (1992), 1.
- [3] KRUSCHKE, J. K. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review* 99 (1992), 22–44.
- [4] KRUSCHKE, J. K. Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 22 (1996), 3–26.
- [5] LOVE, B. C., AND MEDIN, D. L. Modeling item and category learning. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (Mahwah, NJ, 1998), Lawrence Erlbaum Associates, pp. 639–644.
- [6] LOVE, B. C., AND MEDIN, D. L. SUSTAIN: A model of human category learning. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (Cambridge, MA, 1998), MIT Press, pp. 671–676.
- [7] MEDIN, D. L., DEWEY, G. I., AND MURPHY, T. D. Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 9 (1983), 607–625.
- [8] NOSOFSKY, R. M. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115 (1986), 39–57.
- [9] NOSOFSKY, R. M., GLUCK, M. A., PALMERI, T. J., MCKINLEY, S. C., AND GLAUTHIER, P. Comparing models of rule based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition* 22 (1994), 352–369.
- [10] ROSCH, E., MERVIS, C. B., GRAY, W. D., JOHNSON, D. M., AND BOYES-BRAEM, P. Basic objects in natural categories. *Cognitive Psychology* 8, 3 (1976), 382–439.
- [11] TANAKA, J. W., AND TAYLOR, M. Object categories and expertise: Is the basic level in the eye of the beholder. *Cognitive Psychology* 23, 2 (1991), 457–482.

Will Lowe

wlowe02@tufts.edu

Center for Cognitive Studies  
Tufts University; 11 Miner Hall  
Medford MA 02155 USA

Scott McDonald

scottm@cogsci.ed.ac.uk

Institute of Adaptive and Neural Computation  
University of Edinburgh; 2 Buccleuch Place  
Edinburgh EH8 9LW UK

## Abstract

McKoon and Ratcliff (1992) presented a theory of mediated priming where the priming effect is due to a direct but weak relatedness between prime and target. They also introduced a quantitative measure of word relatedness based on pointwise mutual information (Church and Hanks, 1990), and showed that stimuli chosen with the measure produced graded priming effects as predicted by their theory. Using stimuli from Balota and Lorch (1986), Livesay and Burgess (1998a,b) replicated the mediated priming effect in humans, but found that in HAL, a corpus-derived semantic space (Lund et al., 1995), mediated primes were in fact further from their targets than unrelated words. They concluded from this that mediated priming is not due to direct but weak relatedness. In this paper we present an alternative semantic space model based on earlier work (McDonald and Lowe, 1998). We show how this space allows a) a detailed replication of Ratcliff and McKoon's experimental results using their stimuli and b) a replication of Livesay and Burgess's human experimental results showing mediated priming. We discuss the implications for theories of mediated priming.

## Mediated Priming

Mediated priming is an important test for theories of semantic memory (Neely, 1991). According to spreading activation theory (e.g. Anderson, 1983), when a word is presented it activates its representation in a network structure in which semantically related words are directly connected; more generally, the semantic similarity of two words depends on the number of links that must be traversed to reach one to the other. The level of activation controls the amount of facilitation received by the corresponding word. Although ultimately every word can be reached from any location in the network, activation decays during memory access so only a few of the most related words are facilitated. Spreading activation theories predict that a prime word should facilitate pronunciation or lexical decision on a target word directly, for example when "tiger" facilitates "stripes". Spreading activation theory also predicts that "lion" will facilitate "stripes" when activation spreads from the representation of "lion" to that of "stripes", via the related concept of tiger (de Groot, 1983; Neely, 1991).

Small but reliable mediated priming effects have been demonstrated for pronunciation tasks though they are less reliable for lexical decision (Balota and Lorch, 1986). Spreading activation theory explains the size of the priming effect by arguing that "lion" and "stripes" are only indirectly related in semantic memory so that activation has decayed significantly by the time activation from "lion" reaches "stripes".

Theories that do not assume the existence of activation or a network structure in semantic memory, e.g. compound cue theory (Ratcliff and McKoon, 1988; McKoon and Ratcliff, 1998), cannot take advantage of either of the priming explanations above. In compound cue theory, direct priming is explained roughly as follows: the prime and target are joined in a compound cue that is compared to representations in long-term memory. The comparison process generates a 'familiarity' value which controls the size of the priming effect. The essential feature of this explanation is that, unlike spreading activation theory, there is no mention of the intermediate representation "tiger" when explaining how "lion" facilitates "stripes". But is less clear how compound cue theory should explain mediated priming.

In response to this difficulty, McKoon and Ratcliff (1992) have argued that the mediated priming effects are not due to activation spreading through an intervening representation, but are the result of direct but weak relatedness between the prime and target words. To address the issue of priming effect magnitude they provided a quantitative method for generating prime target pairs with various degrees of relatedness. The method is based on pointwise mutual information (Church and Hanks, 1990) computed over a corpus. McKoon and Ratcliff's (1992) Experiment 3 showed that their method produced stimuli that reliably generated a range of priming effect sizes, and that the effect sizes could be controlled. They then argued that mediated priming is simply a special case of graded priming.

Livesay and Burgess (1998a,b) replicated the mediated priming effect in human subjects using a pronunciation task, but had less success with lexical decision (the same situation that was reported in Balota and Lorch's original paper). In an attempt to understand the nature of the priming mechanism they found that mediated primes from the Balota and Lorch stimuli could be divided heuristically into contextually appropriate and contextually inappropriate word pairs. Subsequent analysis revealed that only contextually appropriate pairs were responsible for generated a priming effect.

They then compared distances between each type of prime (direct or mediated) and their targets in HAL, a semantic space model (Lund et al., 1995). Burgess and colleagues have argued that distances in HAL reflect semantic relatedness; shorter distances reflect greater semantic relatedness (Burgess et al., 1998). Directly related primes were on average closer to their targets than the corresponding unrelated primes, so HAL successfully replicated the direct priming effect. However, both contextually appropriate and contextually

ally inappropriate mediated primes were *further* from their targets than unrelated controls. Thus distances in HAL predict that the mediated primes should slow responses to their targets, relative to an unrelated word baseline. Subsequent analysis showed that even for contextually consistent primes, greater distance correlated 0.6 with larger priming effects.

Livesay and Burgess concluded that mediated priming could not be due to direct but weak relatedness between mediated primes and their targets on the grounds that HAL predicted the wrong effect. They then explored the possibility, suggested in McKoon and Ratcliff’s paper, that mediated priming is determined by raw co-occurrence frequencies between prime words and their targets, but found no significant correlations.

Below we present replications of two priming experiments using a semantic space model. In Experiment 1 we replicate human performance on the stimuli generated by McKoon and Ratcliff using pointwise mutual information. We will refer to these stimuli as the mutual information stimuli. These results demonstrate that McKoon and Ratcliff’s direct theory of mediated priming is consistent with explanations of priming based on semantic space. In Experiment 2 we tackle mediated priming directly by replicating the results of Livesay and Burgess’s mediated priming experiment. From these two experiments we argue that our semantic space constitutes a model of mediated priming that is ‘direct’ in the way that McKoon and Ratcliff suggested.

## Experiment 1

### Materials

In this experiment we use materials from McKoon and Ratcliff’s Experiment 3. Each target (e.g “grass”) has a prime taken from association norms (“green”), a high-t prime (“acres”) and a low-t prime (“plane”). High and low-t primes were chosen by first calculating a measure of lexical association based on the T-statistic between each target word and a large number of candidate primes (Church and Hanks, 1990, see Appendix A for details). McKoon and Ratcliff divided the candidate primes for each target into those with high values of the T-statistic (high-t primes) and low values (low-t primes). Unrelated primes were related primes from another target.

### Methods

We constructed a semantic space from 100 million words of the British National Corpus, a balanced corpus of British English (Burnage and Dunlop, 1992). Word vectors were generated by passing a moving window through the corpus and collecting co-occurrence frequencies for 536 of the most reliable context words within a 10 word window either side of each stimulus item. Appendix B describes the method of choosing reliable context words. We used positive log odds-ratios to measure the amount of lexical association between each context word and each of the experimental stimuli.

A brief justification of the positive log odds-ratio as a measure of lexical association is appropriate at this point: Table 1 describes the true co-occurrence probabilities for a stimulus word  $t$  and context word  $c$ .  $p(c, \neg t)$  is the probability of seeing  $c$  with a word other than  $t$ . The odds of seeing  $t$  rather than some other word when  $c$  is present are  $p(c, t)/p(c, \neg t)$  and the odds of seeing  $t$  in the absence of  $c$  are  $p(\neg c, t)/p(\neg c, \neg t)$ , so if the presence of  $c$  increases the probability of seeing  $t$  then

Table 1: The true probabilities of seeing combinations of words  $t$  and  $c$  in text.  $p(c, t)$  is the probability of seeing words  $c$  and  $t$  together in a window.  $p(c, \neg t)$  is the probability of seeing  $c$  together with a word that it *not*  $t$ .

	Target	Non-target
Context	$p(c, t)$	$p(c, \neg t)$
Non-context	$p(\neg c, t)$	$p(\neg c, \neg t)$

the odds ratio

$$\theta(c, t) = \frac{p(c, t)/p(c, \neg t)}{p(\neg c, t)/p(\neg c, \neg t)} = \frac{p(c, t) p(\neg c, \neg t)}{p(c, \neg t) p(\neg c, t)}$$

is greater than 1. When  $\theta > 1$   $c$  and  $t$  are said to be positively associated. In contrast, if the presence of  $c$  makes it *less* likely that  $t$  will occur then  $\theta < 1$  and  $c$  and  $t$  are negatively associated. Finally, when the presence of  $c$  makes no difference to the probability of seeing  $t$  then  $\theta = 1$  and we can conclude that  $c$  and  $t$  are distributionally independent.

An important advantage of the odds ratio for measuring lexical association is that takes into account differing marginal word frequencies. For example, consider two target words  $t_1$  and  $t_2$  that have baseline occurrence probabilities  $p(t_1)$  and  $p(t_2)$ . For simplicity we assume that co-occurrences are counted in a window extending exactly one word to one side of stimulus. When neither word is related to a context word  $c$  then all three words will distributionally independent. Under distributional independence the expected values of co-occurrence counts  $f(c, t_1)$  and  $f(c, t_2)$  depend only on their occurrence probabilities:

$$\begin{aligned} E[f(c, t_1)] &= p(c) p(t_1) N \\ E[f(c, t_2)] &= p(c) p(t_2) N \end{aligned}$$

where  $N$  is the number of words in the corpus<sup>1</sup>. If  $p(t_1)$  is much larger than  $p(t_2)$  then the expected co-occurrence counts may differ substantially, despite the fact that  $c$  has no relation to  $t_1$  or  $t_2$ . In other words if raw co-occurrence counts are used to measure lexical association then a more frequent target word will be judged more strongly associated with  $c$  than a less frequent target word, whether or not they are actually related. Also, the fact that vector elements for two target words with different frequencies will be tend to have different magnitudes will bias the Euclidean distance measure to treat target words from different frequency bands as further away from each other than those in the same band. This occurs because the measure depends on squared differences between vector elements.

The odds ratio is well-known to be a measure of association that takes chance co-occurrence into account (Agresti, 1990). When  $t_1$  and  $c$  are distributionally independent then  $p(t_1, c) =$

<sup>1</sup>Strictly speaking  $N$  is the number of bigrams in the corpus, which is one less than the number of words.

$p(t_1)p(c)$ . The odds ratio is

$$\theta(c, t_1) = \frac{p(c)p(t_1)p(-c)p(-t_1)}{p(c)p(-t_1)p(-c)p(t_1)} = 1,$$

and it is clear that the value of  $\theta(c, t_1)$  does not depend on target and context word frequencies.

$\theta(c, t_1)$  is estimated from a corpus by setting the elements of Table 1 to their Maximum Likelihood values. The odds ratio estimate can then be computed using only occurrence and co-occurrence frequencies (see e.g. Agresti, 1990)

$$\hat{\theta}(c, t) = \frac{f(c, t) f(-c, -t)}{f(c, -t) f(-c, t)}.$$

We log the odds ratio to make the measure symmetric around 0 (denoting distributionally independent words) and set all negative odds-ratios to zero. This reflects our belief that information about the whether a word occurs with another *more* often than chance is psychologically salient, whereas the knowledge that a word tends *not* to occur with some other word (one of, say, 60,000 others in the lexicon) is not psychologically salient and need not be represented in the model. Empirical studies show that neither logging nor truncation of the basic odds-ratio measure make much difference to the results presented below. The most important step seems to be taking into account chance when using co-occurrence to quantify lexical association. The g-score (Dunning, 1993) is another useful measure for this purpose (McDonald and Lowe, 1998).

We created vectors for each of the experimental stimuli by calculating lexical association values between it and each context word. Unrelated primes were primes from the previous target word<sup>2</sup>. We use the cosine of the angle between word vectors as a similarity measure corresponding to semantic relatedness (McDonald and Lowe, 1998).

When modeling priming experiments, the cosine between a prime and its target should be inversely proportional to the corresponding reaction time. The size of a priming effect is calculated by subtracting the cosine between the unrelated prime and target from the cosine between the related prime and target. Cosines are entered directly into analyses of variance.

## Results

McKoon and Ratcliff’s subjects responded fastest to target words preceded by an associated prime, next fastest to a high-t prime, slower to a low-t prime and slowest of all to an unrelated prime (see Table 2, line 1.) Priming effects were reliable in all except the low-t condition.

The cosine similarity measure shows similar results (see Table 2, line 2). The following analyses are for items only since there are no subjects. The prime conditions were significantly different,  $F(3,156)=33.32$ ,  $p<.001$  so we performed pairwise analyses of variance to examine the differences more closely, correcting for multiple comparisons according to the Bonferroni method. There was a reliable associative priming effect: associated pairs were significantly more related

<sup>2</sup>Since the stimuli have no inherent ordering, this will not produce any spurious effects. Other methods of choosing primes have been tested and give equivalent results.

than non-associated pairs (0.412 vs. 0.078),  $F(1,78)=80.645$   $p<.001$  and high-t pairs were significantly more related than unrelated pairs (0.216 vs. 0.078),  $F(1,78)=19.727$   $p<.001$ . The mean value for low-t pairs was higher than the unrelated baseline (0.139 vs. 0.078), but this was not significant  $F(1,78)=5.268$   $p=.024$ .

Table 2: Mean reaction times in msec. (line 1) and cosines on (line 2) for the mutual information stimuli (from McKoon and Ratcliff, 1992)

	Related	High-t	Low-t	Unrelated
M&R	500	528	532	549
Space	0.412	0.216	0.139	0.078

## Discussion

Experiment 1 shows a close fit to human reaction time data. The experiment also demonstrates that semantic space models are capable of representing the kind of weak but direct relatedness that McKoon and Ratcliff argue underlies mediated priming. If we can also account for mediated priming data, we will not only have uncovered additional evidence that direct but weak relatedness is sufficient to explain mediated priming, but also have found a ‘direct’ alternative explanation for the apparent mediation process. We address mediated priming in Experiment 2.

## Experiment 2

### Materials

Materials for Experiment 2 are taken from Balota and Lorch’s (1986) paper. Each target (e.g. “stripes”) has a directly related prime (“tiger”) and a mediated prime (“lion”). One target had to be discarded because it had a prime with very low frequency in the corpus. A randomly chosen prime target combination was discarded from each of the other two prime conditions to maintain balance.

### Method

The semantic space was the same as in Experiment 1.

### Results

In the pronunciation task both Balota and Lorch and Livesay and Burgess’s subjects showed direct and mediated priming (see Table 3, lines 1 and 2). The semantic space measure for related, mediated and unrelated pairs is shown in Table 3, line 3. The prime conditions were significantly different  $F(2,132)=12.065$   $p<.001$  and we performed pairwise analyses of variance to examine the differences in more detail. There was a reliable direct priming effect (0.212 vs. 0.085),  $F(1,88)=24.105$   $p<.001$  and also a reliable mediated priming effect of smaller magnitude (cosines 0.164 vs. 0.084),  $F(1,88)=13.107$   $p<.001$ .

## Discussion

The results of Experiment 2 show that it is possible to model mediated priming using a semantic space. The experiment also demonstrates the plausibility of McKoon and Ratcliff’s theory that direct but weak relatedness underlies mediated priming phenomena. There is no mediation mechanism in

Table 3: Mean reaction times in for the pronunciation experiments of Balota and Lorch (B&L, line 1) and Livesay and Burgess (L&B, line 2) in msec. Similarity measures for the same materials are on line 3.

	Related	Mediated	Unrelated
B&L Pron.	549	558	575
L&B Pron.	576	588	604
Space	0.212	0.164	0.084

the space, so the most parsimonious explanation of mediated priming is that it is due to direct relatedness.

On the other hand, Livesay and Burgess's distinction between contextually consistent and contextually inconsistent prime target pairs suggests an alternative view. Perhaps only some of the mediated priming stimuli are causing priming, and the rest are unnecessary.

Unfortunately the distinction between contextually consistent and inconsistent pairs appears to resist characterization in quantitative terms, e.g. in terms of distance in HAL. To investigate the possibility that a subset of primes were carrying the mediated priming effect we examined the distribution of differences between a) cosines between unrelated primes and their targets and b) mediated primes and their targets. The larger these differences are, the larger the mediated priming effect. If only a subset of materials carry the priming effect then we might expect that some targets have larger differences than the rest. However, we found that differences clustered symmetrically around the mean effect size. Ideally we would correlate priming effect size in milliseconds to the cosine measure to identify a subset of relevant primes; this is further work.

In an attempt to understand why HAL does not produce mediated priming, we attempted to replicate its behaviour on the mediated priming stimuli by changing the parameters of our semantic space. First, we used co-occurrence counts for the 536 reliable context words to create vectors for the Balota and Lorch materials and computed Euclidean distances between each prime and target combination. There were no significant differences between conditions,  $F(2,132)=0.043$   $p=.958$ . We then performed the same analysis with vectors normalized to length 1 to offset the effects of large co-occurrence counts. The conditions were still not reliably different  $F(2,132)=1.257$ ,  $p=.288$ . However, in this case the model hinted at a direct priming effect and a smaller mediated effect. Finally we constructed vectors from 500 higher frequency context words<sup>3</sup>, in case our choice of context words had adversely affected the measure. We used normalized vectors because they had previously given a slightly better match to the priming magnitudes. Again there was no significant difference between the conditions  $F(2,132)=0.493$   $p=0.612$ , but the model suggested a larger direct than mediated priming effect.

In conclusion, we were not able to replicate HAL's behaviour by changing the parameters of our model, so it is

<sup>3</sup>The context words had rank frequencies from 200 to 700. Occurrence frequencies ranged between 61926 to 220 occurrences per million.

not easy to explain why the cosines in the space replicate human mediated priming effects while distances in HAL do not. It is possible that relevant differences between the space and HAL depend on HAL's method of choosing context words, or its window weighting function for collecting co-occurrence counts. Comparisons between the space and HAL are the subject of ongoing work.

## Conclusion

In Experiments 1 and 2 we have presented detailed replications of human performance on graded and mediated priming stimuli using a semantic space. Since there is no mediation mechanism in the space we have argued that direct but weak relatedness, as reflected by the cosine measure in our space, is sufficient to yield a mediated semantic priming effect. This result supports McKoon and Ratcliff's contention that weak relatedness, rather than spreading activation, underlies mediated priming effects.

The results presented here stand in marked contrast to HAL's failure to generate mediated priming effects. However, we were not able to replicate HAL's behaviour in our model, so it is presently unclear why the HAL model does not work for this data.

We conclude by noting that graded and mediated priming can now be added to the list of psycholinguistic phenomena which may be accounted for by semantic space models.

## Acknowledgments

WL is grateful to the Medical Research Council for funding, and to Daniel Dennett and the Center for Cognitive Studies at Tufts for providing a supportive and stimulating research environment. SM acknowledges the support of NSERC Canada and the ORS Awards Scheme.

## References

- Agresti, A. (1990). *Categorical Data Analysis*. John Wiley and Sons.
- Anderson, J. R. (1983). *The Architecture of Cognition*. Harvard University Press.
- Balota, D. A. and Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning Memory and Cognition*, (12):336–345.
- Burgess, C., Livesay, K., and Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, (25):211–257.
- Burnage, G. and Dunlop, D. (1992). Encoding the British National Corpus. In *Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora*.
- Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, (16):22–29.



- de Groot, A. M. B. (1983). The range of automatic spreading activation in word priming. *Journal of Verbal Learning and Verbal Behavior*, pages 417–436.
- Dunning, T. (1993). Accurate methods for the statistics for surprise and coincidence. *Computational Linguistics*, (19):61–74.
- Finch, S. (1993). *Finding Structure in Language*. PhD thesis, Centre for Cognitive Science, University of Edinburgh.
- Livesay, K. and Burgess, C. (1998a). Mediated priming does not rely on weak semantic relatedness or local co-occurrence. In *Proceedings of the Cognitive Science Society*, pages 609–614.
- Livesay, K. and Burgess, C. (1998b). Mediated priming in high-dimensional meaning space: What is mediated in mediated priming? In *Proceedings of the Cognitive Science Society*, pages 436–441.
- Lund, K., Burgess, C., and Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665. Mahwah, NJ: Lawrence Erlbaum Associates.
- Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- McDonald, S. and Lowe, W. (1998). Modelling functional priming and the associative boost. In Gernsbacher, M. A. and Derry, S. D., editors, *Proceedings of the 20th Annual Meeting of the Cognitive Science Society*, pages 675–680, New Jersey. Lawrence Erlbaum Associates.
- McKoon, G. and Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory and Cognition*, (18):1155–1172.
- McKoon, G. and Ratcliff, R. (1998). Memory-based language processing: Psycholinguistic research in the 1990s. *Annual Review of Psychology*, (49):25–42.
- Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In Besner, D. and Humphreys, G. W., editors, *Basic processes in reading: Visual word recognition*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Ratcliff, R. and McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, (95):385–408.

## Appendix A

The pointwise mutual information or *association ratio* between a target word and candidate prime is

$$AR = \log_2 \frac{p(\text{prime and target})}{p(\text{prime})p(\text{target})}.$$

The numerator is estimated by normalizing the number of co-occurrences between prime and target words over the corpus. The denominator is estimated from the occurrence frequencies of the prime and target words separately. When prime and target words are distributionally independent *AR* should, like the log odds-ratio, take the value zero. When the prime word is occurs with the target more than would be expected by chance *AR* is positive with greater magnitude for greater levels of association. The T-statistic may be used to determine whether the ratio is significantly different than 0, although Church and Hanks (1990) use the value of the statistic itself as a lexical association measure. The *AR* measure is called pointwise mutual information in analogy to mutual information, an information theoretic measure which is the expectation of *AR* with respect to the distribution  $p(\text{prime and target})$ . Manning and Schütze 1999 discuss uses and shortcomings of pointwise mutual information as an association measure.

## Appendix B

We assume that the ease that two words can be substituted for one another in text reflects their semantic similarity. Substitutability in context, defined over word pairs or *targets*, is the underlying continuous quantity that a semantic space model needs to capture (Finch, 1993). Measuring substitutability in context entails holding linguistic context constant and swapping in targets. This is equivalent to holding targets constant and examining possible surrounding linguistic contexts because targets that are easily substitutable are those that occur in similar contexts.

Any large balanced corpus, such as the BNC, realizes a subset of the possible linguistic contexts that can surround a target. Given sufficient target instances the subset will be representative because the number of times a context surrounds a target is proportional to how meaningful the resulting sentence is. We represent contexts using finite set of *context words*. The linguistic contexts that surround a target are represented by the number of times each context word occurs within a 10 word window surrounding the target. These co-occurrence counts and the marginal frequencies of each context word and the target are used to create vectors of positive log odds ratios. To represent linguistic context adequately context words should be *reliable*.

To quantify reliability we treat context words like human raters and use standard ANOVA methods to assess their reliability: First, we choose several thousand candidate context words from the high frequency portion of the BNC (excluding stop words). Second, we pick randomly another set of words called meta-context words, and compute log odds ratios as described above for each context and meta-context word combination over  $k$  disjoint sections of the corpus. The resulting  $k$  matrices can be seen either as sets of column vectors describing the positions of the meta-context words in a space defined by the candidate context words, or as a set of row vectors describing the positions of the candidate context words in a space given by the meta-context words. The meta-context words are so-called because they are context words for the candidate context words. Each candidate context word is then associated with  $k$  vectors. We consider the vectors to be the results of  $k$  rating tasks and use a within subjects ANOVA to

test whether each context word generates significant variation in vector elements between the  $k$  tests. Context words that are reliable have  $k$  vectors with similar values so their rating do not vary significantly across corpus sections. Context words for which we cannot reject the null hypothesis of no variation between corpus sections are retained.

In these experiments we chose  $k=4$  sections from the BNC, each containing 10M words, and used the rather conservative critical significance level 0.1. The procedure generated 536 context words.

# Zen in the Art of Language Acquisition: Statistical Learning and the Less is More Hypothesis

**David Ludden** ([david-ludden@uiowa.edu](mailto:david-ludden@uiowa.edu))  
Department of Psychology, University of Iowa  
Iowa City, IA 52242 USA

**Prahlad Gupta** ([prahlad-gupta@uiowa.edu](mailto:prahlad-gupta@uiowa.edu))  
Department of Psychology, University of Iowa  
Iowa City, IA 52242 USA

## Abstract

It seems an obvious truth that children are better language learners than adults. Children seem able to master a second language with ease, while adults are rarely successful at second language acquisition. Newport's (1990) Less is More hypothesis represented an attempt to explain these observations by invoking general cognitive mechanisms. This hypothesis takes as its starting point the observation that children exhibit reduced working memory capacity relative to adults and suggests that this reduction serves as a filter to aid children in deducing the structure of the language they are learning. We present two experiments testing a specific prediction that follows from the Less is More hypothesis, namely that adults will perform better on language learning tasks if their available working memory capacity is reduced. The experiments examined the learning of word boundaries and syntactic agreement, each with and without a concurrent cognitive load. The results of these experiments were contrary to the Less is More prediction, suggesting that other explanations must be found for the observed superior language learning performance of children over adults.

## Introduction

In *Zen in the Art of Archery* (Herrigel, 1953), the German philosopher Eugen Herrigel recounts his endeavor to learn Japanese archery from a great Zen master. During his many years of training, Herrigel struggles with his need to control the bow, to think about the target. The more he tries to control the bow, the Master tells him, the less control he will have over it; the more he thinks about the target, the less accurate his shots will be. Rather, the Master tells Herrigel to strive for less control, less conscious thought of the act of releasing the arrow from the bow, for only then will he gain mastery over the bow. The idea that Less is More is a dominant theme of Zen Buddhism. However, this notion is not limited to the realm of Oriental philosophy; it is one that pops up again in Western developmental psychology, in particular, in the area of language acquisition.

It is generally acknowledged that children's cognitive abilities, working memory in particular, are considerably constrained relative to those of adults (Gathercole, Willis,

Baddeley & Emslie, 1994). At the same time, it has also been widely observed that children are much more successful at learning language than adults. After all, virtually all children learn a first language, while few adults ever master a second. Furthermore, children seem to learn language with such effortlessness, as opposed to the great expense of effort necessary for an adult to acquire even marginal proficiency in a second language. To Newport (1990), this seemed a perfect example of a Less is More situation: children have less cognitive ability yet learn language more easily, while adults have more cognitive ability yet learn language less easily. This observation led to the formulation of the Less is More hypothesis (Newport, 1990), which not only maintained that the restricted working memory of children was an advantage to language acquisition, but also attempted to explain the causal nature of that relationship.

The idea for the Less is More hypothesis came out of studies on critical periods in first and second language acquisition performed by Newport and her colleagues. For example, Johnson & Newport (1989) conducted a study of 46 native Chinese and Korean speakers living in the United States who had learned English as a second language. The participants were divided into two groups, those who had arrived in the US before age 15 (dubbed Early Arrivals), and those who had arrived in the US after age 17 (dubbed Late Arrivals). All participants had spent at least the last three years prior to the experiment in the US. Johnson & Newport (1989) found an inverse linear relationship between age of arrival in the US and ultimate performance in English. In that study, only those participants who had arrived in the US by the age of seven achieved native speaker proficiency as measured on a grammaticality judgment task. Beyond age seven, performance on this task decreased as age or arrival increased. Johnson & Newport (1989) also examined attitudinal variables, but found statistical significance for the age of arrival variable over and above and other variables they looked at.

In a subsequent study, Newport (1990) examined critical periods in first language acquisition, in particular in American Sign Language (ASL). She studied three groups of congenitally or pre-lingually deaf adults who used ASL as their predominant language and had limited skills in English. The first group, dubbed Native Learners, had been exposed to

ASL since birth or shortly thereafter. The second group, dubbed Early Learners, had been first exposed to ASL between the ages of 4 and 6. The last group, dubbed Late Learners, had been exposed to ASL only after age 12. All participants were tested on their knowledge of ASL sentence structure and morphology. While all three groups performed at ceiling on the sentence structure test, there was a significant decline in performance in morphology across the three groups from Native to Early to Late Learners.

Newport (1990) took these findings as clear evidence for a critical period in language acquisition, and in an attempt to explain the mechanism responsible for this critical period she posited the Less is More hypothesis. The hypothesis takes as its starting point the notion that the working memory capacity of children is limited compared to that of adults. The hypothesis then proposes that this limitation is actually advantageous. According to Newport (1990), language acquisition requires a componential analysis. Adults take in too much of the language input at one time because of their expanded working memory capacity relative to children. This wider perceptual window in adults leads to a combinatorial explosion of possible analyses for the language input, hence the likelihood of hitting upon the right analysis is small. Children, with their limited working memories, are constrained by the size of the input they do take in to a more limited number of possible analyses. Hence, the likelihood of hitting upon the right analysis is greatly increased. According to Newport (1990), it is the limitations of the child's ability to process information that provides the basis for successful language acquisition.

Elman (1993) tested the Less is More hypothesis with a connectionist model of syntactic agreement acquisition. He trained a simple recurrent network (Elman 1990) on a corpus of sentences based on a simplified English grammar. In this grammar, subjects and verbs agreed in number, verbs differed in argument expectations, and sentences could contain multiple embeddings. The corpus contained sentences such as *cats chase dogs* and *dogs see boys who cats who mary feeds chase*. The context units of this simple recurrent network represented the working memory of the network, and the capacity of this working memory was a parameter that could be varied. When the network was trained on the entire corpus with working memory at full capacity, the network failed to learn.

Elman (1993) then tried incrementing the capacity of the working memory of the network. Working memory capacity was manipulated by an automatic flushing of the context units after every three or four words. As training progressed, the interval between flushings was gradually increased. The result was that the network was then able to learn how to process the input. Elman (1993) interpreted this finding as consistent with the Less is More hypothesis.

Relevant empirical data came from a study by Santelmann & Jusczyk (1998), who used a headturning paradigm with 15- and 18-month-old infants to test their sensitivity to morphosyntactic dependencies in English. The experimental condition consisted of well-formed English sentences with the structure *...is...<verb>ing*, while the control condition consisted of ill-formed sentences with the structure *...can...<verb>ing*, such as *Everybody is baking* vs. *\*everybody can baking*. Santelmann & Jusczyk (1998) also varied the distance in syllables between auxiliary verb (*is* or *can*) and main verb by the insertion of adverbs, as in *Everybody is often baking*. They found that at distances of 1-3 syllables, 18-month-old infants preferred well-formed over ill-formed sentences. However, at distances over 3 syllables, the 18-month-olds showed no preference for either form, nor did the 15-month-olds at any distance. Santelmann & Jusczyk (1998) concluded that their findings were "consistent with the hypothesis that 18-month-olds are working with a limited processing window, and that they are only picking up relevant dependencies that fall within this window." Although the authors found no evidence to determine whether these limitations in processing space facilitated or hampered language acquisition, the Santelmann & Jusczyk (1998) study nevertheless does lend support to a basic premise of the Less is More hypothesis, namely that infants are processing the language input in shorter chunks than adults are, justifying the approach Elman (1993) took in modeling the syntax-acquisition process.

Work on statistical learning by Saffran and her colleagues has also been relevant to the Less is More hypothesis. Saffran, Newport & Aslin (1996a) asked adult participants to listen to a nonsense language that contained words but no meanings or grammar. The task was to try to figure out where the word boundaries were. At the end of 21 minutes of exposure, the participants were asked to choose which of two items sounded more like a word from that language. The participants performed significantly above chance, with a mean score of 76% (chance was 50%). This type of exposure condition was referred to as the explicit learning condition in this and later Saffran studies.

Saffran et al. (1997) tested the learning of word boundaries in an incidental learning condition. In this condition, participants were asked to draw a picture while the stimulus played in the background. Subjects were told nothing about the stimulus. After 21 minutes of exposure, the participants were administered the same test as in the explicit condition. Saffran et al. (1997) tested two groups, adults (college students) and children (6-7 years old). Mean percent correct identification scores for each group were significantly above chance (50%) at around 59%, with no significant difference between children and adults. Because of the low scores after one exposure period, the experiment was redone with two exposure periods on consecutive days. In this second experiment, adults averaged 73% and children 68%, with the difference between adults and children being nonsignificant. Saffran et al. (1997) concluded that passive exposure was

sufficient at least for some aspects of the language acquisition process.

In her dissertation, Saffran (1997) extended her research in statistical learning to the acquisition of syntax, in particular, hierarchical phrase structure. The stimulus set in each experiment consisted of a sample of sentences from an artificial language, with the only cues to syntactic structure being statistical. In an explicit learning condition, the participants were exposed to the stimulus for 30 minutes a day for two days, and tested on their knowledge of the phrase structure at the end of each listening period. Mean adult performance in this explicit learning task was 68%<sup>1</sup>. (No children were run in this condition.) In an incidental learning condition in which participants listened to the stimulus while drawing a picture, both adults and children (aged 6-9) showed performance significantly above chance after the first exposure period, with no significant improvement after the second session.<sup>2</sup> Children's scores (57%) were significantly worse than those of the adults (67%). There was no significant difference between adults in the explicit and incidental conditions.

The results of these various experiments by Saffran and her colleagues seem inconsistent with the Less is More hypothesis. Specifically, the Less is More hypothesis predicts that children will perform better than adults in language learning tasks, and furthermore that adults will perform better in an incidental learning task than an explicit one. But these predictions are belied by the data. Not only was there no significant difference in performance between children and adults in the incidental word boundary learning task (Saffran et al. 1997), children in fact fared worse than adults in the implicit syntax learning task (Saffran, 1997). Furthermore, there was no significant difference between explicit and incidental conditions in adult performance on the syntax learning task (Saffran 1997).

The various results described above paint an inconsistent picture of the impact of working memory resources in language learning. The studies by Elman (1993) and Santelmann & Jusczyk (1998) appear to support the Less is More hypothesis, while the results of the studies by Saffran et al. (1996) and Saffran et al. (1997) are inconsistent with that hypothesis.

The experiments described below were aimed at examining the following question: Is adult performance on a language learning task superior when working memory resources are reduced, as the Less is More hypothesis would predict? Although the results of Saffran

et al. (1997) are inconsistent in this regard, they are difficult to interpret because they were obtained under different experimental conditions. The present experiments attempt to address this question systematically. Experiment I addresses this question in the domain of word boundary learning while experiment II addresses this question in the domain of syntax learning.

## Experiment I: Word Boundaries

Because the exposure periods in Saffran et al.'s explicit (1996a) and incidental (1997) word boundary learning tasks were not equivalent, a direct comparison cannot be made. Experiment I represents an attempt to replicate these two experiments under identical exposure conditions. To insure this, the difference between these two tasks was reduced to the presence or absence of a concurrent cognitive load (drawing a picture) during the exposure to the stimulus. For this reason, in this and the following experiment, Saffran et al.'s (1996a) explicit condition is referred to by the more theory-neutral term No Load, while Saffran et al.'s (1997) incidental condition is referred to as the Load condition. If the Less is More hypothesis is true, then we would expect superior performance in the Load vs. the No Load condition.

## Method

Thirty-two participants were recruited for the experiment from the University of Iowa Psychology Department subject pool. The participants received partial credit toward fulfilling requirements for an introductory psychology course. These participants were randomly assigned to two groups of 16 each, constituting the No Load and Load groups for this experiment. The exposure and test materials were reconstructed per the specifications given in Saffran et al. (1997).

In the No Load condition, participants were informed that they would be listening to an artificial language that consisted of a small number of words but no meanings or grammar. They were not told the exact number of words or anything about the structure of those words. The participants were asked to listen to the language and try to figure out where the word boundaries were. They were also told that they would be tested on their knowledge of the word boundaries later in the experiment. These instructions were made as similar as possible to those given in Saffran et al. (1996).

In the Load condition, participants were asked to draw a picture using a computer drawing program. They were informed that an auditory stimulus would play while they drew, and that the experimenter was looking at a certain effect that would be explained to them later in the experiment. The participants were told nothing at the outset of the experiment about the content of the stimulus, nor were they informed that they would be given a test based on the auditory stimulus later on. These instructions were made as similar as possible to those given in Saffran et al. (1997).

The exposure procedure was identical for both the No Load and the Load groups, and consisted of three seven-minute listening periods with five-minute breaks between. This exposure procedure is the same as that used in Saffran et al.

---

<sup>1</sup> This and the following three composite scores were calculated from the data in Saffran (1997).

<sup>2</sup> Saffran (1997) acknowledges that this incidental task was not as incidental as it was in the word boundary experiments. In the incidental learning condition of the phrase structure experiment, participants were told about the nature of the background stimulus and the test they would be given at the end of the drawing period.

(1996). After the three listening sessions were finished, the experiment proceeded to the test phase, in which the participants were asked to listen to each pair of sound items and to decide which of the two sounded more like it came from the stimulus.

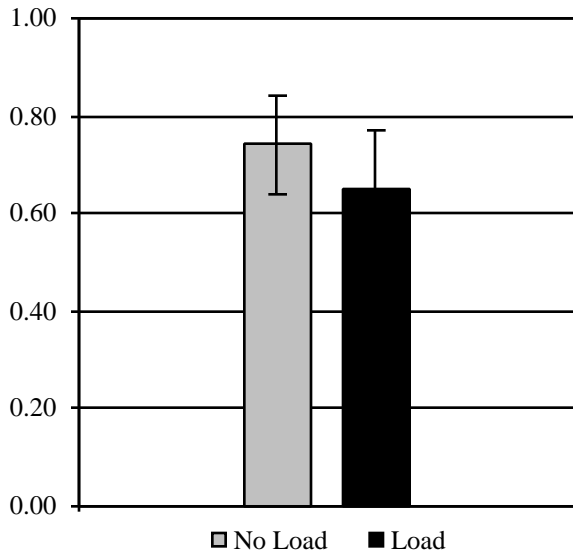


Figure 1: Results of Experiment I, Statistical Learning of Word Boundaries (Mean percent correct on vertical axis).

## Results and Discussion

The results of both groups are in line with the results of the word boundary learning experiments of Saffran et al. (1996) and Saffran et al. (1997), and are shown in Figure 1. The mean score for the No Load group was 74%. A single-sample  $t$  test (two-tailed) showed that performance was significantly above chance,  $t(15) = 9.94, p < .01$ . The mean score for the Load group was 65%, which was significantly above chance as well,  $t(15) = 5.20, p < .01$ . A two-sample  $t$  test of the No Load vs. Load means was significant ( $p < .03$ ), indicating superior performance on the part of the No Load group over the Load group.

In the present experiment, both groups exhibited learning. However, the Load group did not outperform the No Load group, contrary to the prediction of the Less is More hypothesis. Rather, this finding is consistent with a More is More hypothesis, that is, with the idea that enhancing cognitive resources enhances cognitive performance.

## Experiment II: Syntactic Agreement

While the results of Experiment I were inconsistent with the Less is More hypothesis, it could be argued that segmenting words from a stream does not involve the sort of componential analysis that Newport (1990) considered necessary for successful language acquisition. Experiment II consists of a task that would require such a

componential analysis: the learning of the pattern of syntactic agreement in an artificial language. The Less is More hypothesis predicts that participants with a reduced working memory capacity (experimentally induced by the imposition of a cognitive load) will perform better than participants with no reduction of working memory.

## Method

Thirty-two adult college students were recruited from the subject pool of the University of Iowa Psychology Department and randomly assigned to two groups (No Load or Load) as in Experiment I.

An artificial language with a small vocabulary and a simple grammar was created for this experiment. The vocabulary of this language consisted of twenty one-syllable noun roots and twenty one-syllable verb roots. The grammar consisted of two rules. First, all sentences were two words in length, each composed of a noun followed by a verb. Second, the noun and verb of each sentence agreed in number, with singular nouns marked with the suffix *-bo*, plural nouns with *-za*, singular verbs with *-ki*, and plural verbs with *-nu*. Thus, the noun *da* and the verb *me* could form both the sentence *da-bo me-ki* (singular) and the sentence *da-za me-nu* (plural). The exposure and test corpuses were set up such that all the words in the test phase had been heard in the exposure phase, but that all of the test sentences were new.

For the sake of comparison across Experiments, the instructions and procedures in Experiment II were made as parallel as possible to those used in Experiment I.

Subjects in the No Load condition were told that they would hear an artificial language consisting of a series of two-word sentences. They were told nothing about the number of different words or the length of the words. The participants were told that this language was spoken by a computer speech-synthesis program that did not put pauses between words or sentences. Their task, then, was to listen to the language and see if they could figure out where the sentence breaks were supposed to be. They were also told that they would be tested on their ability to find the sentence breaks at the end of the experiment. The rationale for giving the participants this task during exposure was twofold: One was to keep the participants focused on the stimulus, and the other was to keep the procedures in Experiment II as parallel as possible to those in Experiment I. During the test phase, the participants were told to listen to each pair of sound items and decide which of the two sounded more like the training stimulus.

Subjects in the Load condition were given a drawing task and cover story as in Experiment I. During the test phase, they were asked to decide which of the two items in each trial sounded more like the stimulus that played while they were drawing. However, at no time were they told about the content of the stimulus.

## Results and Discussion

The results of Experiment II are shown in Figure 2. Mean performance of the No Load group was 56%. A single-sample

*t* test showed that performance was significantly above chance,  $t(15) = 4.57, p < .01$ . Mean performance of the Load group was 52%. This performance was not significantly above chance,  $t(15) = 1.23, n.s.$  A two-sample *t* test comparing No Load vs. Load means, however, was significant,  $p < .05$ .

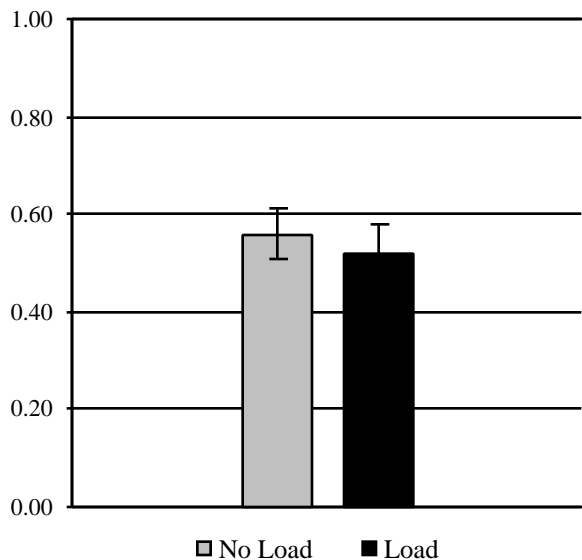


Figure 2: Results of Experiment II, Statistical Learning of Syntactic Agreement (Mean percent correct on vertical axis).

Contrary to the predictions of the Less is More hypothesis, the No Load group did not outperform the Load group; on the contrary, the Load group significantly outperformed the No Load group. However, the null result in the Load condition makes the findings of this experiment hard to interpret. It is not clear whether the Load group failed to learn due to the inherent difficulty of the task, or if they would have exhibited learning had they been given a longer exposure period. When Saffran et al. (1997) increased exposure in their incidental word boundary learning task to two 21-minute sessions on consecutive days, the participants' performance improved significantly. However, the participants in Saffran's (1997) phrase-structure learning experiments showed no significant improvement from Day 1 to Day 2 in either the explicit (No Load) (Saffran, 1997) or incidental (Load) (Saffran, 1997) conditions. At the very least, the results of the present experiment suggest that increased working memory capacity leads to better performance. There could also be a role for attention in the acquisition of syntax, as the null result in the Load condition suggests that syntax may not be learnable at all without attention.

Results across Experiments I and II were analyzed in a two-way ANOVA of task (Word Boundaries vs. Syntactic Agreement) by condition (No Load vs. Load). The results, as shown in Figure 3, indicate main effects for both task and load, but no interaction. In other words, performance

in the Word Boundary task was significantly better than in the Syntactic Agreement task, regardless of condition. Likewise, performance in the Load condition was significantly worse than performance in the No Load condition, regardless of task. The results of this ANOVA suggest, first of all, that the syntactic agreement task was inherently more difficult than the word boundary task was. In addition, they suggest that the imposition of a cognitive load leads to reduced performance in either of these tasks, a finding that runs counter to the predictions of the Less is More hypothesis.

### General Discussion

Experiment I clearly indicates that, at least with regard to the segmentation of words in a speech stream based on statistical regularities, the Less is More hypothesis does not hold. Under that hypothesis, we would expect to see better performance on the part of the Load participants. What we see instead is significantly better performance on the part of the No Load participants, exactly the opposite of what we would expect if the Less is More hypothesis were true. The same pattern of results obtained in Experiment II, suggesting that the Less is More hypothesis does not hold in the domain of syntax acquisition, either.

A key issue in the Less is More hypothesis is the issue of the role of working memory in language acquisition. The Less is More hypothesis posits that the restricted working memory in children aids them in language learning, and furthermore maintains that the larger working memory capacity of adults hinders their language learning ability. An alternative to the Less is More hypothesis would be a More-is-More hypothesis predicting that the greater the cognitive resources available, the better the language learning (or any other) performance will be. The data presented in this paper are consistent with that view.

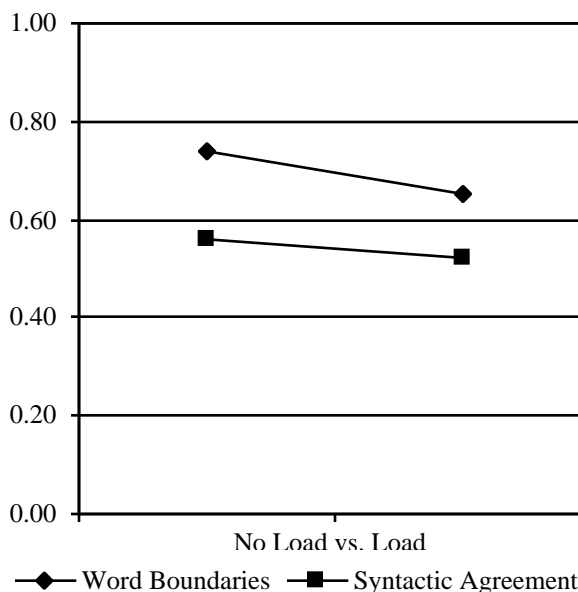


Figure 3: Result of ANOVA across Experiments I and II.

Even if it were the case that the lower performance in the No Load condition were in part because participants were not attending to the stimulus (and not merely because of working memory limitations imposed), this would still in some sense represent a reduction of available cognitive resources and thus, according to the Less is More hypothesis, should still result in better performance. We recognize that there is a potential confound in the two experiments presented here between the effect of the manipulation on working memory and on attention. Research currently underway in our lab is testing the separate effects of working memory and attention on the acquisition of syntax.

We would also like to address three concerns with the present research that have been brought to our attention. The first concern is that Experiment I may not be relevant to the Less is More hypothesis. However, a similar result in both experiments suggests similar mechanisms at work in both the word boundary learning and syntactic agreement learning tasks, making Experiment I relevant to our argument. The second concern is that the dependencies were too close, creating a situation in which the Less is More hypothesis would predict no advantage for a limited working memory capacity. But if this were the case, then we would expect no difference in performance between the Load and No Load groups. Rather, we see that even when dependencies are only a syllable apart, the Load group performs significantly worse than the No Load group. The third concern is that the low performance in Experiment II may be due to having too many words to learn; in other words, syntax acquisition was hindered by the demands of vocabulary acquisition. However, it is not at all clear that it is necessary to learn the words in order to learn the syntax. For example, it is doubtful that the infants in Santelman & Jusczyk (1998) knew all the words in the sentences they heard; nevertheless, they were sensitive to the long-distance dependency being tested for.

To the extent that the Less is More hypothesis is challenged, the question is raised of how to account for the observed critical period effect (Lenneberg 1967) in language acquisition. The Less is More hypothesis makes the implicit assumption that the only relevant difference between children and adults approaching the language learning task is working memory capacity. However, it is very likely that there are other differences, motivational in particular, between the conditions under which children and adults enter a language learning situation besides just working memory capacity (Schuman, 1975, as cited in Johnson & Newport, 1989). In fact, empirical evidence suggests that when motivational factors are held constant in a laboratory situation, children fare worse than adults, as they did, for example, in the experiments reported by Saffran (1997).

## Conclusion

After seven years in Japan, Herrigel (1953) finally masters the bow, learning to send the arrow to its target with apparent effortless. Yet behind that appearance of ease lies seven years of struggle. Seeming effortless is the goal in mastering the bow, not the means to mastering it. Likewise in mastering a language. Facility in a language is achieved only by an arduous, extended process. The language learning process demands a great expense of cognitive effort, and it only stands to reason that the more cognitive resources one has available, the more likely one is to succeed at the task. This premise is borne out by the evidence presented here: adults performed better at language learning tasks when there were no other cognitive demands placed on them. At least for the aspects of language acquisition examined here, it is clear that less is less, not more.

## Acknowledgments

We would like to thank Steven Luck and Rochelle Newman for many helpful discussions.

## References

- Elman, J.L. (1990). Finding structure in time. *Cognitive Science* 14, 179-211.
- Elman, J.L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition* 48, 71-99.
- Gathercole, S.E., Willis, C.S., Baddeley, A.D. & Emslie, H. (1994). The children's test of nonword repetition: a test of phonological working memory. *Memory* 2, 103-127.
- Herrigel, E. (1981). *Zen in the art of archery*. New York: Vintage Books.
- Johnson, J.S. & Newport, E.L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* 21, 60-99.
- Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley.
- Newport, E.L. (1990). Maturation constraints on language learning. *Cognitive Science* 14, 11-28.
- Saffran, J.R. (1997). *Statistical learning of syntactic structure: mechanisms and constraints*. Doctoral dissertation, Departments of Brain & Cognitive Sciences and Linguistics, University of Rochester, Rochester, NY.
- Saffran, J.R., Newport, E.L. & Aslin, R.N. (1996). Word segmentation: the role of distributional cues. *Journal of Memory and Language*, 35, 606-621.
- Saffran, J.R., Newport, E.L., Aslin, R.N., Tunick, R.A. & Barrueco, S. (1997). incidental language learning: listening (and learning) out of the corner of your ear. *Psychological Science*, 8, 101-105.
- Santelmann, L.M. & Jusczyk, P.W. (1998). Sensitivity to discontinuous dependencies in language learners: evidence for limitations in processing space. *Cognition* 69, 105-134.



# Two Views are Better than One: Epistemic Actions May Prime

Paul P. Maglio  
IBM Almaden Research Center  
San Jose, California  
pmaglio@almaden.ibm.com

Michael J. Wenger  
Department of Psychology  
University of Notre Dame  
mwenger1@nd.edu

## Abstract

*Epistemic actions* are physical actions people take more to simplify their internal problem-solving processes than to bring themselves closer to an external goal state. In the video game Tetris, for instance, players routinely over-rotate falling shapes, presumably to make recognition or placement decisions faster or less error-prone. Along these lines, an experimental study was performed to test the hypothesis that it is easier to recognize a two-dimensional shape if it is presented in two different orientations than if it is presented in only one. In particular, we tested whether performance on a shape-based video game task was facilitated by multiple views of a shape, and whether game performance (an indirect test of memory) differed from a direct test of memory for previously presented shapes. Results show that indeed task performance is both faster and more accurate when participants see two views of a shape than when they see one, but that more than two views do not improve performance further. In addition, multiple views lead to faster performance on the video game than on the memory test, but only in the earliest stages of training. We conclude that Tetris players may rotate falling shapes manually to see the shapes in more than one orientation, which leads to faster and more accurate placement decisions.

## Introduction

Studies of people playing the video game Tetris have shown players often take actions in the external environment that are not strictly necessary but that serve to simplify or speed up internal cognitive or perceptual operations (Kirsh & Maglio, 1994; Maglio, 1995; Maglio & Kirsh, 1996). Playing Tetris involves maneuvering falling two-dimensional shapes into specific arrangements on the computer screen (see Figure 1). It was found that even as players become faster with practice, they also tend to over-rotate falling shapes, leading to backtracking in the task environment as these over-rotations are corrected. To make sense of this backtracking, Kirsh and Maglio (1994) argued that sometimes physical rotation can serve the same purpose as mental rotation, effectively offloading mental computation onto the physical world (for other examples, see Clark, 1997; Kirsh, 1995; Maglio, Matlock, Raphaely, Chernicky & Kirsh, 1999). Such physical actions—taken to simplify internal cognitive computation rather than to move closer to the external goal state—are called *epistemic actions*.

Recent work suggests that mental rotation and physical rotation share at least some internal processes (e.g.,

Wexler, Kosslyn & Berthoz, 1998; Wexler & McIntyre, 1997; Wohlschlager & Wohlschlager, 1998). Specifically, physically rotating objects can be shown to facilitate or to inhibit mental rotation under certain conditions. The epistemic function of physical rotation in Tetris, therefore, might be far more complex than is suggested by the simple idea that physical rotation can substitute for mental rotation. In fact, Kirsh and Maglio (1994) speculated that physical rotation might serve the epistemic function of cueing retrieval. Because physically rotating a game piece (which we call a *zoid*) in Tetris provides the player two views of it (i.e., in each of two orthogonal orientations), it is possible that seeing two views makes retrieval of relevant information easier than does seeing just one. This idea makes computational sense; for example, if one conceives of memory in terms of an attractor space, such as a Boltzman machine, the first presentation of the shape is like placing the system near the top of the energy sink that represents the target shape in memory, and the second pushes the system closer to this attractor.

Of course, if shape recognition is orientation-dependent (Tarr & Pinker, 1989; Tarr, 1995; Ullman, 1989), we would not expect multiple views of a single shape to speed up recognition. However, it has been shown that shape identification can be facilitated when primed with orientations different from the target orientation (Cooper, Schacter, Ballesteros & Moore, 1992; Srinivas, 1995). Moreover, numerosity judgments can be facilitated even when test stimuli are not presented at the same orientation as the originally learned patterns (Lassaline & Logan, 1993), suggesting memory for the pattern may not require that the retrieval cue be specifically oriented.

That an epistemic action might cue retrieval raises the possibility that such cueing might be limited to specific types of retrieval demands. In particular, the effects of cueing might depend on whether the task requires *direct* or *indirect* access to memory information. Demands for retrieval while playing Tetris can be thought of as *indirect* tests of memory in that they allow for effects of prior experience to be expressed without requiring explicit memory for the original experience (e.g., Richardson-Klavehn & Bjork, 1988). Tasks requiring explicit memory for the original event—such as old/new recognition or recall—are referred to as *direct* tests of memory. Previous work has shown that direct and indi-

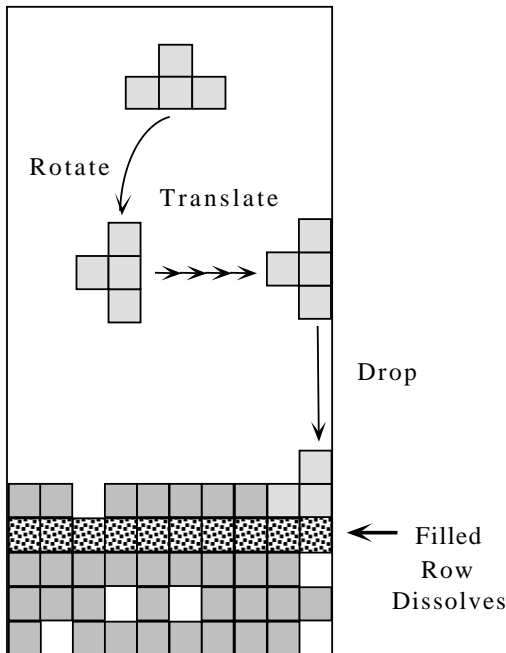









Figure 1: In Tetris, two-dimensional shapes fall one at a time from the top of the screen, eventually landing on the bottom or on top of shapes that have already landed. There are seven shapes, which we call *zoids*—, , , , , , . As a zoid falls, it can be rotated, and moved to the right or left. The object of the game is to fill rows of squares all the way across the screen. When a row is completely filled, it dissolves and all partially filled rows above it move down. The game ends when unfilled rows pile up to the top, blocking new zoids from falling.

rect tests of memory are differentially sensitive to characteristics such as orientation, object symmetry, and other physical aspects of visual objects (Srinivas, 1995, 1996; Srinivas & Schwoebel, 1998). Thus, in the experiment presented here, we used both direct and indirect assessments of memory to determine how effective previews are under different retrieval demands. In addition, because the effectiveness of memory cues generally depends on the time that elapses between presentation of cue and presentation of the item to be retrieved, we investigated the effect of various delays between onset of the first preview and onset of the test zoid by embedding the previews in a sequence of zoids presented prior to test.

In this paper, we empirically test the hypothesis that two different views of a falling zoid are better than one. In addition, we examine whether such a potential benefit might depend on the orientation of the preview relative to the zoid that must be placed, and whether these previews facilitate zoid recognition and Tetris performance.

## Method

To test whether two views of a falling zoid leads to faster or more accurate performance in Tetris than does one, we




created a controlled experimental situation that shared many attributes with the game of Tetris but that allowed fine-grained control over the parameters of interest. In our experimental set up, a Tetris configuration (i.e., a Tetris board and zoid floating above it) is preceded either by none, one, or two previews of the zoid in either the same or different orientations (see Figure 2). The participant's job is to quickly and accurately determine whether the zoid fits snugly on the board. Thus, the task creates situations similar to those faced by Tetris players during an actual game, and also requires responses similar to those required of players during an actual game.

Participants spent three days (one hour each day) playing this experimental version of Tetris. Separate groups of participants were required either (a) to make judgments about whether a target zoid fit in an accompanying board (the indirect test), or (b) to make this judgment *and* indicate whether they remembered seeing the test zoid in the set of zoids that were presented prior to the target (the direct test). Between 0 and 2 previews of the target zoid were presented in a sequence of zoids prior to the target, and the orientation of these previews (when present) varied relative to the target. As noted, by placing the previews in a sequence of events prior to the test, we were able to manipulate the interval over which the preview would have to be retained in memory.

## Participants

A total of 30 participants were recruited from psychology courses and participated voluntarily in exchange for course credit. All participants reported normal or corrected-to-normal vision and unencumbered use of both hands.

## Design

The experiment was conducted as a 3 (number of previews: 0, 1, 2)  $\times$  3 (orientation of the first preview relative to the target zoid: same, clockwise rotation of 90°, counter-clockwise rotation of 90°)  $\times$  3 (retention interval between first preview and target zoid, in frames: 0, 1, 2)  $\times$  3 (zoid type: , , )  $\times$  2 (status of target zoid relative to the board: fit, not fit)  $\times$  3 (day of testing: 1, 2, 3)  $\times$  2 (type of memory judgment at test: direct, indirect) mixed factorial design. All factors except type of zoid and type of memory judgment were manipulated within participants.

## Materials

All zoids and boards were constructed from 20  $\times$  20 pixel squares. Squares were outlined by light gray lines, 1 pixel in width, and were filled in solid black. The background for all displays was solid black as well. All zoid types were composed of four blocks. All receptor boards were six blocks in height and width. Four receptor types were defined for each zoid type, corresponding to four ways in which the zoid could be snugly placed. Each receptor type was used with equal frequency. Materials were displayed on a 33 cm VGA monitor controlled by a PC-compatible microcomputer. Onset and offset of each display was synchronized to the vertical scan of the

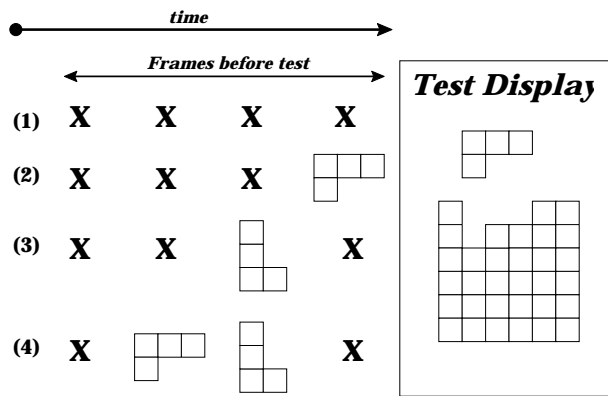


Figure 2: A schematic representation of some of the events in four frames prior to a test display in a single trial. The Xs indicate non-target zoids. (1) The events in a 0-preview trial. (2) The events in a 1-preview trial, with no retention interval (0 frames) between the preview and the test display. (3) The events in a 1-preview trial with a 1-frame retention interval between the preview and the test display. Here the preview is rotated 90° counter-clockwise relative to the test display. (4) The events in a 2-preview trial with a 2-frame retention interval between the *first* preview and the test display. Here the first preview has the same orientation as the test display, while the second preview is rotated 90° counter-clockwise relative to the test display.

monitor. The standard PC keyboard was used to collect and time (to  $\pm 1ms$ ) participant responses.

## Procedure

Participants were tested on three consecutive days, at approximately the same time each day, with each session lasting approximately 1 hour. All sessions were conducted in a darkened room, with participants seated at an unconstrained distance from the monitor, and began with a five min period for dark adaptation. Participants were told that, on each trial, they would see a sequence of zoids, presented very rapidly. At some random point in this sequence, they would see a combination of a zoid and a receptor board, and would need to make one of two types of responses, depending on whether they were in the indirect or direct memory test condition.

In the indirect condition, participants simply had to decide whether the presented piece would fit snugly into the board. Participants responded in the affirmative using the index finger of their dominant hand, and in the negative using the index finger of their non-dominant hand, pressing either the “z” or “/” keys on the lower row of the PC keyboard. In the direct condition, participants had to indicate with a single key-press both their judgment about whether the presented piece fit snugly in the board and their memory for any occurrence of the

test piece (in any orientation) in the sequence of pieces that preceded the target piece. Participants responded with the index finger of their dominant hand if the target piece fit and they remembered seeing this piece in the preceding sequence, with the middle finger of their dominant hand if the target piece fit and they did not remember seeing this piece in the preceding sequence, and with the index finger of their non-dominant hand if the piece did not fit.<sup>1</sup> Speed and accuracy were equally emphasized.

Each trial began with the presentation of between one and eight zoids (“non-target zoids”) designed to be distinct from the target zoid assigned to the participant. The actual number of these non-target zoids shown was randomly determined for each trial. Each non-target zoid was presented for 250 ms and then replaced by the next non-target zoid; the non-target zoids in this sequence did not repeat (i.e., all were unique). Following this, four zoids (between 0 and 2 target zoids, and between 2 and 4 non-target zoids) were presented for 250 ms each. After the last of these were presented, a target zoid and a receptor board were presented for 250 ms. Following the participant’s response, a tone was briefly sounded (100 ms) indicating a correct (880 Hz) or incorrect (440 Hz) response.

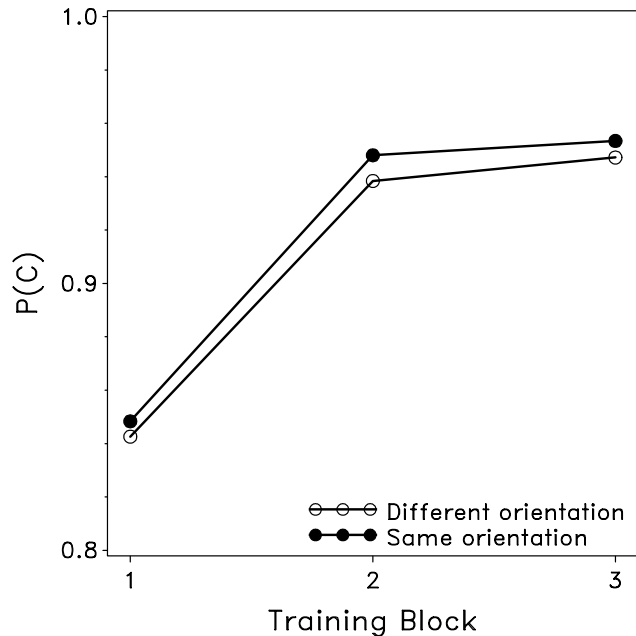
A total of 480 trials were presented in each session. Participants were allowed short breaks after every 80 trials. Feedback on overall accuracy and mean response time was provided at the end of each session.

## Results

First, we asked whether having one preview improved performance over having no previews, and found a pronounced effect in both accuracy and response time (RT). When participants were presented with a single preview, the resulting level of accuracy was significantly higher (0.86) than when they were not presented with a preview (0.53),  $t_{(1,59)} = 33.85, p < 0.001$ . Similarly, when participants were presented with a single preview, the resulting RTs were significantly shorter (869 ms) than when they did not see a preview (1791 ms),  $t_{(1,59)} = 2.01, p < 0.05$ .

Given that providing a preview had an effect on performance, we moved on to determining whether having *more than one* preview had an additional effect, and whether the provision of previews interacted with our other experimental factors. Our analysis of the accuracy data indicated that zoid, number of previews (1 vs. 2), and retention interval all failed to have an effect on accuracy (all  $F_s < 1.00$ ). However, test type did have a significant impact on performance, with participants in the direct test condition performing at a higher level of accuracy (0.95) than participants in the indirect condition (0.88),  $F_{(1,25)} = 4.59, MSE = 0.05$ . Orientation of the prime exerted a statistically significant effect on accuracy,  $F_{(1,25)} = 4.01, MSE = 0.01$ , but the magnitude of the difference between the previews presented in the

<sup>1</sup>We did not ask for a memory judgment on trials in which the piece was judged not to fit, as our primary concern was with the effects of previews on accurate placement of pieces in the board.



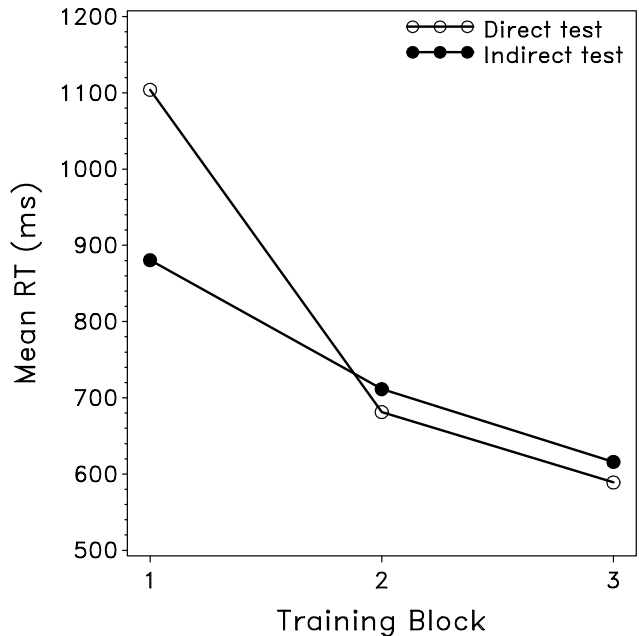
**Figure 3:** Effects of orientation of preview and block on accuracy. Practice affects the probability of making a correct response. However, whether the zoid was previewed in the same orientation or in a different orientation (as the test zoid) does not affect the probability of making a correct response.

same orientation (0.92) and those presented in a different orientation (0.91) suggests that the difference may not be meaningful. Exploration of these data across blocks of experience (see Figure 3) suggests that the difference between the two forms of preview was induced by the fact that performance with previews in a different orientation did not improve quite as quickly from the first to the second training block as did performance with previews in the same orientation, though this interaction was *not* significant. Finally, as expected, performance improved consistently across blocks,  $F_{(2,50)} = 6.67$ ,  $MSE = 0.03$ , as can be seen in Figure 3.

Analysis of the RT data indicated that test type, zoid, number of previews, orientation of the preview, and retention interval all failed to affect the speed of responding (all  $F_s < 1.00$ ). Although RTs consistently improved across the experiment,  $F_{(2,50)} = 57.56$ ,  $MSE = 44847.84$ , the form of improvement was dependent on test type (direct vs. indirect),  $F_{(2,50)} = 7.03$ ,  $MSE = 44847.84$ . As shown in Figure 4, the direct test condition (which required *two* response judgments) was slower than the indirect test condition (which required one response judgment), but only in the first block of trials.

## Discussion

Our results show that if participants are presented with two views (i.e., one preview) of the falling zoid (a two-dimensional shape), response times are faster than if only a single view (i.e., no previews) is presented. This sup-



**Figure 4:** Effects of test type and block on mean RT. Participants in the indirect test condition (i.e., deciding whether the zoid fits snugly) respond faster than participants in the direct condition (i.e., deciding whether the zoid fits *and* whether the zoid had been previewed) only on the first day of practice.

ports our hypothesis that *two views are better than one*. Nevertheless, it was a bit surprising to find that three views provide no advantage over two views. In terms of the simple Boltzman machine model mentioned previously, this would mean that the second view of the zoid pushes the system so close to the attractor that it is trapped, and so the third view is rendered irrelevant. Alternatively, the effect of the first preview might be to accelerate the system toward the attractor state to such an extent that a second preview provides no appreciable additional acceleration.

Note that response time was speeded up by a preview in any of the three orientations relative to the test zoid. The benefit was not restricted to previews that shared orientation with the test display. This finding is consistent with priming studies in which it was found that a prime need not be presented in the same orientation as the target to facilitate recognition or identification (e.g., Cooper, Schacter, Ballesteros & Moore, 1992; Srinivas, 1995). It is surprising, however, to find that different orientations prime just as strongly as the test orientation does. One possible explanation is that participants have stored multiple views of the zoids and so seeing one view is just as good as seeing another (Tarr & Pinker, 1989).

The only difference between the direct and indirect tests of memory was observed on the first day of training, and restricted to the latency data. On the first day, participants in the direct test condition required more time

to respond than did the participants in the indirect test condition. This difference may be easily accounted for by the fact that participants in the direct test condition had to make two response decisions and choose among three response alternatives. The *lack* of a difference in either accuracy or latency as a function of memory test suggests that the benefits obtained by having a preview do not depend on the manner in which memory for that preview is assessed.

Returning to the idea of epistemic action in Tetris, these data suggest that by rotating the falling zoids, players may be able to effectively cue themselves, enabling quicker responses in a Tetris situation. Previous research has established various ways in which Tetris players take actions for their epistemic effects (Kirsh & Maglio, 1994; Maglio, 1995; Maglio & Kirsh, 1996). The data reported here show that a preview of the falling zoid at least speeds up performance on a Tetris-like task, but the hypothesis that Tetris players over-rotate zoids in order to speed up performance is not directly tested. It remains to be seen whether actually taking the action of orienting the preview (i.e., physically rotating the falling shape) is a critical component of performance, independent of the presentation of the preview itself.

In the end, we can conclude that two sequentially presented views of the falling zoid lead to faster and more accurate performance than a single view of the falling zoid. In addition, it appears that having this single preview is sufficient to boost performance to something of a limit, as more than one preview adds little if any additional help. It also appears that the benefit of the preview is robust across the retention intervals considered here. Thus, if players are able to use rotations to self-cue, they may be able to get all they need from a single rotation, even one that is somewhat separated in time from the eventual judgment. The payoff associated with a small number of additional steps more than compensates for the temporal and physical costs of executing additional steps. The epistemic functions of physical rotations in Tetris, then, might not be merely to substitute for mental rotation or to provide a visual means for matching the contour of the board with contour of the falling shape, but also to cue or prime retrieval from memory of information associated with the falling shape, enabling faster recognition and faster placement decisions.

### Acknowledgments

Thanks to Chris Campbell and Teenie Matlock for many thoughtful comments on a draft of this paper. Thanks also to Khara Guttierrez, Nicole Silva, Rhonda Czapla, and Nathan Shaver for assistance in data collection.

### References

Clark, A. (1997). *Being there: Putting body, brain, and world together again*. Cambridge, MA: MIT Press.

Cooper, L. A., Schacter, D. L., Ballesteros, S., & Moore, C. (1992). Priming and recognition of transformed three-dimensional objects: Effects of size and reflection. *Journal of Experimental Psychology: Learning Memory and Cognition*, *18*, 43–57.

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence*, *73*, 31–68.

Kirsh, D. & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*, 513–549.

Lassaline, M. E. & Logan, G. D. (1993). Memory-based automaticity in the discrimination of visual numerosity. *Journal of Experimental Psychology: Learning Memory and Cognition*, *19*.

Maglio, P. P. (1995). *The computational basis of interactive skill*. Doctoral dissertation, University of California, San Diego.

Maglio, P. P. & Kirsh, D. (1996). Epistemic action increases with skill. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, pages 391–396, Mahwah, NJ. Lawrence Erlbaum.

Maglio, P. P., Matlock, T., Raphaely, D., Chernicky, B., & Kirsh, D. (1999). Interactive skill in Scrabble. In *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society*, pages 326–330, Mahwah, NJ. Lawrence Erlbaum.

Richardson-Klavehn, A. & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology*, *39*, 475–543.

Srinivas, K. (1995). Representation of rotated objects in explicit and implicit memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, *21*, 1019–1036.

Srinivas, K. (1996). Contrast and illumination effects on explicit and implicit measures of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1123–1135.

Srinivas, K. & Schwoebel, J. (1998). Generalization to novel views from view combination. *Memory & Cognition*, *26*, 768–779.

Tarr, M. & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, *21*, 233–282.

Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin and Review*, *2*, 55–82.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, *32*, 193–254.

Wexler, M., Kosslyn, S. M., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition*, *68*, 77–94.

Wexler, M. & McIntyre, J. A. (1997). Is mental rotation a motor act. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, pages 808–813, Mahwah, NJ. Lawrence Erlbaum.

Wohlschlagel, A. & Wohlschlagel, A. (1998). Mental and manual rotation. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 397–412.

# Preschool Children's Use of Category Information to Interpret Negations

Bradley J. Morris (bmorris@andrew.cmu.edu)

Dept. of Psychology, Carnegie Mellon University  
5000 Forbes Ave., Pittsburgh, PA 15213

## ABSTRACT

Two experiments examined 4 and 5 year old children's use of semantic category to interpret negations. In an 'I Spy' game, children were given a hint in the form of a negation then instructed to select a referent in a forced-choice procedure. Children used category information to infer that the referent was semantically related (near-neighbor) rather than semantically unrelated (far-neighbor) to the negated item, though both were logically correct choices. The results suggest that one type of 'pragmatic interpretation' children use for understanding negations is semantic relatedness that reduces the scope and indeterminacy of negations.

An important question in the development of reasoning and communication is how children develop an understanding of logical connectives such as AND, OR, and NOT. Logical connectives are unique problems in language acquisition since they do not directly refer to an object or action but are relational in nature. Negations are particularly vexing since a negation "has no referent...and is inherently indeterminate," (Pea, 1980, p. 156) referring to the absence of an object or set of objects. Negations pose a unique problem in reasoning since they do not specify a clear referent. For example, the statement 'Flipper is not a fish' does not indicate what 'Flipper' is; only what it is not.

A negation is a simple syntactic marker for changing the truth-value of an affirmed statement (Johnson-Laird, 1983). Previous research on understanding negations indicates that processing a negation involves two cognitive operations: creating a representation of an item, then inhibiting this representation (Johnson-Laird, 1983; MacDonald & Just, 1989).

The 'classical' interpretation of a negation is that it refers to anything outside of a designated set equally (Horn, 1989). In a classical interpretation of the previous example, 'Flipper' might refer to a whale, a human, or a car. A classical interpretation of negations presents two main difficulties in communication and reasoning: a) scope, or the limit to the objects or set of objects to which the negation refers and b) indeterminacy of reference, or determining to what a negation refers.

Negations are interpreted classically in formal reasoning- i.e.; a negation includes all objects except those that have been negated (Horn, 1989). However, cognition is bounded; that is, people consider only a small number of possibilities at any time. Thus, since cognition is bounded, people do not consider all possibilities for a negation since this would create a problem set far too

large to be cognitively tractable. Thus, to make operations tractable, children (and adults) must reduce the number of possible solution states, though they are capable of considering more than one possibility (Horobin & Acredolo, 1989). Therefore, either the number of states is reduced randomly or there is a process of determining which states are to be maintained. Such a heuristic must operate within young children's knowledge of negations.

In problem solving, an understanding of logical connectives is necessary to establish the mapping between evidence and form in order to draw a correct conclusion. An understanding of logical connectives is crucial to performance on logic problems. However no current theory gives a principled explanation of how syntactic and semantic information influences their interpretation (Johnson-Laird, 1983).

## Young children's knowledge of negation

By the end of preschool, children have some understanding of negations. Children are capable of assigning truth-values for negations (Kim, 1985). For example, when shown a banana and told that the object is not an apple, children are capable of judging the statement as true. Children are also sensitive to syntactic markers and how these markers limit the scope of negations (De Boysson-Bardies, 1977; Romaine, 1988). For example, children understand that negations refer to particular parts of speech (e.g., noun phrases) due to the position in a sentence. Finally, children have several functional uses for negations such as denying a request ("do you want some juice?" "No") or expressing disappearance ("No juice" when glass is empty) (Bloom, 1970; Pea, 1980).

One question that remains is how children infer a referent for a negation. That is, given a negated noun phrase, syntax alone is insufficient for determining a referent because all 'nouns' could be equally plausible. Another strategy seems necessary inferring a referent. Previous research suggests that children may use linguistic and non-linguistic cues to help them resolve problems of indeterminate reference (Oaksford & Stenning, 1992).

One such cue may be provided by semantic categories. A well-documented finding in developmental research is that young children have the capacity for category-based reasoning because "members of object categories...share deep, underlying commonalties" (Waxman et al., 1997, p. 1074). This category-based information can be used for induction in which the

properties of one entity are extended to another on the basis of similarity.

### **Semantic Category and Memory**

Previous research has also indicated that there is a reciprocal relationship between semantic categories and the structure of memory (Brainerd, Reyna, & Kneer, 1995; Ackerman, 1997). Encoding and recall of items in memory are related to semantic category because accessing an item tends to activate items within the same category more strongly than functionally related items (Brainerd, Reyna, & Kneer, 1995). Category-based information may also function as context, making similar items more salient, aiding retrieval because the process of categorization itself may form associations between concepts (Ackerman, 1997).

Category information may also interfere with retrieval. In the False Recognition Paradigm when similar items are activated increases in errors are directly related to the level of similarity between the distracter and the negated item (Brainerd, Reyna, & Kneer, 1995). Further, negating an item does not seem to reduce the salience of related items. Activation levels of similar items were increased when a target item was negated; even when the items themselves were contextually suppressed (MacDonald & Just, 1989).

To summarize, understanding negations involves a combination of syntactic and semantic/pragmatic processes. Syntax determines the part of speech that a negation modifies (e.g., noun phrase) providing a limit on the scope of a negation. Semantic understanding of negations is a two-step cognitive operation in which an item is represented, then inhibited. When the represented items are accessed in working memory, they activate similar items (i.e., other category members). Because activated items are inhibited in negations, then it is possible that category information guides induction of possible referents by providing a contextual 'frame' in which pragmatic inferences can be drawn. Thus, negating an item may provide a cue to the range of items to which a negation refers by providing context.

There is indirect evidence that category information may provide context for interpreting a negation. First, in a study examining conditional reasoning, phonological cues focused reasoners on intended contrasts (Oaksford & Stenning, 1992). For example, in the sentence *Tim did not travel from Chicago to Pittsburgh by car*, if one stresses the word *car* participants tended to infer probabilities of the mode of transportation *Tim* used mediated by the knowledge of the trip (i.e., plane, train).

A second example is taken from an early study of young children's understanding of negation. In a series of class-inclusion experiments, children were given a collection of objects that could be classified on various

dimensions such as shape or color and given instructions (phrased as negations) to sort these objects on one dimension (e.g., things that are NOT green) (Inhelder & Piaget, 1964). The results indicated that 5-9 year old children did not sort objects using class-inclusion rules (e.g., failing to understand the hierarchical inclusion of 'blue triangles' within the class of 'triangles'). A closer examination of the data indicates that a prominent error pattern was to sort objects on one dimension (e.g., shape). For example, when told to select objects that were 'NOT red circles,' children would often select only red triangles, ignoring other possible responses (such as other circles or non-red triangles). Perhaps 'circle' guided participants to infer that they should focus on a shape-based set of objects. A final example is drawn from a semi-structured interview in which a child implicitly states how such category-based inferences are useful for interpreting negations (from Inhelder & Piaget, 1964, p. 141):

Piaget: "And is it more correct to say that a cow isn't a bird, or that a house isn't. Or are both equally correct?"

Ros: "It's a little ridiculous to say that a house isn't a bird."

Piaget: "And a cow?"

Ros: "Well, it is an animal!"

There seem to be three possibilities for how young children might use category information to interpret negations. The first is that they simply do not use this information. However, if children do use this information, then there are (at least) two possibilities that reflect contrary pragmatic interpretations. One interpretation infers that the referent is something like the negated item. This would result in a 'near-neighbor' inference in which children would look for something within the same category as the negated item. A second interpretation infers that the referent is something unlike the negated item. This would result in a 'far-neighbor' inference in which children would infer that the referent is something outside the category of the referent. Using category information may help reduce the number of possibilities corresponding to a negation by providing a framework for evaluating which items may be relevant - either items that are closely related to the negated item (near-neighbor) or items that are unrelated to the negated item (far-neighbor). Using the category information to infer either type of relationship between the negated item and the referent demonstrates a structured understanding of pragmatics and an attempt to infer the meaning of the speaker.

The present study examines two questions. First, do young children use category information to infer a probable referent for a negation? If children do not use category information then the number of far-neighbor and near-neighbor choices should not differ from chance.

Second, if children use category information, do they infer that the referent is something like the negated item (near-neighbor) or something unlike the negated item (far-neighbor)? If the number of choices differs from chance, then this tendency reflects one of two possibilities. Above chance selection of near-neighbor items suggests that the child inferred that referent is something like the negated item. An above chance selection of far-neighbor items suggests that the child inferred that the referent is unlike the negated item.

Two experiments were conducted to examine young children's use of category-based inferences to induce the referent of a negation. A forced-choice paradigm was used for all experiments in which participants were presented with three objects: a negated item and two choices. Children were instructed that they were playing an "I Spy" game. Children were given a hint as to what the experimenter was "spying" phrased as a negation. They were then asked to infer the referent of the negated item by making a choice between the objects and placing it into a basket. The objects represented *logically correct* choices but differed only in the *degree of relatedness* to the negated object.

Three semantic categories were used for both experiments: animals, vehicles, and foods. Three objects were chosen from each category. Three factors were examined in the series of experiments: the impact of the familiarity of the objects, the number of near-neighbor choices, and the reference set.

In Experiment 1, the experimenter provided three familiar objects from each category and presented a 'hint' in the form of a negation. Children were then asked to choose between two objects: one from within the same category of the negated item and one from a different category. Experiment 2 used the same design as Experiment 1 but used objects that were unfamiliar yet fell into the categories.

## Experiment 1

### Methods

#### Participants

The participants were twenty-one 4-year-olds and 20 5-year-old children from two preschool classrooms. Children ranged in age from 4.1 to 5.5 years (25 girls, 16 boys). Most children were from middle class, white families. Children were selected on the basis of receipt of parental permission.

#### Materials

A total of nine objects were used. The objects were chosen to represent three semantic categories: foods-apple, banana, orange; animals- dog, cat, bunny; vehicles-car, plane, boat. The objects were chosen as familiar

based on rankings taken from the MacArthur Communicative Development Inventory (CDI) that established each object as being in receptive vocabulary before year 2 (Fenson et al., 1994). Each object was similar in size. All children named each object spontaneously.

#### Procedure

The procedure was a forced-choice selection task framed as an "I spy" game in which children were presented three objects: two within the same semantic category and one outside the semantic category. Children were told to guess to which object the experimenter was referring and to place that object in a basket. The child was told "What I spy is *NOT* (emphasized) the (negated object)" and asked to place one object in the basket. Each child was tested individually in a quiet room and took approximately 10 minutes. There were two phases: a warm-up and an experimental phase. The warm-up phase consisted of three questions intended to familiarize each participant with the game and to check understanding of basic negations.

#### Warm-up

The warm-up phase began by asking participant to name all objects and to correct any mistakes. Most participants named each object correctly and all correctly named the object set before warm-up tasks began. The same objects were used in the warm-up and experimental phases. Participants were then presented with three warm-up questions to learn the rules of the 'I spy' game. All children demonstrated an understanding of negations by not choosing target object on three of three trials and continued into the experimental phase.

#### Experimental Phase

Once the child was familiarized with the procedure, each array was presented and the child was told "What I spy is NOT the x" and instructed to place one object in the basket. Once the question was asked, eye contact with the participant and materials was avoided until after the selected object was placed in the basket in order to minimize nonverbal cues. The negated object (A1) always was one of two within the same category. The two possible choices included one from within the semantic category of the negated object (A2) and one from outside the semantic category of the negated object (B1). A total of nine trials were performed in which each object was negated only once and appeared in two other arrays, once as a near-neighbor choice and once as a far-neighbor choice. The placement of the negated object and possible referents was systematically varied. Order of



presentation was counterbalanced. All participants completed all nine trials.

### **Coding**

Responses were coded in one of two categories: within the same category as the negated item (near neighbor) or outside the category of the negated item (far-neighbor). Children could pick one or two objects per trial. If one object was chosen, then it was coded as either within the same category or outside the category of the negated object

## **Results**

Children's choices were examined two ways: across-individuals and within-individuals. Across-individual analyses examined overall response trends while within-individual analyses examined patterns of responses for consistency across the experiment. Responses were coded in one of two categories: near-neighbor or far-neighbor. Preliminary analyses indicated no gender or age differences thus were combined for further analysis.

### **Across-Individual Results**

Seventy-one percent of choices were within the category of the negated object leaving twenty-nine percent outside the category. A confidence interval test was conducted to compare whether children's item selection was at a level different than would be expected if they were choosing items randomly. Seventy-one percent of item choices were within the category of the negated item leaving twenty-nine percent of choices outside the category of the negated item. The selection pattern was significantly different than would be expected by chance ( $p < .01$ , confidence interval 66-75%). This indicates that children selected items from within the same category as the negated item at an above chance level.

### **Within-Individual Responses**

In order to evaluate the consistency of individual participants, a within-individual analysis was performed. A participant was coded as adhering to a pattern if they used the same selection pattern on seven of nine trials. Seven of nine trials represent an above chance pattern of responses whose conditional probability was less than .10. Twenty-eight children were coded as using a consistent response pattern and of these participants, 23 used a near-neighbor selection pattern while 5 used a far-neighbor pattern.

## **Discussion**

The results indicated that young children demonstrated a preference for choosing an object from within the same category as the referent for a negation (though both choices were logically equal). For example, given an apple and a boat and asked "What I spy is NOT a banana", children overwhelmingly selected the apple. Individual analysis revealed a large number of children responded consistently across tasks, primarily using a near-neighbor strategy, in which a near-neighbor object was chosen 7 of 9 times. No age-related differences were found between the 4 and 5 year olds.

These findings suggest that children are sensitive to the semantic information provided in a negation as providing a context for pragmatic interpretation. This information was used most frequently to infer a near-neighbor relationship between the item negated and the referent. Thus, inferring that 'not a cat' is a dog was more frequent than inferring that 'not a cat' referred to a car. It is also plausible that this same marker may indicate that the object is outside of the category of the referent, as demonstrated by the five children who made such an interpretation. However both are clearly category based inference patterns.

Although the results of the study are clear their interpretation could be limited by the familiarity of the materials. Perhaps with familiar objects there are thematic relationships (e.g., dogs and cats are often in the same house) along with the taxonomic relationships, and these additional links increased near-neighbor choices. Thus, a second study was designed to examine the influence of less familiar materials to eliminate the possibility that labels and thematic relations may have influenced the results.

## **Experiment 2**

In order to address the familiarity bias that may have influenced the results of Experiment 1, Experiment 2 extended the same procedure and categories of Experiment 1 using unfamiliar stimuli. A similar procedure was utilized using novel materials (yet within the same semantic categories as Experiment 1) to reduce the possibility that the familiarity of materials might be influencing the results. A second procedural change was introduced to reduce the focus on familiar labels: only naming the target object only during the experimental phase.

## **Methods**

### **Participants**

The participants were 21 4 and 21 5-year-old children from two preschool classrooms in a different preschool than in Experiment 1. Children ranged in age from 4.4 to 5.3 years (22 girls, 20 boys). Most children

were from middle class, white families. Children were selected on the basis of receipt of parental permission.

### **Materials**

Nine objects were chosen as unfamiliar, within the same three semantic categories as Experiment 1: food-eggplant, zucchini, cabbage; animals- lynx, tapir, gazelle; vehicles- diving bell, seacopter, hovercraft. Each object was similar in size.

### **Procedure**

The basic procedure was similar to Experiment 1. The procedure was slightly different in that only the target objects were named. This procedure was used to reduce the emphasis on labels. Responses were coded as in Experiment 1.

## **Results**

### **Across-Individual Results**

Preliminary analyses indicated no significant differences between four and five-year-olds and the two ages were combined for further analysis. A confidence interval test was conducted to compare whether children's item selection was at a level different than would be expected if they were choosing items randomly. Seventy percent of choices were within the category of the negated item and thirty percent of choices outside the category of the negated item. The number of within-category selections was above chance ( $p < .01$ , confidence interval 64-76%).

### **Within-Individual Responses**

As in Experiment 1, a participant was coded as adhering to a pattern if the same selection pattern was used on seven of nine trials. Twenty-two of 42 participants used a near-neighbor selection pattern on at least 7 of 9 trials.

## **Discussion**

Experiment 2 was conducted to replicate the findings of Experiment 1 and examined the possibility that the results of Experiment 1 may have been influenced by the familiarity of the materials. As in Experiment 1 almost all children chose only one object per trial and this object was most often (66%) within the same taxonomic category as the target. Once again there was considerable individual consistency, with 22 children choosing the near-neighbor objects at least 7 of 9 trials. One interesting difference from the previous experiment was that no child consistently chose the far-neighbor object.

These data suggest that when given a choice among unfamiliar objects as the referent of a negation, there is a tendency to choose an object from the same taxonomic category. Thus, the semantic information in a negation provides one clue as to how the negation might be interpreted.

## **General Discussion**

The findings indicated that most children used category information to interpret negations and they used this information to infer that a referent was related to the item negated rather than unrelated. The findings suggest that children tend to make these inferences regardless of whether objects are familiar or unfamiliar.

The first research question investigated the possibility that children used category information to infer the intended referent of a negation. Children selected items at levels above chance; that is, they demonstrated a preference for one type of item, presumably due to category information.

The second research question examined the type of selection preference. There were two possibilities for a selection preference: selecting items within the category of the negated item (near-neighbor) or selecting items outside the category of the negated item (far-neighbor). The two patterns involve different assumptions about the pragmatics of negation. A near-neighbor pattern uses category information to find an item similar to the negated item. For example, NOT CAT would mean DOG (rather than CAR) since both are animals. A far-neighbor pattern uses category information to find an item unrelated to the negated item. Using the previous example, NOT CAT would mean CAR (rather than DOG). The results clearly demonstrated that children selected an object from within the same category as the negated item. Individual analyses indicated that children's near-neighbor selection patterns were quite consistent across the problem set with roughly half of the children selecting near-neighbor items on at least 7 of 9 trials.

The findings suggest that semantic information may help reduce the scope and indeterminacy of negations by providing a contextual 'frame' for pragmatic inference. That is, choosing to negate an item may provide a cue to its interpretation: by choosing to negate item  $x$ , some property of item  $x$  may be relevant to understanding the referent. For example, the sentence 'Whiskers is not a cat' provides a clue that there is something about this object (cat) that is relevant to figure out what 'Whiskers' is- otherwise another object might have been negated. For example, we would probably be more surprised if 'Whiskers' was a book than if 'Whiskers' was a hamster since a near-neighbor interpretation favors the latter. Thus, the pragmatics of a near-neighbor interpretation may reduce the search for

possible referents to items within the category of the negated item (e.g., other pets). This strategy can be formalized using a simple inference rule: given NOT X, then search for items within the category of X as possible referents. Thus, this inference rule combined with semantic category information may provide a powerful tool for inferring the referent for a negation.

The findings suggest a principled explanation of interpreting negations. Since the structure of categories and memory are well established by previous research (Brainerd, Reyna, & Kneer, 1995; Ackerman, 1997), all that is required is a simple pragmatic rule easily derived from experience in which a negation indicates a near-neighbor relationship between the negated item and referent. These findings suggest that the semantic information from negations may provide one source of information with which one reduces the scope and indeterminacy of negations, thus reducing the number of possibilities while maintaining information. A near-neighbor relationship may be common in young children's language environments in word acquisition. For example, when children overextend labels onto unfamiliar objects (e.g., labeling a CAT a DOG) adults often implicitly utilize a near-neighbor negation in correcting the error ("No, that is not a dog, it is a cat).

Finally, these findings may provide one explanation for the interaction between pragmatics and deviations from normative reasoning lacking in current theories of logical development. Understanding children's interpretations of negations is important since children and adults do not appear to use classical logical reasoning (Johnson-Laird, 1983; Sharpe et al., 1996). That is, children and adults rarely solve logical problems as a trained logician would solve them. As noted earlier, current theories of logical development rely on pragmatics to explain performance, yet do not provide explanations of how pragmatics is achieved. Therefore, understanding how pragmatics influences reasoning solutions and strategies is useful for understanding performance and how to improve performance through instruction. This study provides evidence for one type of pragmatic interpretation- a near-neighbor interpretation of negations in which the category of the negated item provides context that guide item selection to an item within the same category.

## References

- Ackerman, B.P. (1997). The role of setting information in children's memory retrieval. *Journal of Experimental Child Psychology*, 65, 238-260.
- Bloom, L. (1970). *Language development: Form and function in emerging grammars*. Cambridge, MA: MIT Press.
- Brainerd, C.J., Reyna, V.F., & Kneer, R. (1995). False-recognition reversal: When similarity is distinctive. *Journal of Memory and Language*, 34, 157-185.
- De Boysson-Bardies, B. (1977). On children's interpretations of negation. *Journal of Experimental Child Psychology*, 23, 117-127.
- Fenson, L., Dale, P.S., Reznick, J.S., Bates, E., Thal, D.J., & Pethick, S.J. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development*, 59(5, Serial No.242).
- Horobin, K., & Acredolo, C. (1989). The impact of probability judgments on reasoning about multiple possibilities. *Child Development*, 60, 183-200.
- Horn, L. R. (1989). *A Natural History of Negation*. University of Chicago Press: Chicago.
- Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child*. New York: Norton.
- Johnson-Laird, P.N. (1983). *Mental Models*. Cambridge, MA: Harvard Press.
- Kim, K.J. (1985). Development of the concept of truth-functional negation. *Developmental Psychology*, 21 (3), 462-472.
- MacDonald, M. & Just, M. (1989). Changes in activation levels with negation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 45-68.
- Oaksford, M., & Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 835-854.
- Pea, R. D. (1980). The development of negation in early child language. In D. Olson (Ed.), *The Social Foundations of Language and Thought* (pp. 156-186). New York: Norton.
- Sharpe, D., Eakin, L., Saragovi, C., & Macnamara, J. (1996). Resolving apparent contradictions: adult's and preschoolers' ability to cope with non-classical negation. *Journal of Child Language*, 23, 675-691.
- Rumain, B. (1988). Syntactics of interpretation of negation: A developmental study. *Journal of Experimental Child Psychology*, 45, 119-140.
- Waxman, S.R., Lynch, E.D., Casey, K.L., & Baer, L. (1997). Setters and samoyeds: The emergence of subordinate level categories as a basis for inductive inference in preschool-age children. *Developmental Psychology*, 33 (6), 1074-1090.

# An Inquiry Into the Function of Implicit Knowledge and its Role in Problem Solving

Timothy J. Nokes (tnokes@uic.edu)  
Stellan Ohlsson (stellan@uic.edu)

Department of Psychology  
The University of Illinois at Chicago  
1007 West Harrison Street (M/C 285)  
Chicago, IL 60607, U.S.A.

## Abstract

Research on implicit learning has shown that the knowledge generated from memorizing patterned symbol sequences can be used to make familiarity judgements of novel sequences with similar structure. However, the degree to which these knowledge representations can be used for subsequent cognitive processing is not known. In this study, participants memorized either patterned number strings (patterned training) or random number strings (random training) and then solved either a number or letter sequence extrapolation problem. Patterned training participants performed significantly better on number problems than on letter problems, thus implying that patterned training influences performance, but only on near transfer problems.

## Function of Implicit Knowledge

To support successful performance on complex, unfamiliar tasks, knowledge must be both abstract and generative. The origin of such knowledge is a central question for cognitive psychologists, developmental psychologists, educators, machine learning researchers and philosophers of science.

Many theoretical proposals conceptualize the acquisition of deep knowledge as a deliberate, effortful and constructive process. For example, one frequently stated hypothesis with roots in both philosophy (Popper, 1972/1959) and psychology (Thorndike, 1898) claims that learners replace or revise their knowledge when the latter is *falsified* by contradictory information; on this view, deep learning is driven by the evaluation of evidence (Gopnick & Meltzoff, 1997; Posner, Strike, Hewson & Gertzog, 1982). The hypothesis of *analogical learning* (e.g., Holyoak & Thagard, 1995) claims that the learner retrieves a possible analog to his or her current problem from memory and discovers their shared structure by constructing a mapping between them. According to the idea of *representational redescription* (Karmiloff-Smith, 1992), the learner reflects on his or her knowledge and, as a consequence, generates a higher-order representation of it. Many other proposed learning mechanisms share this active character (Ram & Leake, 1995).

In contrast, research on implicit learning of artificial grammars (Reber, 1989, 1993) suggests that learning is a

passive, inductive process which is independent of any intention to learn and which creates knowledge that cannot be deliberately recalled. In the *training phase* of the standard artificial grammar learning paradigm, the participants memorize letter strings, one by one. The strings have been generated with an artificial grammar and hence embody some very abstract properties, but the participants are not informed of this fact. In the *test phase*, the participants encounter new letter strings which are derivable from the relevant grammar, mixed with distractors which are not. The task is to decide whether the test strings are of the same type as the strings seen during the training phase. A large body of evidence (Stadler & Frensch, 1998) shows that people perform better than chance in the test phase, indicating that they have acquired knowledge of the underlying grammar.

Servan-Schreiber and Anderson (1990) and Perruchet and Gallego (1997) have attempted to explain artificial grammar learning in terms of the learning of substrings. However, Manza and Reber (1997) report a series of six experiments in which the strings encountered in the test phase were expressed in different letters than the strings encountered in the training phase. People perform better than chance in this condition as well, indicating that what is learned is abstract enough to transfer and hence does not consist of knowledge about the relative frequencies of individual substrings. In short, the data imply that what is learned in the artificial grammar learning paradigm is an abstract representation of the relevant grammar.

This finding is counterintuitive, because string memorization is passive, incidental and purely inductive and so stands in contrast to the constructive learning mechanisms hypothesized in other areas of cognitive research. This leads us to inquire into the nature of the knowledge generated by the string memorization procedure. How does that knowledge function in subsequent processing? Can it support problem solving, text comprehension and other higher-order cognitive processes?

To investigate this question, we revised the standard artificial grammar learning paradigm by replacing the string classification task typically used in the test phase with a letter sequence extrapolation problem. Sequence extrapolation problems were first introduced into psychology by Louis L.

Thurstone and they have been studied from a cognitive perspective by Simon (1972), Greeno and Simon (1974), and Kotovsky and Simon (1973). In this type of problem, the problem solver is given a sequence of letters generated in accordance with some pattern and asked to extrapolate it. To solve the problem, he or she must thus first uncover the pattern in the given segment of the letter sequence and then use that pattern to generate the next N letters in the sequence.

The goal of the present study was to determine whether implicit learning of the pattern embedded in a sequence improves the ability to extrapolate that sequence. In the training phase, our participants memorized strings of double-digit numbers generated in accordance with a pattern. In the test phase, they tried to extrapolate a letter or number sequence that followed that same pattern. If string memorization produces an abstract and generative representation of the pattern underlying the strings and if people can access that representation during problem solving, string memorization should improve performance on sequence extrapolation.

To investigate the levels of abstraction we instantiated the extrapolation tasks in both numbers (near transfer) and letters (far transfer). If the knowledge generated from string memorization is encoded in terms of the surface features of the training strings, then that knowledge should not be available for problem solving. In contrast, if the knowledge gained during training is of limited abstraction, then it should be available to solve number problems (near transfer) but not letter problems (far transfer). Finally, if the knowledge gained is completely abstract it should be available to solve both number and letter problems.

## Method

**Participants** Ninety-eight undergraduate students from the University of Illinois at Chicago participated in return for course credit.

**Materials** The target tasks were two sequence extrapolation problems with a periodicity of six items. The target tasks were instantiated in both numbers (near transfer) and letters (far transfer); see Table 1. To enable the participants to induce the pattern, the given segments were 12 items long. That is, they covered two complete iterations of the underlying pattern. Problems were created specifically for this study with patterns similar to those used by Simon (1972) and Kotovsky and Simon (1973).

For example, pattern 1 in Table 1 can be described as follows: The pattern consists of two groups of two letters, separated by X and ending with Z. Within the first group of two, the second letter is two steps forward in the alphabet from the first. In the second group of two, the first letter is one step forwards from the last letter in the first group, and the second letter is one step backwards from that same letter. The second period has the same internal structure but begins with the letter that is one step forward from the second letter in the first group of two in the previous period.

Table 1. Two sequence extrapolation problems expressed in both letters and numbers.

Symbol Type	Given letter or number sequence & the correct 8-step extrapolation
<i>Problem 1</i>	
Letter	B D X E C Z E G X H F Z H J X K I Z K M
Number	25 27 47 28 26 49 28 30 47 31 29 49 31 33 47 34 32 49 34 36
<i>Problem 2</i>	
Letter	C D B E A M D E C F B N E F D G C O F G
Number	63 64 62 65 61 73 64 65 63 66 62 74 65 66 64 67 63 75 66 67

There were 24 training strings consisting of 12 double-digit numbers, twelve for each problem. The twelve strings associated with a problem followed the same pattern as the given letter or number sequence; see Table 2 for examples. In addition, there were 24 strings of random double-digit numbers used in the control condition. Participants in both number and letter problem solving conditions received the same training.

Table 2. Two training strings for Problem 1.

Example	String
1	13 15 35 16 14 37 16 18 35 19 17 37
2	59 61 81 62 60 83 62 64 81 65 63 83

Each participant received a booklet with two parts. Within each part, there were twelve sheets presenting the strings to be memorized, twelve blank recall sheets, one sheet for assessing the result of the training, one sheet presenting the sequence extrapolation problem, and one blank sheet to assess the participants knowledge of the pattern. Problems were counterbalanced across all conditions.

**Design and procedure** The participants were randomly assigned to one of four groups created by pairing training (patterned vs. random) with problem-type (letter vs. number): patterned near ( $n = 26$ ), patterned far ( $n = 27$ ), random near ( $n = 21$ ), and random far ( $n = 24$ ). In the patterned training groups, the participants memorized the strings that conformed to the same patterns as those in the extrapolation problems; see Table 2 for examples. In the

random groups, the participants memorized random number sequences. In the near transfer groups, the target problems were number extrapolation problems; see Table 1. In the far transfer groups, the target problems were letter extrapolation problems; see Table 1.

The participants were tested in groups of 25. The procedure consisted of two *cycles*. Each cycle was composed of training followed by problem solving. The participants memorized and recalled twelve strings, one by one. They were given 60 seconds to memorize each string. They were then told to turn the page and write down the string. This procedure was repeated through the twelve training strings. Next, the participants were told to turn the page and solve the sequence extrapolation problem. They were given 5 minutes to solve the problem. They were then asked to turn the page and describe the pattern in the extrapolation sequence as best they could. The second cycle proceeded in the same way. The procedure took approximately 70 minutes.

## Results

**Training** The first question is whether the participants in the patterned training group extracted the pattern embedded in the patterned training strings. If they did, they should perform better on the memorization task than the participants in the random training group. Knowledge of the pattern can be used to reconstruct the number sequence so it should improve recall performance.

The *memory score* for each participant was the number of double-digit numbers correctly recalled in the memorization task. Because there were 12 numbers to memorize, the memory score varied between 0 and 12. Mean memory scores for both patterned and random groups for each pattern are presented in Figure 1.

A 2 (training, patterned vs. random) by 2 (pattern-type, 1vs. 2) mixed analysis of variance revealed that there were main effects for both training and pattern-type. The patterned training group performed significantly better than the random training group,  $F(1, 96) = 88.28$ ,  $MSE = 8.68$ ,  $p < .05$ , indicating that the former benefited from the patterns embedded in the training sequences. As Figure 1 shows, this effect is present for each training pattern. There was also a main effect of pattern-type,  $F(1, 96) = 4.25$ ,  $MSE = 1.48$ ,  $p < .05$ , indicating that pattern 2 was easier to detect than pattern 1. Finally, type of training interacted significantly with pattern-type,  $F(1, 96) = 5.38$ ,  $MSE = 1.48$ ,  $p < .05$ , indicating that the advantage of the patterned training group was larger for pattern 2 than for pattern 1.

In summary, the data show that the patterned training group performed better on the string memorization task than the random training group. We infer that the participants in the patterned group learned the pattern embedded in the relevant training strings. It is noteworthy that the memorization strings did not share any substrings. Hence, this result contradicts that predicted by the substring hypothesis (e.g., Perruchet & Gallego, 1997).

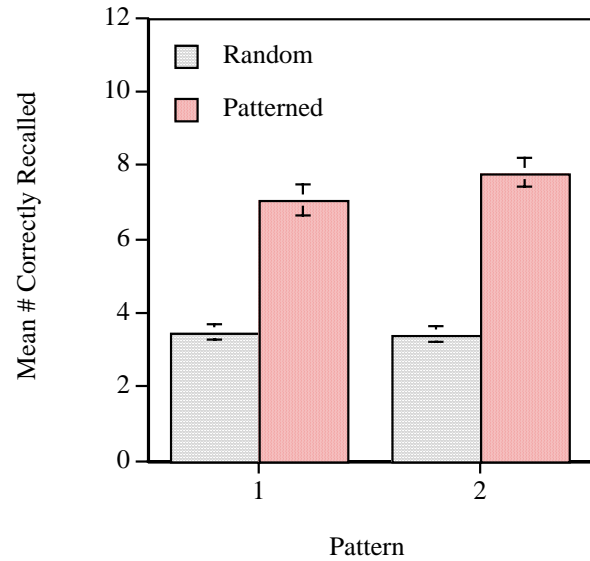


Figure 1. Mean memory scores for both training groups on pattern 1 and 2.

**Problem-solving** The second question is whether the relevant training group performed better on the problem solving tasks. The *problem solving score* was the number of letters or numbers correctly extrapolated in each problem solving task. Because the participants were asked to continue the sequence to eight places their problem solving scores varied between 0 and 8. Figure 2 shows the mean problem solving scores for both patterned and random groups on each problem.

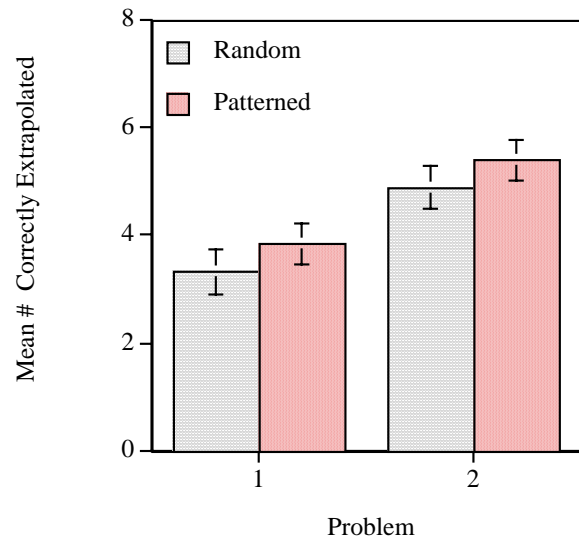


Figure 2. Mean problem solving scores for patterned and random groups on problems 1 and 2.

Although the patterned training group performed better than the random group on each problem, a 2 (treatment, patterned vs. random) by 2 (problem-type, 1 vs. 2) by 2 (transfer, near vs. far) mixed analysis of variance revealed no main effect for treatment condition,  $F(1, 94) = 1.03$ ,  $MSE = 15.33$ , *ns*. However, there was a main effect of transfer,  $F(1, 94) = 10.42$ ,  $MSE = 15.33$ ,  $p < .05$ , indicating that participants in the near transfer groups performed significantly better than participants in the far transfer groups. There was also a main effect of problem-type,  $F(1, 94) = 16.99$ ,  $MSE = 6.34$ ,  $p < .05$ , indicating that problem 2 was easier than problem 1. This is consistent with the higher memory performance on pattern 2; see Figure 1.

In addition, the interaction of treatment by transfer was marginally significant,  $F(1, 94) = 3.94$ ,  $MSE = 15.33$ ,  $p = .05$ , indicating that the advantage for participants in the patterned group was larger on near transfer problems than on far transfer problems. Figure 3 shows the mean problem solving scores for both patterned and random groups as a function of transfer. Main comparisons show that the patterned group performed significantly better than the random group on near transfer problems but not on far transfer problems,  $F(1, 94) = 6.41$ ,  $p < .05$ , and  $F(1, 94) = .87$ , *ns* respectively. These results show that participants in the patterned group only benefited from training when solving near transfer problems.

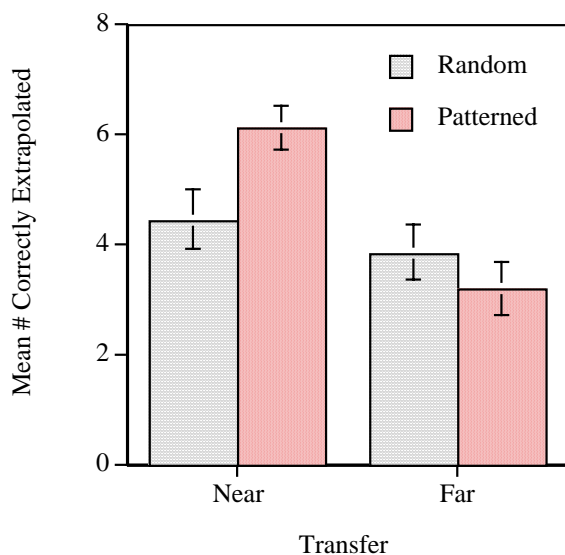


Figure 3. Mean problem-solving score for patterned and random groups on near and far transfer problems.

**Individual differences** To further investigate the relationship between string memorization and problem solving, we compared mean memory performance for each position in the sequence to the number of participants who correctly solved that position in problem solving. Participants were classified as either high or low memory

based on a median split of the memory scores for both patterned and random training. Median splits were calculated at each position of the pattern and the number of participants to correctly extrapolate each position was recorded. Table 3 shows the average number of participants to solve any given position correctly for both patterned and random groups as a function of memory.

Table 3. Percentage of subjects to solve any given problem position correctly

Memory Performance	Training Condition	
	Patterned	Random
Low	39%	56%
High	79%*	48%

In addition, chi square tests were calculated at each position of the problem to compare the number of high memory participants to correctly solve a particular position to the number of low memory participants to correctly solve that position. Chi square tests revealed that for patterned training, significantly more high memory participants solved corresponding extrapolations than low memory participants,  $\chi^2(1, N = 53) = 10.43$ ,  $p < .05$ . Chi square tests also showed that high-low memory groups with random training did not significantly differ in problem solving performance,  $\chi^2(1, N = 45) = .20$ , *ns*.

Similar position by position analyses were conducted comparing participants who solved near transfer problems to those who solved far transfer problems for both training groups. Table 4 shows the average number of participants to solve any given position correctly for both patterned and random groups as a function of transfer.

Table 4. Percentage of subjects to solve any given problem position correctly

Problem	Training Condition	
	Patterned	Random
Near transfer	76%*	55%
Far transfer	41%	48%

Chi square tests revealed that for patterned training, significantly more participants solved near transfer (number) problems than far transfer (letter) problems,  $\chi^2(1, N = 53) = 5.84$ ,  $p < .05$ . Chi square tests also revealed that participants in the random group did not significantly differ when solving near and far transfer problems,  $\chi^2(1, N = 45) = .64$ , *ns*.

## Discussion

As expected, the patterned training group performed significantly better than the random training group on the memorization task. The number strings were equivalent in the two conditions except for the fact that the strings memorized by the patterned group contained a pattern, while

the strings memorized by the control group were random. The higher performance of the patterned group is strong evidence that they acquired a schema for the underlying pattern during memorization. This replicates the common result found in implicit learning experiments (Berry, 1997; Reber, 1993; Stadler & Frensch, 1998).

The question asked here is whether the participants could apply this implicitly learned schema in deliberate problem solving. The patterned group was slightly better than the random group on the problem solving tasks, but the difference was small in magnitude (see Figure 2). However, significant differences appear when we take the type of problem and individual differences into account. There was a significant interaction between type of training and type of problem solved. The patterned group performed significantly better than the random group on near transfer problems but not on far transfer problems, indicating that the knowledge generated from the memorizing the pattern facilitated problem solving, but only when solving problems instantiated in surface features similar to those used in the training sequences.

This conclusion is also supported by the position-by-position analyses. Participants in the patterned training condition who performed above the median on the memorization tasks were consistently more likely to solve any one position during sequence extrapolation than those who performed below the median. This result was true for both problems 1 and 2 (see Table 3). In addition, the number of subjects who correctly solved any one position during extrapolation was consistently larger for participants in the patterned training condition who solved near transfer problems than for those who solved far transfer problems. Again, this result was true for both problems (see Table 4). No effect was observed in the random training conditions.

A plausible explanation for why the participants could not apply what they learned during training to the letter strings is that the relations in the patterns are less obvious on the alphabet than on numbers. For example, to solve pattern 1, the subject needs to realize that the letters E and C are the predecessor and successor, respectively, to D, a fact which is less obvious than the fact that the numbers 26 and 28 have those positions with respect to 27. This explanation implies that memorization of letter strings might produce different results. We are currently conducting studies to explore this implication.

In summary, our results are consistent with the hypothesis that what is acquired by memorizing patterned symbol sequences is a knowledge representation that is potentially generative but of limited abstraction. Such a representation might not be available for recall, conscious inspection or verbalization, but is nevertheless available to other high-level cognitive processes such as problem solving.

Although people do not go through life memorizing symbol strings, they do experience sequences, repetitions, and recurring events. Everyday tasks like starting a car has an intrinsic sequential structure: A person has to insert the

key before he or she can turn it; he or she must be inside the car in order to insert the key; he or she must open the door in order to get inside the car; and so on. In symbolic domains, sequential patterns of various kinds are perhaps even more prevalent. An example is the set of computer commands for accomplishing an elementary task such as a writing and sending an email message. Sequential patterns are consequences of the fundamental fact that actions have preconditions.

Given the importance and prevalence of sequential patterns, it is plausible that human beings have evolved cognitive mechanisms for identifying and encoding them. The output of this mechanism are what cognitive scientists often call schemas (Marshall, 1995). The data presented in this paper are consistent with the hypotheses that this mechanism operates even when the learner is not deliberately trying to extract a schema. We find this conclusion compatible with everyday experience: We doubt that human beings walk around and deliberately attempt to find patterns in experience; they find those patterns anyway.

If this conclusion is supported in future studies, the problem for cognitive theory is to elucidate the mechanism by which a schema that is not available for deliberate recall nevertheless influences problem solving, decision making, conceptual change and other cognitive processes. Hybrid models that combine symbolic representations with subsymbolic operations on activation levels (e.g., Anderson & Lebiere, 1998) seem the right kind of model, but the precise specification of such a model has to await replication and elaboration of the empirical observations reported in this paper.

## References

- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, New Jersey: Lawrence Erlbaum.
- Berry, D., (Ed.), (1997). *How implicit is implicit learning?* Oxford, UK: Oxford University Press.
- Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.
- Greeno, J. G., & Simon H. A. (1974). Processes for sequence production. *Psychological Review*, 81, 187-198.
- Holyoak, K., & Thagard, P. (1995). *Mental leaps*. Cambridge, MA: MIT Press.
- Karmiloff-Smith, A. (1992). *Beyond modularity*. Cambridge, MA: MIT Press.
- Kotovsky, K., & Simon, H. (1973). Empirical tests of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology*, 4, 399-424.
- Manza, L. & Reber, A. S. (1997) Representing artificial grammars: Transfer across stimulus forms and modalities. In D. Berry (Ed.), *How implicit is implicit learning?* (pp. 73-106). Oxford, UK: Oxford University Press.
- Marshall, S. (1995). *Schemas in problem solving*. Cambridge, UK: Cambridge University Press.



- Perruchet, P. & Gallego, J. (1997) A subjective unit formation account of implicit learning. In Berry, D. C., (Ed.), *How implicit is implicit learning?* (pp. 124-161). Oxford, UK: Oxford University Press.
- Popper, K. (1972/1959). *The logic of scientific discovery*. London, UK: Hutchinson. [Orig. *Logik der Forschung*, Vienna, 1935.]
- Posner, G., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, 66, 211-227.
- Ram, A., & Leake, D. (Eds.), (1995). *Goal-driven learning*. Cambridge, MA: MIT Press.
- Reber, S., R. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, 118, 219-235.
- Reber, A. S. (1993) *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. New York: Oxford University Press.
- Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 592-608.
- Simon, H. (1972). Complexity and the representation of patterned sequences of symbols. *Psychological Review*, 79, 369-382.
- Stadler, M., & Frensch, P. (Eds.), (1998). *Handbook of implicit learning*. Thousand Oaks, CA: SAGE.
- Thorndike, E. (1898). *Animal intelligence*. Unpublished doctoral dissertation. New York: Columbia University.

# Hypothesis-Testing Method in a Community of Psychologists

**Takeshi Okada (j46006a@nucc.cc.nagoya-u.ac.jp)**

School of Education, Nagoya University; Furo-cho, Chikusa-ku,  
Nagoya, 464-8601, JAPAN

**Takashi Shimokido (m47031a@nucc.cc.nagoya-u.ac.jp)**

School of Education, Nagoya University; Furo-cho, Chikusa-ku,  
Nagoya, 464-8601, JAPAN

## Abstract

Cognitive scientists have become increasingly interested in doing research on the nature of interdisciplinary collaboration. This new direction raises questions about the role that discipline specific practices may have when scientists from different disciplines collaborate. In this paper we explore one such discipline-specific belief: the importance of hypothesis testing in psychology research.

## Introduction

We interviewed scientists about their collaborative processes (Okada, Schunn, Crowley, Oshima, Miwa, Aoki, & Ishida 1995), and a computer scientist who had collaborated with a cognitive psychologist mentioned the following:

“The most important benefit of participating in this interdisciplinary collaboration is that there are scientists who have a different sense of value on science. For example, when working with researchers in the same discipline, we share a common ground and a common language. We can make progress in our project very quickly without wondering about what the co-researcher meant. Now, I think that my co-researcher in this interdisciplinary project and I didn’t share that common ground when we started our project. Therefore, we could not make any progress for about one year. We could not understand what confused us... The difference of disciplines related to the differences in the sense of value on science, methodology that we use, and more concretely speaking, evaluation criteria. Those differences made me feel this interdisciplinary collaboration was very interesting!” [Translated from Japanese]

We have also had similar experiences when working with computer scientists. It seemed that the computer scientists were more concerned with creating phenomena on a computer system, while we, as psychologists, were concerned with understanding phenomena in the real world, through experimental design, hypotheses, and manipulating variables.

These episodes suggest that scientists have beliefs about how science should proceed and be evaluated, and that

these beliefs are discipline-specific. These beliefs influence scientists’ research activities, such as conducting research projects, writing research articles, and advising students’ research projects. In this paper, we will focus on a discipline-specific belief about science that is widely shared in the psychology community. Recently, research interests in interdisciplinary collaboration have been growing in the cognitive science community (Derry & Gernsbacher, in press). An interdisciplinary collaboration, by nature, is an enterprise by collaborators with different beliefs from different backgrounds. Thus, it would be extremely important for us to know what kind of beliefs each discipline brings into the collaboration.

## Hypothesis-Testing as a Prescriptive Method

Scientific psychology has emphasized the importance of justification as a measure for being scientific. The hypothesis-testing style (i.e., entertaining clear hypotheses and testing them based on data) has been adopted as a prescriptive means for justification. When conducting scientific research, many psychologists believe that they must first develop clear hypotheses before testing them against the available data.

This hypothesis-testing style seems to be an offshoot of several dominant movements that emerged in Western psychology in the 1930s: logical positivism and operationism, hypothetico-deductive method, and inferential statistics.

Logical positivism aims to clarify the language of science and investigate the conditions under which empirical propositions are meaningful, then verify the propositions by means of a concrete procedure. This movement in the philosophy of science was introduced to the psychology community in the 1930s by Stevens (1939). Operationism (Bridgeman, 1928), which claims that a scientific concept should be defined by concrete operations to achieve the concept, was integrated into the logical positivism movement. These movements served as a strong theoretical background for the formation of scientific psychology.

The hypothetico-deductive method (H-D method) is a scientific method in which investigators are required to adopt a postulate tentatively and deduce its logical implications, and then check the validity of them by observation. Hull, a founder of neo-behaviorism, adopted it

as a core research method for his studies (Hull, 1943). As neo-behaviorism became dominant in psychology for the next several decades, Hull's emphasis of the H-D method had a strong influence in the psychology community.

With the development and introduction of inferential statistics to psychological research, investigators then had tools to implement logical positivism and the H-D method in their research (Fisher, 1935).

These movements had strong influences on the psychology community, the effects of which are still being felt today. In the rest of this paper, we will address the following questions related to the hypothesis-testing style of research in the community of psychology: How and when was such a belief formed in the psychological community in Japan? What kind of role does this belief about science play in shaping research activities?

The primary data are from the Japanese psychology community. However, we feel that this data reflects the situation of psychology in the Western community as well, since the Japanese psychology community has been strongly influenced by Western psychology, particularly by the USA. Moreover, Kerr (1998) found a similar pattern of researchers' beliefs on the hypothesis-testing style in the psychology community in the USA using a similar questionnaire survey with somewhat a different focus.

### Three Aspects of Research Activities

We will focus on three important aspects of scientific research: 1) Writing journal articles; 2) educating psychology students, and, most importantly; 3) conducting research projects.

#### On Writing Journal Articles

When submitting articles to psychology journals, authors sometimes receive comments that may have been motivated by the belief that research papers without hypotheses are unscientific. Following are examples of comments that our colleagues received from journal reviewers:

"The authors do not make any predictions or provide the foundation for predictions." (Cognitive Science)

"The most serious problem of this paper is that there is no clear hypothesis mentioned. ...You should predict what kind of result you would acquire and describe what the paper would contribute if the result is obtained." (Japanese Journal of Psychology) [Translated from Japanese]

In order to verify whether or not these examples reflect the current situation of the psychology community in Japan, we conducted a questionnaire survey of psychology researchers in 1998. Participants were first and second authors of articles published in the Japanese Journal of Psychology and the Japanese Journal of Educational Psychology over the previous year. Those two journals are bulletins of the two major scientific psychology societies in Japan. A questionnaire was mailed to 137 authors. We received replies from 111 authors—a response rate of

81.2%! The questionnaire included questions about the timeline of developing the hypotheses mentioned in each article and authors' past experiences of hypothesis formation in research activities. Each question will be described in detail in later sections of this paper.

Participants were asked if they had ever received reviewer comments that recommended revising the article to clarify the hypothesis: 25.7% of respondents answered yes. Considering the fact that this question only applies to authors who have previously submitted at least one paper to a journal without including any hypotheses, this rate should be regarded as higher than it appears. This suggests that the Japanese psychology community encourages researchers to write articles with clearly stated hypotheses. On the same issue, Kerr (1998) conducted a similar study, giving a questionnaire to 156 behavioral scientists in the USA. It asked them to estimate what percentage of publishable research articles should state an explicit hypothesis, according to journal editors and reviewers. Respondents thought journal editors and reviewers would say that research articles should state an explicit hypothesis about 80% of the time. Though this research did not focus on respondents' actual experience with reviewers, it does suggest that beliefs about the hypothesis-testing style in journal review processes are widely shared among psychologists, not only in Japan, but also in the USA.

In order to see how such journal review processes affect the style of journal publications, we coded the empirical articles (i.e., articles with data) in the 1997 volume in the Japanese Journal of Psychology (Okada & Shimokido, in press). If any hypotheses, predictions, or expectations were stated in an article, it was coded as an "article with hypothesis." Sometimes, hypotheses were clearly stated in the articles: "The hypothesis of this research is..." or "We have three hypotheses. The first one is..." Sometimes, the expression in an article was more subtle such as, "...was expected" or, "If it is true, this result would happen." We included all of them as "article with hypothesis" because, with this analysis, we wanted to capture how authors were influenced by the hypothesis-testing style of writing. Using this criterion for hypotheses, we divided the empirical articles into four categories. The first category is articles with no hypotheses mentioned. The second category is articles with hypotheses mentioned after the first experiment. The third category is articles with one or more hypotheses mentioned in the introductory section. The fourth category is the articles with two or more hypotheses mentioned in order to distinguish a correct one from wrong ones (i.e., a diagnosis test). The third and fourth categories were regarded as "articles with hypotheses."

The results showed that, in 1997, 58.8% of the empirical research articles in the Japanese Journal of Psychology had some kind of hypotheses written in the introductory section. Note that the other empirical articles, that didn't have any hypotheses, focused mainly on clinical case studies, testing the validity of a questionnaire, or psychophysics, which traditionally are types of articles written without hypotheses. Taking this into account, we can say that the hypothesis-testing style of writing articles

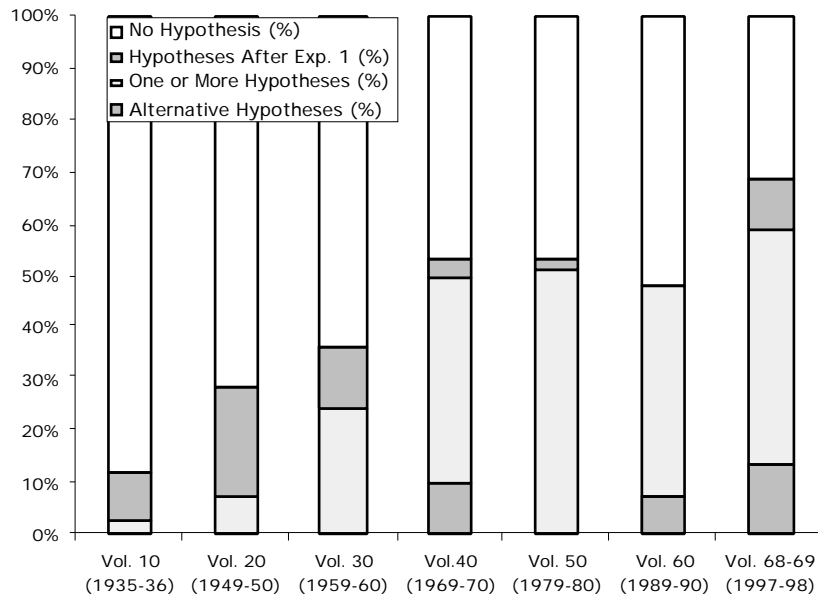


Figure 1: Hypotheses in the Japanese Journal of Psychology

is currently dominant in the Japanese community of psychology.

Do research articles in other disciplines follow the same hypothesis-testing style? We checked the 1996 volumes of Japanese science journals—these were the most recent volumes available in our university library at the time. We looked at the publication lists of the faculty members of each discipline in our university, then chose journals in which they were frequently published. We used the same coding scheme as the one previously mentioned, for the 1997 volume of the Japanese Journal of Psychology.

Table 1. Articles with hypotheses in various disciplines.

journal	Articles with hypotheses
Solid-state physics	0% had hypotheses
Earth Science	0% had hypotheses
Analytical chemistry	0% had hypotheses
Environmental medicine	0% had hypotheses
Neuroscience	0% had hypotheses
Ocean & Sky (Meteorology)	0% had hypotheses
Analytical Chemistry	2.8% had hypotheses
Polymer chemistry	24% had hypotheses

The main result of this analysis is shown in Table 1. As we can see, many research articles in other scientific disciplines do not follow the hypothesis-testing style of writing. Despite the belief about a scientific writing style that our psychology community shares, it seems that many scientists in natural science disciplines do not adhere to the hypothesis-testing style of writing. Are we willing to say that these articles without hypotheses are unscientific?

The next question that occurred to us was whether or not psychology articles have always used the hypothesis-testing style. If movements such as logical positivism, H-

D method, and inferential statistics had influenced research activities in psychology, the hypothesis-testing style of writing should have emerged at some point thereafter and spread throughout the psychology community. In order to answer this question, we conducted a historical analysis of the Japanese Journal of Psychology. It has been published since 1925, is the official journal for the Japanese Psychological Association, and is the oldest and most prestigious psychology journal in Japan. We coded the journal articles using the same coding scheme as previously described.

Figure 1 shows the results of the historical analysis of empirical articles in this journal. We examined every tenth volume of the journal from 1935 to 1998. (The most recent volume was Volume 69 at the time of this analysis.) As shown: 1) There were almost no articles with hypotheses in the introduction published prior to WWII; 2) the number of articles with hypotheses gradually increased after WWII; and, 3) the majority of the articles in the current volumes have hypotheses.

These results suggest that after WWII Japanese psychology researchers formed the standard that scientific articles should have hypotheses clearly stated in the introduction. This standard is quite different from that of the journals in other scientific disciplines. (Currently, we are conducting the similar analysis with American scientific journals so we can verify whether or not this trend is unique to the Japanese psychology community.)

The results of the historical analysis of the Japanese Journal of Psychology agree with the historical evidence regarding the import of the aforementioned movements—logical positivism, hypothetico-deductive method, and inferential statistics—from Western psychology to Japan. Although those movements emerged in Western psychology in 1930s and 1940s, World War II prevented Japanese psychologists from being exposed to

them. When World War II ended, these movements were introduced and gradually adopted into the Japanese psychology community. Theoretical articles on logical positivism and operationism appeared in the 1948 volume (the first volume after WWII) of the Japanese Journal of Psychology. Before WWII, there was very little research in behaviorism in Japan. After WWII, neo-behaviorism was introduced. For example, symposia on behaviorism were held at the 13th annual conference of the Japanese Psychological Association in 1949. Inferential statistics were introduced to Japan right after WWII. In the Japanese Journal of Psychology, the first theoretical article on inferential statistics appeared in the 1948 volume of the Japanese Journal of Psychology and the percentage of articles with inferential statistics increased dramatically during the 1950s (Omi, 1997).

From this evidence, we believe that it would be a fairly valid inference that a new belief about the hypothesis-testing style of writing articles among Japanese psychology researchers was influenced by logical positivism, hypothetico-deductive method, and/or inferential statistics.

### On Teaching How to do Research

During four years of teaching in a psychology department in Japan, the first author found that many psychology majors were taught that they must form clear hypotheses before collecting data. An undergraduate student in a research methods course complained to him, "Although I want to study this topic, I cannot come up with a clear hypothesis. So, I cannot study this topic." A graduate student writing a master thesis came to his office one day and confessed, "Though I conducted three experiments for my master thesis, I could only come up with a clear hypothesis in the last experiment. So, I may not have the ability to conduct scientific research."

In the process of learning about psychology, many students seem to acquire the idea that they have to form a clear hypothesis in order to conduct a psychological research study. In the questionnaire survey mentioned in the last section, we asked the following questions:

1. When you were a student, had you ever received advice from someone telling you that you should start a research project by developing clear hypotheses? *77.4% of respondents answered yes.*
2. When writing papers, had you ever received advice telling you that you should write clear hypotheses in the paper? *65.1% of respondents answered yes.*
3. Have you ever read a textbook on research methodology of psychology suggesting that you should start with clear hypotheses when conducting research? *70.8% of respondents answered yes.*
4. Have you ever given advice to someone telling him or her that when conducting research they should develop clear hypotheses before collecting data? *69.8% of respondents answered yes.*

Overall, the percentage of the respondents who answered yes to at least one of the above questions was

90.1%. Thus, the hypothesis-testing style seems to be the dominant practice in Japanese psychology.

The results of this survey are not surprising because many research methodology textbooks in psychology also mention that psychological research should proceed by finding questions and entertaining clear hypotheses first, then by collecting data. This is an excerpt from popular Japanese textbooks about research method in psychology.

*How to conduct research and write a paper*, Sirasa (1987): "The research process is the process of testing hypotheses... Thus, entertaining hypotheses is a very important first step to start research. If you think that you can discover something when conducting a survey or experiment with vague ideas, you will never succeed in your research."

The same trend was found in textbooks in the USA (Kerr, 1998). It seems that both in Japan and in the USA, psychology undergraduate and graduate students have been taught to use the hypothesis-testing style of research and to write articles following that style.

### On Conducting Research Projects

Our questionnaire survey was individualized for each respondent. We identified hypotheses in an article they had published and asked the authors specific questions about the hypotheses. If no hypothesis had been stated in their article, the same question was asked without identifying any specific hypothesis. The question was regarding whether they had developed the hypothesis written in their paper before they had collected the data. Respondents had to choose one of the following answers: a) The same hypothesis was entertained throughout; b) a different hypothesis was entertained; c) a vague hypothesis was entertained; d) no hypothesis was entertained; or (e) others.

In the case of articles with hypotheses, 70.6 % of the respondents said that the same hypotheses had been entertained throughout the study. However, 23.5% of the respondents admitted that they had different hypotheses, vague hypotheses, or no hypotheses at all before collecting data. Thus, we found that even if there are hypotheses clearly written in journal articles, it does not necessarily mean that the authors used the hypothesis-testing style when conducting their research. That is, in some cases hypotheses may have been developed between data collection and the writing of the paper. When interpreting the data, we have to consider that this survey was addressed to authors who have successfully published articles in mainstream psychology journals in Japan. It is highly possible to imagine that many psychologists who conducted research without hypotheses either could not publish their work in those mainstream journals or did not have the courage to submit them.

These results tell us somewhat contradictory stories about psychologists' research activities. While researchers in psychology conduct research in diverse ways (i.e., sometimes starting with a hypothesis and sometimes without), when they write journal articles they often imply that they had conducted the hypothesis-testing style of

research. When researchers teach others how to conduct research, they strongly emphasize employing the hypothesis-testing style. Does this mean that those psychologists who could not come up with any hypothesis before collecting data are not practicing the “correct” method of scientific research? Is the hypothesis-testing style really the best and the most scientific method of conducting research?

### **Potential Problems with the Hypothesis-Testing Style of Research**

Many philosophers of science have pointed out that scientists are not necessarily using the hypothesis-testing style of research when conducting scientific research (e.g., Hanson, 1958). Scientific discovery processes have two main phases: discovery of an explanation and justification for it. The hypothesis-testing style of research is strongly related to the justification side of scientific discovery processes, but not as much to the discovery side. Therefore, it does not completely reflect the actual process of scientific discovery. For example, Hanson (1958) stated in his famous book, *Patterns of discovery*:

“Physicists do not start from hypotheses: They start from data. ...H-D accounts begin with the hypothesis as given. ...The H-D account describes what happens after the physicist has caught his hypothesis; but it might be argued that the ingenuity, tenacity, imagination and conceptual boldness which has marked physics since Galileo shows itself more clearly in hypothesis-catching than in the deductive elaboration of caught hypotheses.”

Like Hanson, it seems that the majority of philosophers of science abandoned the concept that the hypothesis-testing style of research was the ideal scientific method a long time ago. However, as we have described, many psychology researchers still believe that this method is the best (and sometimes the only) scientific method that psychology should follow.

### **Various Styles of Research in Science**

Some scientists have pointed out that they are actually conducting research and producing prominent findings using other research styles. For example, Herbert A. Simon (1991), one of the founders of the fields of cognitive science, artificial intelligence, and cognitive psychology, has written about his research style as follows:

“When I examine my other experimental research, I find to my embarrassment that this fundamental condition for sound experimentation is seldom met. What have I been up to? What can I possibly have learned from ill-designed experiments? The answer (it surprised me) is that you can test theoretical models without contrasting an experimental with a control condition. And apart from testing models, you can often make surprising observations that give you ideas for new or improved models...”

“Perhaps it is not our methodology that needs revising so much as the standard textbooks on methodology, which perversely warn us against running an experiment until precise hypotheses have been formulated and experimental and control conditions defined. Perhaps we need to add to the textbooks a chapter, or several chapters, describing how basic scientific discoveries can be made by observing the world intently, in the laboratory or outside it, with controls or without them, heavy with hypotheses or innocent of them.” (pp. 383-385.)

Simon (in press) describes a case study of Faraday and further argues that curiosity and careful observation, which often lead to surprising results, are centrally important values to the scientific enterprise.

It seems that, at least in some scientific disciplines, scientists conduct research without using the hypothesis-testing style. They form hypotheses after observing phenomena.

### **The Cognitive Psychology of Scientific Thinking**

This point is supported by further evidence from studies in cognitive psychology. In the field of cognitive psychology, there have been substantial numbers of studies focused on scientific discovery processes (e.g., Klahr & Dunbar, 1988; Okada & Simon, 1997; Schunn, 1995). These studies suggest that: 1) Subjects frequently design experiments without hypotheses; 2) the frequency with which subjects design experiments without hypotheses is higher at the beginning of research; 3) there are individual differences in whether people tend to design experiments without hypotheses (experimenters who start experiments without hypotheses versus theorists who start experiments with hypotheses). These results fit with our findings from the questionnaire survey. Together, they converge to tell us that there are various research methods and styles in science.

### **Advantages and Disadvantages of Scientific Styles**

Hypothesis-testing styles of research, which are based on the H-D method and strong inference (Platt, 1964: i.e., develop alternative hypotheses and devise a crucial experiment that excludes one or more of the hypotheses, then carry out the experiment so as to get a clean result), are probably useful when the research field has been well-developed or the research project has progressed up to the level that the researchers do not need to create any new paradigm or theory. Although the percentage of articles which used strong inference in psychology journals is not high (see Figure 1), Platt (1964) claimed that this scientific method is the most productive way to conduct scientific research.

However, we feel that his claim is probably too strong to generalize. In certain situations, the H-D method (especially strong inference) might not work well. For example, when the research field is not well formed yet, or the research project is at the starting stage, the hypothesis-testing style of research might force researchers to form a hypothesis prematurely. Toyoda (1998) pointed out that

even a study with a precise statistical analysis to distinguish rival hypotheses might only be able to distinguish the rival hypotheses that are located very close to each other in a highly complicated hypothesis space. Therefore, when there is no valid reason to form hypotheses with the currently available data and theory, there is a possibility that the researchers will focus on hypotheses that are far apart from the correct hypothesis. In such a case, they might be stuck with irrelevant questions or irrelevant hypotheses that might not lead to any major discovery.

### Conclusion

In the historical and social context of Japanese psychology, many psychology researchers in Japan acquired the belief that the hypothesis-testing style was the best, and sometimes the only, scientific way. Such a belief creates a cognitive constraint (Miyake & Hatano, 1991; Siegler & Crowley, 1994) on the way that psychology researchers participate in research activities such as conducting research, writing research articles, and teaching research methods. Such a belief, on one hand, has a positive effect in enhancing effective research activities—many research articles have been published using this hypothesis-testing style. However, on the other hand, there could be situations in which such a belief has negative effects on research activities. As we have shown above, it was suggested that some articles without hypotheses have been rejected by journal reviewers as non-scientific even though such articles might have made a great contribution to the community of psychology, had they been published. It was also suggested that such a belief shaped types of research procedures that might have distorted researchers' views on scientific discovery. As for the educational aspect, it was suggested that some of the psychology students felt discouraged to explore new research directions because they received advice emphasizing the hypothesis-testing style of research.

We believe that the information from these analyses about researchers' beliefs in psychology would be useful when we try to understand cognitive processes among psychologists and other scientists in an interdisciplinary collaboration.

### Acknowledgments

Partially supported by the Japan Foundation, 1999, the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for encouragement for young scientists, 1997-1998, and the grant of the Tokai Science Academy, 1999.

### References

- Bridgeman, P. W. (1928). *The logic of modern physics*. New York: Macmillan.
- Derry, S.J. & Gernsbacher, M.A. (Eds.) (in press). Problems and promises of interdisciplinary collaboration: Perspectives from cognitive science. Mahwah, NJ: Erlbaum.
- Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, 17, 397-434.
- Fisher, R. A. (1935). *The design of experiments*. London, England: Oliver & Boyd.
- Hanson, N. R. (1958). *Patterns of discovery*. Cambridge, MA: Cambridge University Press.
- Hull, C. L. (1943). *Principles of behavior*. New York: Appleton-Century-Crofts.
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2, 196-217.
- Klahr, D. & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, 12, 1-48.
- Miyake, N. & Hatano, G. (1991). Nichijoteki ninchi katsudo no shakaiteki bunkateki seiyaku [Socio-cultural constraints and beyond]. In Japanese Cognitive Science Society (Ed.), *Ninchi Kagaku no hatten*, 4. Tokyo: Kodansha.
- Okada, T., Schunn, C. D., Crowley, K., Oshima, J., Miwa, K., Aoki, T. & Ishida, Y. (1995, June). *Collaborative scientific research: Analyses of historical and interview data*. Paper presented at the 12th Annual Conference of the Japanese Cognitive Science Society.
- Okada, T. & Shimokido, T. (in press). The role of hypothesis formation in a community of psychology. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implication from everyday, classroom, and professional settings*. Mahwah, NJ: Erlbaum.
- Okada, T. & Simon, H. A., (1997). Collaborative discovery in a scientific domain. *Cognitive Science*, 21, 2, 109-146.
- Omi, Y. (1997). Kenkyuhou no hensen [Change of research methods]. In T. Sato & H. Mizoguchi (Eds.), *Tsushi Nihon no shinrigaku*. Kyoto, Japan: Kitaoji-Shobo.
- Platt, J. R. (1964). Strong inference. *Science*, 146, 347-353.
- Schunn, C. D. (1995). A goal/effect trade-off theory of experiment space search. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA, USA.
- Shirasa, T. (1987). *Kenkyu no susumekata matomekata* [How to conduct research and write a paper]. Tokyo: Kawashima-Shoten.
- Siegler, R. S. & Crowley, K. (1994). Constraints on learning in nonprivileged domains. *Cognitive Psychology*, 27, 194-226.
- Simon, H. A. (1991). *Models of my life*. New York: Basic Books.
- Simon, H. A. (in press). "Seek and ye shall find": How curiosity engenders discovery. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implication from everyday, classroom, and professional settings*. Mahwah, NJ: Erlbaum.
- Stevens, S. S. (1939). Psychology and the science of science. *Psychological Bulletin*, 36, 221-263.
- Toyoda, H. (1998). *Kyoubunsankouzobunseki nyumonhen: Kozohoteishiki modeling* [Introduction to Covariance Structure Analysis: Structural Equations Modeling]. Tokyo: Asakura-Shoten.

# Fast and Frugal Use of Cue Direction in States of Limited Knowledge

**Magnus Persson (Magnus.Persson@psyk.uu.se)**

Department of Psychology, Uppsala University  
Box 1225, SE-751 42, Uppsala, Sweden

**Peter Juslin (Peter.Juslin@psy.umu.se)**

Department of Psychology, Umeå University  
SE-901 87, Umeå, Sweden

## Abstract

An exemplar-based algorithm, PROBEX, (Juslin & Persson, 1999) is shown to make robust decisions in multiple-cue inference tasks when very few exemplars are known. We demonstrate the crucial role of knowledge of cue directions for performance and confront PROBEX with an artificial environment specifically construed to favor a non-compensatory algorithm like Take The Best (Gigerenzer & Goldstein, 1996). PROBEX is demonstrated to perform well even in these unfavorable conditions. The explanation for the robust performance is that PROBEX approximates Dawes Rule by using information of the cue directions for all cues, while yet making few a priori assumptions about the structure of the environment.

## Introduction

In this paper we will discuss the importance of knowledge of cue directions for probabilistic inference. The cue direction refers to the valence of the relationship between two variables, for example, as represented by the sign of a correlation, or a coefficient in a linear equation (e.g., a beta weight in linear regression). We will concentrate on a simple binary choice task format of the following sort, “Which German city has the higher population: a) Bonn, b) Hamburg?”. In this case, knowledge of a cue direction corresponds to knowing if a binary probability cue, say, that Hamburg, but not Bonn, has a soccer team in the Bundesliga, increases or decreases the probability that alternative (a) (Bonn) is the correct answer to the question.

The constraints on a plausible cognitive mechanism for learning cue-directions are complex and multifaceted. In some tasks, learning cue-directions is not important because they are known a priori. In those cases, it is the relative weight of the cues that is important. On the other hand, it is sometimes proposed that knowledge of cue directions is *the* crucial aspect of learning to make probabilistic inferences (Dawes & Corrigan, 1974). Moreover, inferences have to be made for new and unexpected tasks for which little previous experience is available. This means that the system cannot rely extensively on pre-computed knowledge—which cues are predictive, and the direction of predictive validity, has to be detected on the spot. Finally, a flexible algorithm should map both linear and nonlinear aspects of an environment.

These constraints boil down to whether we can find an algorithm integrating the ability to represent non-linearity with

on-the-spot detection of cue directionality. Is this possible without violating psychological plausibility? In this paper, we show that PROBEX (PROBABILITIES from EXemplars: Juslin & Persson, 1999) is such an algorithm. It relies on similarity-guided retrieval of exemplars to capture nonlinear relationships and to estimate cue directions. An added bonus is that because PROBEX belongs to the class of *lazy algorithms* (Aha, 1997), which do not require pre-computed knowledge, this is achieved in a fast and frugal fashion.

An important development in the judgment literature is the concern with evaluating cognitive algorithms within real environments (Gigerenzer & Goldstein, 1996; Gigerenzer, Todd, & the ABC-group, 1999). Gigerenzer and Goldstein (1996) demonstrate that, when applied to the structure of a real environment, simple heuristics that only rely on a single cue perform on par with complex algorithms that integrate multiple cues. Take-The-Best (TTB) relying on the single most valid cue that is applicable, performed as well as linear multiple regression that integrates 9 cues. It was concluded that although TTB falls short of classical norms of rationality, it provides the same accuracy at a minimum of computation: It is “fast and frugal”. One shortcoming of the simulations in Gigerenzer and Goldstein was that all algorithms were provided with a priori knowledge of cue directions.

Connectionist, exemplar-based and decision-tree architectures have been shown to compete evenly with TTB in regard to accuracy (Chater et al, 1999). Specifically, in Juslin and Persson (1999) it was shown that PROBEX outperformed TTB and linear multiple regression in regard to accuracy while relying on no pre-computed knowledge. PROBEX further provided a good quantitative fit to the quantitative point-estimates, binary decisions, and probability judgments made by human participants (see Dougherty, Gettys, & Ogden, 1999, for a similar approach).

In this paper, we complement Juslin and Persson (1999) in three respects: First, we illustrate the crucial role of estimating cue directions for the performance of any algorithm. Second, we explore boundary conditions for the robust performance of PROBEX by exposing it to an environment deliberately construed to favor a non-compensatory algorithm like TTB and a linear, additive algorithm like linear multiple regression. Finally we present a simple demonstration to elucidate why evolution should favor decisions algorithms that are robust in states of limited knowledge.



## PROBEX—The Algorithm

Many theories in cognitive science stress the storage of exemplars (traces, instances) (e.g.; Kruschke, 1992; Logan, 1988; Medin & Schaffer, 1978; Nosofsky, 1984, Nosofsky & Palmeri, 1997). One property of exemplar-based models is that they describe algorithms that respond to both frequency and similarity. In this respect, they map onto well-known properties of human probability judgment (Juslin & Persson, 1999). PROBEX was developed from one of the well-known and successful exemplar-based model, the *context model* (Medin & Schaffer, 1978; Nosofsky, 1984). PROBEX amends the context model in the following respects (see Juslin & Persson, 1999): (a) With a sequential sampling mechanism that allows prediction of response times (as such PROBEX provides a humble cousin of the EBRW model presented by Nosofsky & Palmeri, 1997); (b) A dampening in order to predict pre-asymptotic performance (see also Nosofsky et al., 1992); and (c) response rules that allows prediction of point-estimates and subjective probability judgments. The simple stopping rule and the mechanisms for point-estimation and probability judgment are the main differences from previous exemplar-based models.

Knowledge of the environment is modeled by an  $R \times C$  matrix, with  $R$  exemplars,  $C$  cue dimensions and one vector with  $R$  target values. The exemplars in the knowledge matrix represent distinct psychological entities, either traces of dated events from episodic memory or semantic knowledge. Exemplars are described in terms of binary feature values, except for the continuous target dimension  $t$ . Each exemplar is represented by  $D$  binary features  $x_i = [x_{i1}, x_{i2}, \dots, x_{iD}]$ , where  $1$  denotes presence of the feature and  $0$  its absence. The participant is presented with a new exemplar  $\bar{t}$  and is required to make a judgment or a decision. The similarity between  $\bar{t}$  and stored exemplar  $y$  is computed by the multiplicative similarity rule of the context model,

$$S(\bar{t}, y) = \prod_{j=1}^D d_j, \quad d_j = \begin{cases} 1 & \text{if } t_j = y_j \\ s & \text{if } t_j \neq y_j \end{cases}, \quad (1)$$

where  $d_j$  is 1 if the values on a feature match and  $s$  if they mismatch. Similarity  $s$  is a parameter in the interval  $[0, 1]$  for the impact of mismatching features. For low values of  $s$ , the similarity is close to one only for an exemplar that is almost identical to the new exemplar, but for high values of  $s$  all of the stored exemplars are deemed very similar.

We examine PROBEX with  $s=0.5$  for all cue dimensions which is a compromise between these extremes. The idea is that this compromise, referred to as *similarity-graded probability*, is a particularly robust and efficient way to exploit states of limited knowledge (Juslin & Persson, 1999).

The stored exemplars race to determine the response (Logan, 1988; Nosofsky & Palmeri, 1997). The stored exemplars are retrieved one-by-one from an initial set  $K$  to yield a sequence  $x_1, x_2, \dots, x_N$ . The probability that exemplar  $y$  is the sampled exemplar  $x_n$  at iteration  $n$  is:

$$\forall y (y \in K_n), \quad P_n(x_n = y) = \frac{S(\bar{t}, y)}{\sum_{z \in K_n} S(\bar{t}, z)} \quad (2)$$

The summation in the denominator is performed across exemplars not yet sampled. A response is generated at iteration  $N$ , where the decision rule specified below terminates the sampling process.  $N$  is a random variable, the distribution of which can be used to predict response times.

To estimate the target value  $v'(\bar{t})$  of the new exemplar  $\bar{t}$ , the target values  $v(x_i)$  of the retrieved exemplars  $x_i$  are considered. The estimate of the target value at iteration  $n$  is,

$$v'(\bar{t}, n) = \frac{\sum_{i=1}^n S(\bar{t}, x_i) v(x_i)}{\sum_{i=1}^n S(\bar{t}, x_i)}, \quad (3)$$

a weighted average of the retrieved target values, where the similarities are weights. The final estimate is  $v'(\bar{t}) = v'(\bar{t}, N)$  where  $N$  is the first iteration where the conditions for the stopping rule are satisfied. Eq. 3 can also be produce probability assessments (Juslin & Persson, 1999).

The sampling of exemplars is terminated at the first iteration  $N$  where the following condition has been satisfied. The stopping rule is:

$$|v'(\bar{t}, n) - v'(\bar{t}, n-1)| < k \cdot |v'(\bar{t}, n)|. \quad (4)$$

The free parameter  $k$  decides the sensitivity of the stopping rule. One can interpret this rule as a way of judging when the change in the point estimate from  $v'(\bar{t}, n-1)$  to  $v'(\bar{t}, n)$  is too small to merit further sampling.

## Is PROBEX Frugal?

Gigerenzer and Goldstein (1996) proposed the notion of 'fast and frugal' as a conceptual threshold that a model of human decision making must reach in order to be plausible. At first sight, the mathematics of PROBEX that model retrieval of exemplars seems too complex to be "fast and frugal". But there are two major issues that resolve this dilemma. First, the complex part of PROBEX models quick and effortless memory processes that operate in parallel. Secondly, it is not enough to prove that an algorithm is efficient at the moment the decision is made, without concern for the requirements on pre-computation. PROBEX belongs to the class of *lazy algorithms* (Aha, 1997) in artificial intelligence, that avoid the processing of data before the task is given. On the other hand, all algorithms discussed in this paper, except PROBEX, rely on pre-computed representations. TTB (Gigerenzer & Goldstein, 1996) for example need to compute a sorted list of cue validities for the specific task, which in it self is computationally demanding. Arguably, if PROBEX makes good decisions by retrieving few exemplars and without using pre-computed representation, it is fast and frugal in a more general and important sense.

## The Ecological Rationality of Five Algorithms

**The German City-population Task** The task in the initial study on ecological rationality was the German city-population task (Gigerenzer & Goldstein, 1996; see Gigerenzer, et al., 1999, for applications to other environments). The task is to answer questions such as "Which city has the larger population: Heidelberg or Erlangen?" The decision process is modeled by an algorithm that relies on some strategy to make intelligent guesses about the populations of German cities. The simulation also requires an environment-model containing facts about German cities which—once known to the algorithm—can be used to infer city-populations. The environment is represented by nine binary cues that characterize each city, for example, whether a city is a state capital or not, whether it has a university or not, where the nine cues vary in predictive validity.

**The Algorithms** PROBEX was compared to four algorithms for guessing which of two objects have the largest value on the target dimension: (1) A linear multiple regression model with cues as independent variables and population as dependent variable. In a pair-wise comparison task, the algorithm decides on the city with the higher estimate. The direction of the cues is the sign of the regression weights. Linear multiple regression is included because it is routinely claimed to provide robust and accurate predictions. However, it cannot handle situations with few observations unless one amends it with a method such as Ridge Regression to compensate for cue dimensions with no information<sup>1</sup>.

(2) Dawes' Rule (Gigerenzer, et al., 1999): A heuristic version of the linear model that counts how many of the cues support each of the two cities and decides on the city implied by more cues. Cue direction is represented as 1 or -1 if there is positive or negative correlation in the training data respectively and zero if it was not computable. As detailed below, two versions of Dawes' Rule were implemented in order to investigate the importance of a priori knowledge of the cue directions. (3) TTB (Gigerenzer & Goldstein, 1996): In pair-comparisons, TTB decides on the task implied by the first most valid cue that differentiates between the pair.

(4) QUICKEST (Gigerenzer, et al., 1999) is an algorithm appropriate for skewed distributions like the German city-populations, where most cities have small populations. For each cue, the mean population for cities with negative cue-values is computed (negative cue values are those that go with small populations, e.g., not being a state capital). Cues are rank-ordered from the cue with lowest mean given a negative cue value to the cue with the highest mean given a

negative cue value<sup>2</sup>. To estimate a population the algorithm starts by checking if a city has the cue value that is first in this rank-order, then the next, and so on until a match is encountered. Then the mean population for the cities which have this cue is the estimate. Given the skew of the city-population distribution with mostly small cities, this algorithm is frugal in the sense of minimizing the number of cues that have to be accessed (i.e., for most cities the algorithm will stop for the first negative cue values in the rank-order). In pair-comparisons, QUICKEST decides on the city with the larger estimate.

Gigerenzer and Goldstein (1996) tested the algorithms by feeding them with all the pair-wise comparisons between German cities with more than 100 000 inhabitants. One weakness of this procedure was that the knowledge of the algorithms was assumed to consist of all the cities. A better test (Gigerenzer, et al., 1999) is to split the set of German cities into a training set and a test set. The training set is used to train the algorithm and the test set is the pool from which the test questions are constructed. This cross-validation is a true test of the robustness and detects any over-fitting to the training set. Moreover, it highlights the issue of learning the cue directions.

**Learning Cue Directions** In the simulations, all algorithms except A Priori Dawes' Rule have no a priori knowledge of the cues, but have to learn them from the known exemplars. In some tasks it, may be possible to infer cue direction by reasoning, but this topic is not addressed in this paper.

To illustrate the importance of knowing the cue directions, two versions of Dawes' Rule were implemented. The first version is the A Priori Dawes' Rule and it assumes that cue directions are known a priori. Its purpose is to show the theoretical upper limit of Dawes' Rule with a priori knowledge of cue directions. The second version, Dawes' Rule, relies on observed training exemplars to estimate cue directions by calculating whether each cue is positively or negatively correlated with the target dimension among the training exemplars. The difference between the variants indicates the importance of a priori knowledge of cue directions.

It is important to make special solutions for several of the algorithms below. With Dawes' Rule, insufficient data can make the correlation between a cue and the target variable undefined and then this cue is never used in the test phase. TTB must also be treated with care, because it use a sorted list of cue validities. When there are few exemplars it is often the case that several cue validities get the same numerical value and then these have to be listed in random order within the list. If this is not done the computer implementation can lead to a biased cue order which may either increase or decrease the accuracy of the algorithm.

---

<sup>1</sup> Ridge regression has the drawback of biasing the predictions towards the mean, and thus lowers the predictive accuracy when the weights are calculated from many observations without problems with correlated variables. We hand-picked an intermediate ridge constant, 0.1, that increased accuracy with limited information (small training set) but did not lower performance with much information (large training set).

---

<sup>2</sup> We did not implement QUICKEST exactly as Gigerenzer et al. (1999) did as we did not use approximations to natural numbers which probably implies that our implementation gives slightly better predictions.

## Study 1: Pair-Comparisons in the German City-Population Task

From an evolutionary perspective, it is important that an algorithm is good also when information is limited, because a decision maker has to survive as a "beginner" (see Study 3 below). Performance with small training sets is therefore a most important aspect of the robustness of an algorithm. Also, each algorithm soon reaches an asymptote that depends on both the cue structure and the algorithm itself. In the first simulation we thus compared the algorithms in the standard binary choice task with a particular eye to their performance in states of severely limited knowledge.

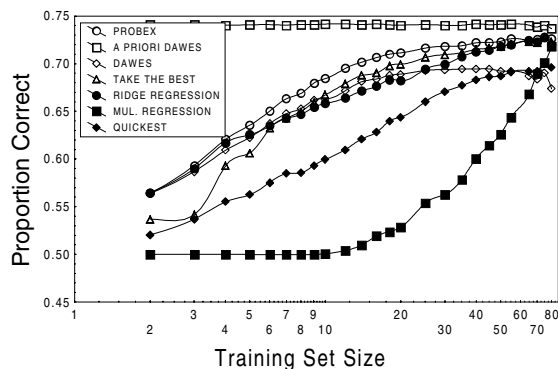


Figure 1. Accuracy and robustness of the algorithms. For each training set size, 1000 sets were randomized.

**Method** The dependent variable was proportion of correct inferences among the pair-comparisons of the test set. For each training set-size (2-80), as many as 1000 participants were simulated in order to make the errors of the means negligible. For each simulated participant, the German cities were randomly partitioned into training and test sets. Each algorithm was given the training set and the remaining cities were combined into all possible unique pairs and used as the test set. The data for the 83 German cities were collected from Gigerenzer and Goldstein (1996).

**Results and Discussion** Aside from A Priori Dawes' Rule that gets a head-start by virtue of its initial knowledge, Probex seems to be the winner in Figure 1. At minimum knowledge, Probex, Ridge Regression, and Dawes' Rule utilize the information equally well. TTB seems to have problems sorting the cues in a good order with little information, but does very well with 6 or more training exemplars. Quickest trade off accuracy for speed which is seen clearly in the cross validation paradigm. Note that A Priori Dawes' Rule does not define the asymptote that the other algorithms converge on, because it does not depend on the training set and does not suffer from over-fitting. This effect of cross-validation explains why it is constant at the high proportion correct of .74. It is surprising that Dawes' Rule performs worse with more information, but this is because the correlation between two cues in the full training

set is negative. With less information there is a greater chance to get a training set with only positive correlations. If cue validities had been calculated instead of correlations, Dawes' Rule would have done better.

In sum: The algorithms that use all information, such as Probex, perform robustly in states of limited knowledge. Probex, however, also performs best when more information is available. The superior performance of A Priori Dawes' Rule shows that a crucial aspect to be acquired by any algorithm is knowledge of the cue directions.

## Probex and Dawes Rule

In order to understand why Probex performs better than TTB with few training exemplars, it is instructive to consider only two optimal training exemplars: Big-city with all cues set to 1 and Small-city with all cues set to 0. For simplicity, all cues are assumed to be positively correlated with population. For Probex the similarity of the probe to Big-city decreases monotonically as a function of the number of cues in the probe that are not 1. The opposite holds for the similarity to Small-city which increases for each cue not set to 1. Because it is the number of cues set to 1 that differentiate the probes, Probex has the same high accuracy as Dawes' Rule.

TTB computes the cue direction for all cues but cannot apply this information, because the order of the cues is selected at random when the cue validity is the same for all cues. The data point with two training exemplars in Figure 1 suffers from this problem because the search order will be picked at random from those cues that have a well defined direction. Probex, Ridge Regression and Dawes' rule, on the other hand, integrate cue direction information from all cues in every decision in a similar way, which explains why they have identical proportions correct for the case of two training exemplars in Figure 1.

## Study 2: A Non-Compensatory Data Set

When is Probex outperformed by other algorithms, like linear multiple regression and TTB? A good guess is in an environment with additive and non-compensatory cues. The cues are non-compensatory if the optimal regression weights are such that the largest weight is bigger than the sum of the smaller weights. The second largest cue should likewise be larger than the sum all the remaining cues, and so on (Gigerenzer et al., 1999). An environment with linear, additive relations favors the regression model and a non-compensatory cue-structure favors TTB. A simple environment for which this is true can be defined by ordinary binary numbers. Each binary number can be used as a cue. For example, 5 is written as 00101 in binary numbers, and gives the cues  $c_1=0$ ,  $c_2=0$ ,  $c_3=1$ ,  $c_4=0$  and  $c_5=1$ . The optimal weights here are  $\{16, 8, 4, 2, 1\}$ , respectively (i.e.,  $16 \cdot 0 + 8 \cdot 0 + 4 \cdot 1 + 2 \cdot 0 + 1 \cdot 1 = 5$ ).

How much better than PROBEX will linear multiple regression and TTB perform in this environment specifically construed to fit the latter two algorithms?

**Method** The same procedure was used as in Study 1, except that the data were binary numbers between 0 and 32. Analogously with the German city-population task, the task was to guess which of two binary vectors has the highest number.

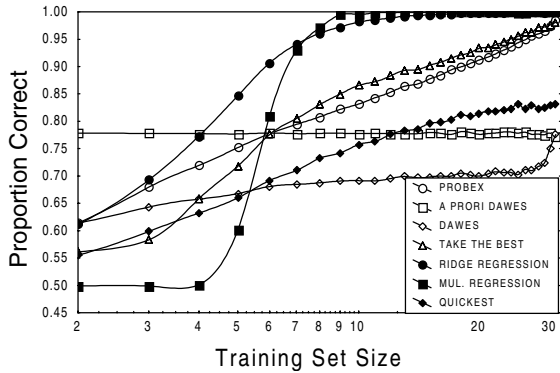


Figure 2. PROBEX in an artificial environment composed of the binary numbers.

**Results and Discussion** Ridge regression is clearly superior in Figure 2, but suffers slightly from biased estimates with 8-15 exemplars in the training set. TTB does not reach the asymptote unless it is trained with almost every exemplar, and is easily beaten by PROBEX when there are few training exemplars. The two variants of Dawes Rule cannot use the non-compensatory nature of the cues and are stuck at about the same level of performance as in Study 1.

As expected, multiple regression is perfectly suited for this environment, but both TTB and PROBEX converge on the same asymptote. More surprisingly, PROBEX is better than TTB with few exemplars. TTB only evens the score when more training exemplars are available, despite the fact that this cue structure is optimal for TTB. Thus, linear multiple regression converges more rapidly on the asymptote, but TTB enjoys no clear advantage over PROBEX.

### Study 3: An Unforgiving Environment

Does it matter if a decision algorithm is a few percentages better for decisions made in states of limited knowledge? The answer to this question, of course, depends on the consequences of these decisions. In this final section, we provide a simple demonstration that in an environment where poor decisions are fatal, and experience is only gained conditional on the survival of previous decisions, a small difference in decision quality may add up quickly. Arguably, these are the living conditions of many animals, including those of humans for a large portion of the evolutionary history.

**Method** In order to cover a wider range of possibilities and simplify the demonstration, we modeled three ideal deci-

sions strategies, *Early Learner* (EL), *Late Slow Learner* (LSL) and *Late Fast Learner* (LSF), as linear functions roughly similar to the functions in Figure 1 and 2. EL is PROBEX-like in the sense that it has a slight advantage early in the learning process. LSL and LSF are more TTB-like in that they start at a lower level but increases in accuracy, where LSF increases at a higher speed.

$$p_{EL}(\text{correct}) = 0.55 + 0.010 \cdot i \quad (5)$$

$$p_{LSL}(\text{correct}) = 0.50 + 0.015 \cdot i \quad (6)$$

$$p_{LFL}(\text{correct}) = 0.50 + 0.020 \cdot i \quad (7)$$

Equation 5, 6 and 7 define the probability of a correct decision if the decision maker has  $i$  training exemplars as guidance. Two simulations were made where EL was pitted once against LSL and once against LFL. Ten generations were simulated, where the relative proportion of surviving decision makers decided the relative proportion in the next generation. Each generation made 11 decisions, where  $i$  was varied from 0 to 10. For example, a member of the EL-species had a 0.55 chance to survive the first decision and 0.65 chance to survive the last. LSL and LFL both had to start off at 0.5, but ended with 0.65 and 0.7, respectively.

LSL is an example of a decision strategy that starts off poorly and barely catches up with EL. LFL, on the other hand, is better than EL for the last 5 trials. Note that, if the decisions were not fatal, LFL and EL would both make the same overall amount of correct decisions.

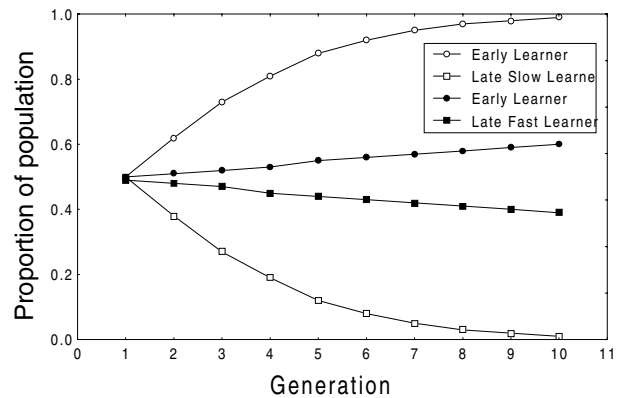


Figure 3: Two simulations of decision making in a unforgiving environment. Open symbols present Early Learners vs. Late Slow Learners, filled symbols present Early Learners vs. Late Fast Learners.

**Results and Discussion** Figure 3 presents the relative population proportions in a competition between EL and LSL, on the one hand, and EL and LFL, on the other. LSL would vanish very quickly in such a harsh environment. LFL are better but the losses in the beginning of every generation cannot be repaired despite the superior performance at the end of each generation (experienced decision makers).

This example is artificial and simplified, but shows that it is important to be a few percent better with little information than a few percent better with a lot of information, if learning is potentially dangerous. Indeed, the differences need not be large if they sum up over thousands of generations.

### General Discussion

We propose that a plausible model of the cognitive processes that underlie memory-based judgment and decision making should have at least three properties: *First*, the model should be consistent with—and preferably extend on—previous models with independent support in the cognitive science literature. PROBEX is a moderately modified version of one of the most successful models from the categorization literature—the context model (Medin & Schaffer, 1978; Nosofsky, 1984). *Second*, algorithm-details that pertain to implementation in judgment and decision making needs to be tested. First steps along these lines have been taken by fitting the predictions by PROBEX to empirical judgment data (Juslin & Persson, 1999).

*Third*, as implied by the research on ecological rationality (Gigerenzer et al., 1999), a cognitive algorithm should make sense also from an evolutionary perspective. An algorithm favored by natural selection should produce accurate judgments when applied to the constraints of a real environment, require a minimum of mental effort, and be robust in states of limited knowledge. In this paper we have scrutinized the ability of PROBEX to infer and use cue directions, in comparison with a number of fast-and-frugal algorithms discussed by Gigerenzer et al. (1999).

PROBEX provides a flexible and efficient way to compute and use directions of many cues on the spot, that is, without requiring any pre-computed knowledge. Moreover, TTB enjoys no systematic advantage over PROBEX in an environment specifically designed to favor a non-compensatory strategy. Importantly, in contrast to the other algorithms, PROBEX brings no strong commitment in regard to the presumed structure of the environment (e.g., linear, compensatory), but applies equally well to nonlinear environments (such as the classic X-OR-problem). Arguably, this is the kind of flexibility favored by evolution in adaptation to a complex and uncertain environment.

### Acknowledgments

The research reported in this paper was supported by the Swedish Council for Research in the Humanities and Social Sciences.

### References

Aha, D. W. (1997). *Lazy Learning*. Kluwer, Norwell, MA.  
Aha, D. W., & Goldstone, R. L. (1992). Concept learning and flexible weighting. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp.534-539). Mahwah, New Jersey: Lawrence Erlbaum.

Chater, N., Oaksford, M., Nakisa, R., Redington, M. (1999). *Fast, frugal and rational: How rational norms explain behavior*. Submitted for publication.  
Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making, *Psychological Bulletin*, *81*, 95-106.  
Dougherty, M. R. P., Gettys, C. F., Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*, 180-209.  
Gigerenzer, G. (1993). The bounded rationality of probabilistic mental models. In K. I. Manktelow, & D. E. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 129-161). London: Routledge.  
Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650-669.  
Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506-528.  
Gigerenzer, G., Todd, P., & the ABC Research Group (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.  
Juslin, P., & Persson, M. (1999). *PROBABILITIES from EXEMPLARS: On the role of similarity and frequency in probability judgment*. Submitted for publication.  
Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22-44.  
Logan, D. G. (1988). Towards an instance theory of automatization. *Psychological Review*, *95*, 492-527.  
Medin, D. L., & Schaffer, M. M. (1978). Context model of classification learning. *Psychological Review*, *85*, 207-238.  
Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104-114.  
Nosofsky, R. M., Kruschke, J., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 211-233.  
Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266-300.

# Effects of Presentation Format on Memory for Order

Margaret J. Peterson (mpeters2@gmu.edu)

Erik M. Altmann (altmann@gmu.edu)

Human Factors & Applied Cognition

George Mason University

Fairfax, VA 22030

## Abstract

Memory for order is important in everyday settings, for example in eyewitness testimony about who started a conflict. Although current theories claim that memory for order is different from memory for the to-be-remembered items themselves, many of the same manipulations that affect item memory also affect order memory. One manipulation that has shown effects in item memory, but that has not been tested in order memory, is presentation format. The hypothesis tested here is that order memory will be better for actions described as pictures than for actions described by text. The hypothesis is not supported by a main effect. However, presentation format does interact with serial position, suggesting that format effects are analogous to modality effects across auditory and visual presentation. Moreover, primacy is greater than recency, undermining Estes's (1997) perturbation model of memory for order.

## Introduction

Suppose that several witnesses to a knife fight agree that two knives were pulled, that a fight ensued, and that one person was killed. However, the witnesses' recollection of who drew a knife first is less exact. Although the to-be-remembered items are not in question, the recollection of order spells the difference between life in prison and acquittal on grounds of self-defense. If we can understand under what conditions people are most likely to be accurate about the order in which things happened, it could help us make better decisions about when to believe eyewitness and other accounts of history.

Understanding order memory also has important implications for our understanding of memory in general. Order memory is central to skill and language acquisition, both of which depend on sequence information. Also, order memory underlies the construct of episodic memory. For example, in some models of forgetting, people do not forget the word "bird", they only forget that "bird" was presented to them in a particular temporal window (Nairne, 1990b).

Memory for order is often claimed to be separate from memory for to-be-remembered items themselves (Bjork & Healy, 1974; Nairne, 1990a). Indeed, Whiteman, Nairne and Serra (1994) go so far as to say that order memory operates in a way that is fundamentally different from processes seen in both item recall and recognition. However, many manipulations that affect item memory have similar effects on order memory (Glenberg & Swanson, 1986; Nairne, 1990b; Naveh-Benjamin, 1990), even when the manipulation was designed to factor out the effects of item memory (Neath, 1997). For example, Glenberg and Swanson (1986) found that auditory presentation produces higher accuracy than vis-

ual presentation, especially for the last items in a list. This pronounced recency effect for auditory modality is similar to the effect seen in item memory. Similarly, Neath (1997) showed that set-size, presentation modality, and word concreteness all have similar effects on order memory as they do on item recognition and recall.

If order memory is similar to item memory, one might expect it to differ across other manipulations that produce differences in item memory. One such difference is between pictures and text. In recall and recognition tasks, pictorial presentation of items produces higher accuracy than verbal presentation (Baggett, 1979; Snodgrass, Wassner, Finkelshtein, & Goldberg, 1974). Dual-coding theory provides one possible explanation for these results. People usually encode pictures verbally as well as pictorially, but may or may not encode verbal material pictorially. This dual encoding of pictures creates a stronger memory trace, which leads to better item recall or recognition (Snodgrass et al., 1974).

Further evidence that memory for pictures is better than memory for text comes from Baggett (1979), who compared memory for a story across presentation formats. The story was presented either as a movie or as text, and was structurally equivalent in both formats. After subjects saw or read the story, they were asked to free-recall as much detail as they could about a specific episode, either immediately after study or after seven days. For those who saw the movie, still pictures were used to cue the beginning and end of the to-be-recalled episode. For those who read the story, textual cues were given. Although there was no effect of format in the immediate recall condition, participants in the seven-day condition recalled significantly more detail if they had seen the movie (Baggett, 1979).

This paper examines the effects of presentation format – pictures versus text – on accuracy of order memory. Our hypothesis is that memory for order should be better when items are presented as pictures. This hypothesis is primarily based on the findings cited above that suggest that item memory is better for pictures than for text. However, the hypothesis is also based on the intuition that order information is fundamental to how people function in the world. In particular, as we suggested in the knife-fight scenario above, order memory is a cornerstone of causal inference. Similarly, a picture is often worth a thousand words because it can make functional inferences easier to generate (Larkin & Simon, 1987). Thus, there is reason to ask whether order memory, given its basic nature, is facilitated by presentation formats more primitive than language.

## Experiment

The most common way to test order memory, while factoring out the effects of item memory, is to show participants a sequence of items at study time, and then show them the same items at test time in scrambled order. Their task is then to place the items back in their original order. Nairne (1992) claims that this type of task can be seen as a pure test of position memory without being confounded by item memory processes because all of the item information is made available at the time of recall.

The current study follows Nairne (1992) in testing incidental memory for order. In that study, subjects were asked to make pleasantness ratings on each word in five lists of five words each. The purpose of the rating task was to ensure that subjects would not be expecting any type of memory test and that the learning of order would be incidental. This type of deception is necessary because if participants are expecting any kind of memory test, even if they do not know that it will be a test of order, it is not truly incidental learning (Naveh-Benjamin, 1990). Tests of incidental memory for order have the highest ecological validity for assessing how order memory functions in everyday life.

Nairne's participants were brought back at time intervals ranging from 30 seconds to 24 hours and were then given a surprise test on the order of the items. The time delay variable is important to study not only to see the effects of decay over time, but also to be able to generalize any findings across short and long term memory. We adopted the three intervals (30 seconds, 4 hours and 24 hours) that Nairne cites as being the most representative of decay of order memory over time.

## Method

### Participants

The participants were 76 undergraduates at George Mason University who participated in the experiment for course credit in psychology classes. One subject's data were excluded due to failure to follow instructions. The experimental sessions were conducted in small groups ranging from four to fourteen participants each with four singletons.

### Materials and Design

The experiment consisted of 30 actions arranged into six thematic groups. The themes were primarily place-oriented; for example, things that might happen in an office setting, or buying items at a supermarket. Care was taken to ensure that actions had no logical or causal sequence of order (e.g. having to knead the dough before baking the bread). All actions in a group were presented either in picture format or in text format. Order of groups, and order of actions within groups, was determined randomly by sampling without replacement. The picture format consisted of a silent, color video segment that depicted an actor performing the target action. The actor's gender for each group was determined randomly before shooting and the same actor in the same clothing was used for each picture within a group. A text group consisted of a set of one-to-three-word phrases that described the actions in a corresponding picture group.

Black and white still shots representative of the video sequences and the corresponding text phrases are given in the Appendix.

Both text and picture actions were presented on videotape played on 27 to 35 inch standard televisions. Phrases were presented in large white block letters against a black background. Participants saw a video with three groups presented in picture format and three groups presented as text. Two different videos were used, each with a different mapping of groups to format. That is, if group was presented as text in one video, it was presented as pictures in the other, and vice versa. The purpose of this counterbalancing was to control for any interaction of format and the theme of the group.

Delay between presentation and test was manipulated between groups at levels of 30 seconds, 4 hours and 24 hours of delay. The within-subject manipulations of format and serial position combined with the between-subject manipulation of delay to produce a 2x5x3 mixed factorial design.

### Procedure

Participants were asked to watch the video and make pleasantness ratings about each picture or phrase they saw on a scale ranging from 1 (unpleasant) to 3 (pleasant). The rating task was used as a decoy to make intentional learning of the items or of the order unlikely. Each picture and phrase was visible for five seconds and was followed by 2.5 seconds of black screen. Between each group of five items, five seconds of blue screen was shown. Participants were not informed about the subsequent memory test, nor were they given any information as to why some of the items were presented as pictures and others as text. They were simply led to believe that they were participating in a rating task evaluating their affect toward everyday actions. Participants wrote their ratings on a response sheet containing six rows of five blanks; one row was designated for each group.

After completing the rating task, participants in the 4 and 24 hour delay groups were excused and instructed to return at the designated time for further rating exercises. Participants in the 30-second condition were asked to turn their rating sheets over and write down numbers counting backwards from 100 by threes. After thirty seconds, they were told to stop and their rating sheets were collected as test sheets were handed out. At the time of test, participants in all delay conditions were handed a response sheet that corresponded to the videotape that had been presented to them. Each test sheet consisted of six rows of five blanks labeled 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, 4<sup>th</sup> and 5<sup>th</sup> in that order. Above each row of blanks were the items that had been presented, but in a new random order. For groups presented as text, phrases were typed with the letters A through E to the left of each phrase. For groups presented as pictures, black and white still pictures representative of the action were presented and were labeled A through E directly above each picture. Participants were told that each group on the text sheet contained all five actions that were in that group at study. They were then asked to put the letter corresponding to each action in the correct blank to reconstruct the original order of presentation. Because all of the to-be-remembered actions were presented at test time, this type of free reconstruction task can be seen as a pure test of position or order memory

## Results

The measure we focus on in this analysis is the proportion of items placed correctly in their original order. These data are shown in Figure 1 (by format, aggregated over delay) and Figure 2 (by format and delay). All analysis of variance (ANOVA) was repeated measures on the format variable.

There was a significant effect of delay ( $F(2,73) = 19.4$ ,  $p < .0001$ ). However, there was no main effect of format ( $F(1,73) < 1$ ), nor was there an interaction between presentation format and delay ( $F(3,72) = 1.9$ ,  $p > .05$ ).

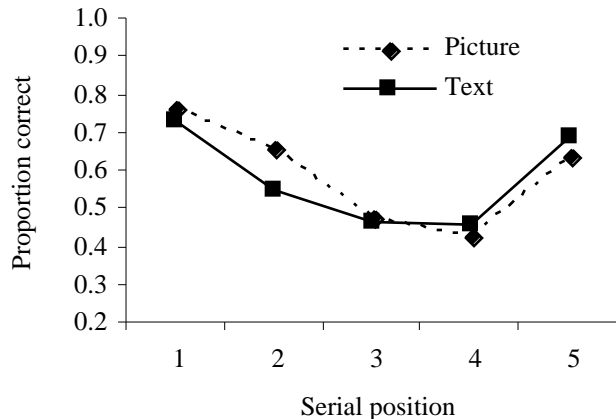
Characteristic primacy and recency effects are reflected in a main effect of position ( $F(3,72) = 41.0$ ,  $p < .0001$ ) and a significant quadratic trend ( $F(1,73) = 115.0$ ,  $p < .0001$ ). However, the linear trend was also significant, ( $F(1,73) = 21.2$ ,  $p < .0001$ ), and a post-hoc comparison of accuracy on the first and last items shows that primacy is greater than recency ( $t(75) = 3.0$ ,  $p < .005$ ). The linear trend and post hoc comparison were significant for five out of six combinations of format and delay, the exception being 24 hour text. There was no significant interaction between position and delay time ( $F(2,73) = 2.7$ ,  $p > .05$ ), indicating that the curves were roughly the same shape at each level of delay.

Although format did not have a main effect, it did interact with position ( $F(3,72) = 3.2$ ,  $p < .05$ ). As Figure 1 shows, the difference between primacy and recency is more pronounced in the picture format (that is, the serial position curve for pictures is rotated slightly clockwise compared to the curve for text). This interaction remains significant across all of the three time delays, as shown in Figure 2.

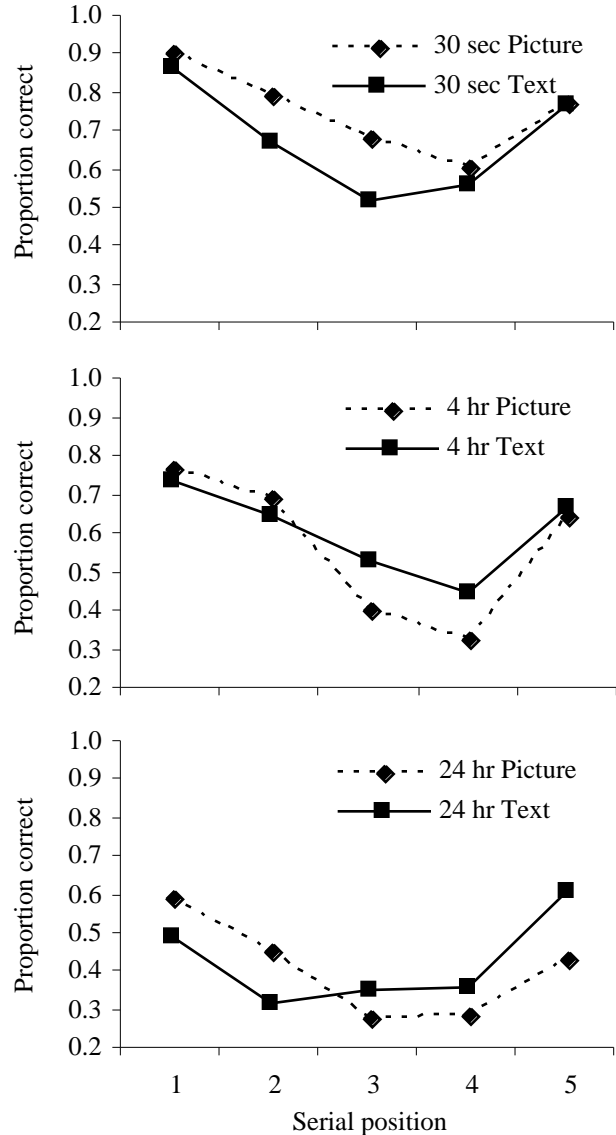
## Discussion

The hypothesis that pictures produce more accurate memory for order than text was not supported. This surprised us, given that pictures are consistently better than text in memory for items (e.g. Baggett, 1979; Snodgrass et al., 1974) and that many item-memory manipulations transfer to order memory (Naveh-Benjamin, 1990; Neath, 1997).

Our results point to a possible confound in how the effects of presentation format have been interpreted with respect to item memory. Another variable that may be correlated with format is whether sequential information is im-



**Figure 1:** Accuracy as a function of format and serial position.



**Figure 2:** Accuracy as a function of format and serial position, by delay (panels).

portant in understanding the stimulus. For example, Anderson (1976) found that order retention for linguistic materials (spoken words) was actually higher than for static pictures. Anderson's explanation for this superiority in his study depended on the sequential properties of the stimuli. In his view, language (in particular, the verbal stimuli in his study) is highly dependent on sequential information, and is therefore processed sequentially. In contrast, Anderson argued that the line drawings he used as pictures were not dependent on sequential processes for interpretation. He maintains that this sequential processing of linguistic material is robust enough to continue even when strings of non-related words are presented as stimuli. Thus, whether pictures are remembered better than words, or vice versa, may depend on the sequential structure (or other structure) of the stimuli across the two conditions.

The sequential structure of stimuli could easily be confounded with presentation format. For example, actions



presented as movies may be perceived to be more coherent than stills of the kind used by Anderson, because they fill in the details of natural action. If movies do communicate sequential structure more effectively, and if sequential structure facilitates memory, then our null effect is consistent with Anderson because the benefit of movies works against the benefit of text. However, the benefit of pictures in other studies (e.g., Baggett, 1979; Snodgrass et al., 1974) then becomes a puzzle. Our results suggest that these studies should be re-examined for other structural aspects of stimuli that may confound the effect of presentation format.

Another factor that may explain our null effect of format is that participants may use the same dual codes to represent both formats. An informal debriefing of participants after the experiment supports this view. Most of those questioned reported that they not only thought about the actions verbally while watching the video, they also visualized themselves doing the actions when they were presented textually. When asked why they had visualized the textual material, participants indicated it was because they needed to see it in their minds to be able to judge its pleasantness. Changing the distracter task used at study time (for example, asking for frequency counts rather than pleasantness ratings) may produce the asymmetrical recoding (pictorial to verbal) seen by Snodgrass et al. (1974). Thus the nature of the distracter task will have to be manipulated in future studies to isolate its effects on memory for order.

Despite the null effect of format, there was an interaction between format and serial position. A similar interaction has been observed between modality (visual and auditory) and position (Glenberg & Swanson, 1986; Neath, 1997), raising the possibility that these interactions are related. In the modality interaction, auditory presentation produces better order memory than visual presentation, but only for last one or two items (Neath, 1997). Early theories of this interaction implicated differences in sensory storage mechanisms across modalities. However, Gardiner and Gregg (1979) showed that the interaction was still pronounced when auditory distracter information was presented during the retention interval. These findings suggest that the modality interaction is caused by a variable that was confounded with echoic memory in earlier studies. The implication for our results is that the format interaction and the modality interaction may in fact stem from the same underlying process.

The interaction between format and serial position could also be due to a primacy benefit for pictures that is related to release from proactive interference. As we noted above, participants reported visualizing themselves performing the actions across all of the text groups. This may have made the text groups less distinct from one another than the picture groups, which each had a new actor and a new context. If a new picture group is more distinct by virtue of these visual cues, one would expect a stronger release from proactive inhibition and hence improved memory immediately afterwards. Thus the format interaction could stem from a recency benefit for text (similar to a modality effect), from a primacy benefit for pictures, or from some combination.

The difference between primacy and recency in our data helps to distinguish among formal models of order memory (see Brown, 1997, for a review of such models). The most

widely discussed is the perturbation model (e.g., Estes, 1997; Nairne, 1992). In this model, primacy and recency occur because there is only one direction in which the first and last items can “perturb”, namely toward the middle of the list. In contrast, middle items can drift in both directions, increasing the likelihood that they will be placed incorrectly at test. Importantly, the model predicts that primacy and recency should be symmetrical, because there is no difference between the two ends of the list. Although Estes (1972) presented this model as pertaining to short-term order memory, Nairne (1992) applied this model to the study of long-term, incidental learning for order memory, with fair results. What the model fails to predict, however, is the difference between primacy and recency effects, in which the first item in a group is placed more accurately than the last item. Nairne (1992) suggested that this difference may prove significant, and the support for this result found in our data raises a substantial problem for the perturbation model.

Two later models do make the prediction that primacy should be greater than recency in memory for order. The primacy model (Henson, Norris, Page, & Baddeley, 1996) is based on the ad hoc assumption that items earlier in a list have higher activation levels. These items are then suppressed by another ad hoc mechanism as they are output at test time. Recency arises from this model because there are fewer remaining choices nearer the end of the list, and therefore fewer chances to make an error based on noise in the activation levels. In addition to being largely ad hoc, this model makes the problematic assumption that participants place items in forward order at test time. However, informal debriefing of several participants in our study suggests that items are often placed in orders other than strictly forward. Indeed, others have observed a pattern in which participants initially place the first and last items, and only then place the middle items (Lee & Estes, 1977). Because sequential placement of order is a basic assumption of the primacy model, further study on the order in which participants actually place items is necessary.

The second model that predicts greater primacy than recency is the dual-code associative model (Altmann, in press). This model represents each item with two codes – one for the item itself and one for its location – and links them together in a chain at encoding time. Errors in linking codes produce order errors at test. Items at either end of a list have an advantage in that they can only be linked to an incorrect code in one direction. However, items at the start of the list have a greater advantage, because they suffer less interference at encoding time (Altmann, 2000). This model has the benefit of explaining how memory for order is encoded, an issue that the primacy model fails to address.

To advance our theoretical understanding of order memory, follow-up studies will have to differentiate among the primacy and dual-code associative models. In addition, existing models fail to account for modality, set-size, presentation format or any of the other effects that appear in studies of both item and order memory. Our goal should be to integrate these phenomena across item, order, short-term, long-term, semantic, episodic and all of the other pigeonholes in which we classify memory, into one unified theory.

## Conclusion

We expected that pictures would produce better memory for order than text, based on similar effects in memory for items and on considerations of the foundational nature of memory for order. This hypothesis was not supported, but we did find an interaction between format (picture vs. text) and serial position. Recency effects were greater for text across a range of retention intervals, an effect that may be related to the modality interaction observed by others. We also found that primacy was greater than recency in aggregate, and, in a finer-grain analysis, in almost all combinations of presentation format and time delay. This finding, foreshadowed by Nairne (1992), suggests that the perturbation model is incorrect and lends credence to models that capture directional processing of items at study.

## References

- Altmann, E. M. (in press). Memory in chains: A dual-code associative model of positional uncertainty. In N. Taatgen and J. Aasman (Eds.), *Proceedings of the 3<sup>rd</sup> international conference on cognitive modeling*.
- Altmann, E. M. (2000). The strength of chains: Why primacy dominates recency in memory for order. *Manuscript submitted for publication*.
- Anderson, R.E. (1976). Short-term retention of the where and when of pictures and words. *Journal of Experimental Psychology: General*, 105, 378-402.
- Baggett, P. (1979). Structurally equivalent stories in movies and text and the effect of the medium on recall. *Journal of Verbal Learning and Verbal Behavior*, 18, 33-356.
- Bjork, E. L., & Healy, A.F. (1974). Short-term order and item retention. *Journal of Verbal Learning and Verbal Behavior*, 13, 80-97.
- Brown, G. D. A. (1997). Formal models of memory for serial order: A review. In M. A. Conway (Ed.), *Cognitive models of memory*. Cambridge, MA: MIT Press.
- Estes, W.K. (1972). An associative basis for coding and association in memory. In A.W. Melton & E. Martin (Eds.), *Coding Processes in Human Memory*. Washington D.C.: V.H. Winston & Sons.
- Estes, W. K. (1997). Processes of memory loss, recovery, and distortion. *Psychological Review*, 104, 148-169.
- Gardiner, J.M., & Gregg, V.H. (1979). When auditory memory is not overwritten. *Journal of Verbal Learning and Verbal Behavior*, 18, 705-719.
- Glenberg, A.M., & Swanson, N.G. (1986). A temporal distinctiveness theory of recency and modality effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12, 3-15.
- Henson, R.N.A., Norris, D.G., Page, M.P.A., & Baddely, A.D. (1996). Unchained memory: Error patterns rule out chaining models of immediate serial recall. *The Quarterly Journal of Experimental Psychology*, 49A (1), 80-115.
- Larkin, J. H., & Simon, H. A. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11, 65-99.
- Lee, C. L., & Estes, W. K. (1977). Order and position in primary memory for letter strings. *Journal of Verbal Learning & Verbal Behavior*, 16(4), 395-418.
- Nairne, J.S. (1990a). A feature model of immediate memory. *Memory & Cognition*, 18(3), 251-269.
- Nairne, J.S. (1990b). Similarity and long-term memory for order. *Journal of Memory and Language*, 29, 733-746.
- Nairne, J.S. (1992). The loss of positional certainty in long-term memory. *Psychological Science*, 3, 199-202.
- Naveh-Benjamin, M. (1990). Coding of temporal order information: An automatic process? *Journal of Experimental Psychology: Learning, Memory and Cognition*, 16, 117-126.
- Neath, I. (1997). Modality, concreteness and set-size effects in a free reconstruction of order task. *Memory & Cognition*, 25(2), 256-263.
- Snodgrass, J.G., Wassner, B., Finkelstein, M., & Goldberg, L.B. (1974). On the fate of visual and verbal memory codes for pictures and words: Evidence for a dual-coding mechanism in recognition memory. *Journal of Verbal Learning and Verbal Behavior*, 13, 27-37.
- Whiteman, H.W., Nairne, J.S., & Serra, M. (1994). Recognition and recall-like processes in the long-term reconstruction of order. *Memory*, 2(3), 275-294.

## Appendix

Pictures (representative stills taken from video) and corresponding textual phrases for the six groups of actions.

### Group 1

#### Picture Presentation Format



#### Text Presentation Format

Wash Dishes  
Take Out Trash  
Vacuum Floors  
Make Bed  
Dust Cabinet

Group 2

Picture Presentation Format



Text Presentation Format

Buy Bread

Buy Potato Chips

Buy Soup

Buy Eggs

Buy Milk

Group 3

Picture Presentation Format



Text Presentation Format

Type Document

Talk On Phone

File Document

Copy Document

Dial Fax Machine

Group 4

Picture Presentation Format



Text Presentation Format

Rest In Bed

Take Temperature

Eat Soup

Drink Medicine

Sneeze

Group 5

Picture Presentation Format



Text Presentation Format

Hug Teddy Bear

Change Diaper

Drink Bottle

Read Story

Throw Ball

Group 6

Picture Presentation Format



Text Presentation Format

Highlight Text

Daydream

Take Notes

Study Flashcards

Read Text

# Teaching and Supporting the Use of Qualitative and Quantitative Concepts in Classical Mechanics

Rolf Ploetzner

ploetz@psychologie.uni-freiburg.de  
Department of Psychology; University of Freiburg  
D-79085 Freiburg, Germany

Siegward Beller

beller@psychologie.uni-freiburg.de  
Department of Psychology; University of Freiburg  
D-79085 Freiburg, Germany

## Abstract

Though very often quantitative problem solving is accentuated in physics instruction, psychological as well as educational research indicates that this emphasis is misleading. In an experimental study, we compared physics instruction with a focus on quantitative problem solving to physics instruction with a focus on qualitative problem solving. Initially, students were taught quantitative as well as qualitative concepts of classical mechanics by means of concept maps. Thereafter, the students attempted to solve four problems whose solutions demanded the coordinated application of knowledge about quantitative and qualitative concepts. During problem solving, the students received support from tutors. While one group of students was supported in qualitative problem solving, the other group was supported in quantitative problem solving. Before and after the problem solving, the students worked on tests. In accord with our expectations, students who were supported in qualitative problem solving improved significantly more from the pretest to the posttest than students who were supported in quantitative problem solving.

## Introduction

Very often, students are not able to successfully approach problems in classical mechanics by means of the knowledge they have acquired during physics instruction. Classical mechanics embodies concepts and relationships between concepts which allow for the description, explanation and prediction of motion. Many concepts and relationships between concepts involve qualitative as well as quantitative information.

Quantitative information is frequently expressed by means of laws which are formalized as algebraic or vector-algebraic equations. Students frequently approach problems which ask for a quantitative solution by only making use of their knowledge about quantitative information. Usually, they start from the variable whose value is in question. Afterwards, they attempt to apply dynamics and kinematics laws in order to determine the variable's value. Very often, however, the students get lost in a muddle of algebraic equations with no means at hand in order to guide their application effectively and efficiently (e.g., Chi, Feltovich & Glaser, 1981; Larkin, 1983).

In contrast to students, experts make use of both their knowledge about qualitative and their knowledge about quantitative information. Initially, they attempt to qualitatively identify the concepts relevant to the problem posed.

Subsequently, they take advantage of the qualitative information in order to select the appropriate dynamics and kinematics laws which quantitatively relate the identified concepts to each other (e.g., Chi, Feltovich & Glaser, 1981; Larkin, 1983). Finally, they apply the selected dynamics and kinematics laws in order to determine the value in question.

While experts seem to possess knowledge structures in which knowledge about qualitative and quantitative information is closely related, students' knowledge frequently is not only fragmentary and weakly related but also includes conceptualizations which are inconsistent with the concepts taught during physics instruction (cf. Pfundt & Duit, 1994). Due to these deficiencies, students seem not to be able to take advantage of their knowledge in the same way that experts do. As a consequence, students have to fall back on so-called weak problem solving methods such as operator subgoalings and means-ends analysis (cf. VanLehn, 1996). These methods, however, provide little guidance for solving problems in classical mechanics.

How can students be supported to acquire and to flexibly apply both knowledge about qualitative and quantitative information on classical mechanics? Though very often the emphasis in physics instruction is on quantitative problem solving, this emphasis seems to be misleading (e.g., Hestenes, 1987; Reif & Heller, 1982). Because very often successful quantitative problem solving presupposes qualitative understanding, physics instruction with an emphasis on qualitative problem solving might be more beneficial (e.g., Ploetzner, 1995; White, 1993).

In this paper we present an experimental study in which physics instruction with a focus on quantitative problem solving is compared to physics instruction with a focus on qualitative problem solving. Because psychological research (e.g., Chi, Feltovich & Glaser, 1981; Larkin, 1983), educational research (e.g., Hestenes, 1987; Reif & Heller, 1982) as well as research in artificial intelligence (e.g., de Kleer, 1977) indicate that successful quantitative problem solving presupposes qualitative understanding, we hypothesize that emphasizing qualitative problem solving is more effective than emphasizing quantitative problem solving.

## Knowledge about Qualitative and Quantitative Concepts in Classical Mechanics

The application domain is made up of textbook problems which refer to one-dimensional motion with constant accel-

eration. The knowledge investigated is on qualitative and quantitative information involved in concepts of dynamics (e.g., gravitational and normal force) and kinematics (e.g., displacement, velocity and acceleration).

With respect to qualitative information, the focus is on the conditions under which concepts are applicable, the attributes possessed by concepts and the values which concept attributes might have. For instance, knowledge about the kinetic friction force might comprise the qualitative information that a kinetic friction force acts on a body, whenever a normal force acts on the body and the body is moving on a surface which is not frictionless.

With respect to quantitative information, the emphasis is on dynamics and kinematics laws which are formalized as algebraic or vector-algebraic equations. For example, knowledge about the kinetic friction force might comprise the quantitative information that the magnitude  $F_f$  of the kinetic friction force on a body equals the magnitude  $F_N$  of the normal force on the same body times the coefficient of friction  $f$ :  $F_f = F_N \cdot f$ .

Qualitative and quantitative information can be conceptualized as complementary information (e.g., de Kleer, 1977). Qualitative information refers to essential features to be taken into account as well as to important distinctions to be drawn. While quantitative information frequently helps to resolve ambiguities inherently involved in qualitative information, the appropriate use of quantitative information very often seems to presuppose the utilization of qualitative information.

Ploetzner (1995) implemented formal representations of qualitative and quantitative information on classical mechanics in a simulation program. If the program is applied to the formal representation of a problem, it simulates how a qualitative problem representation can be taken advantage of to guide the construction of a quantitative problem representation. The program coordinates qualitative and quantitative problem representations in two different ways. Firstly, the information included in a qualitative problem representation is partially transformed into algebraic expressions in order to construct additionally required quantitative information. Secondly, the information contained in a qualitative problem representation is exploited to constrain the use of already available quantitative information.

## Method

### Design

The study comprised two groups of students and was made up of five sections.

In the first section, all students worked on an introduction to concept maps as well as on an introduction to a computerized concept mapping tool. In the second section, all students studied the same instructional unit which described qualitative and quantitative information on classical mechanics by means of concept maps. In the third section, all students worked on a multi-component test which assessed the knowledge about qualitative and quantitative information the students had acquired during the study of the instructional unit.

In the fourth section, the students attempted to solve four problems which demanded the coordinated use of knowledge about qualitative and quantitative information. During problem solving, the students took advantage of the computerized concept mapping tool. In addition, the students received support from tutors. While one group of students was supported in qualitative problem solving, the other group of students was supported in quantitative problem solving.

Finally, all students worked on a parallel multi-component test which assessed the knowledge about qualitative and quantitative information the students had acquired due to the support from tutors.

## Materials

**Introduction to Concept Maps** To be knowledgeable in a domain means to know the relevant concepts as well as the relationships between them. This structural aspect of knowledge can be represented by means of concept maps (e.g., Jonassen, Beissner & Yacci, 1993). Concept maps form an external representation in which information is structured by means of graphs. Individual nodes represent concepts; the directed and undirected links between the nodes represent relationships between the concepts. In an earlier study, Ploetzner, Fehse, Kneser and Spada (1999) demonstrated that concept maps can be equally well employed to teach qualitative as well as quantitative concepts in classical mechanics.

Because qualitative and quantitative information on classical mechanics were taught to the students by means of concept maps, in the first section of the study, the students worked on an introduction to concept maps in order to learn how concept maps are structured. The concepts addressed in the introduction referred not to classical mechanics but to well-known household furniture.

**Computerized Concept Mapping Tool** When concept maps are constructed by paper and pencil, they are frequently difficult to extend and to modify. Furthermore, the construction of concept maps can hardly be reconstructed by conventional observation methods. The use of a computerized concept mapping tool, however, allows one to overcome these drawbacks. Therefore, whenever the students had to construct concept maps, they took advantage of such a tool (cf. Ploetzner, Hoppe, Fehse, Nolte & Tewissen, 1996).

In a computerized concept mapping tool, the concepts and relationships relevant to the domain under scrutiny may be made available to the students in advance by means of menus, for example. If needed, the students may fill in additional concepts and relationships at run time. Complete concept maps as well as parts of concept maps may be selected by the mouse and subsequently be moved, copied or deleted. Concept maps are easily re-arranged as well as saved and re-loaded. In addition, every step taken to construct, extend or modify a concept map can be saved for later analysis.

In order to learn how to use the computerized concept mapping tool, in the first section of the study, the students worked on an introduction to the tool. As in the introduction to concept maps, the concepts addressed in the introduction to the concept mapping tool referred not to classical mechanics but to well-known household furniture.

**Instructional Unit** We designed an instructional unit to teach the students qualitative and quantitative information on classical mechanics by means of concept maps. It was made up of three parts. In the first part, coordinate systems and vectors as well as the addition and resolution of vectors were described. In the second part, qualitative and quantitative information on kinematic concepts such as displacement, velocity and acceleration was presented. In the third part, qualitative and quantitative information on dynamic concepts such as gravitational force, normal force, friction force and resultant force was delineated.

The qualitative and quantitative information on the different concepts was described by means of concept maps. One or more concept maps were followed by several examples and exercises. The solutions to the exercises were also presented. In 100 pages total, the unit comprised 30 concept maps, 18 examples and 20 exercises along with their solutions.

The students worked on the instructional unit in the second section of the study. In a first step, they attempted to elaborate the information included in a concept map. In a second step, the students had the opportunity to consider an example. It illustrated the consequences of applying the information included in a concept map to a certain arrangement. In a third step, the students themselves exercised the application of the information included in a concept map to other arrangements. While some of the exercises asked for the construction or completion of diagrams, other exercises asked for the construction of concept maps. The students always constructed diagrams by paper and pencil. Concept maps were always constructed by taking advantage of the computerized concept mapping tool. Finally, the students were allowed to compare their solution to an exercise with the solution presented in the instructional unit.

**Problems to be Solved with Support from Tutors** Four different problems for problem solving with support from tutors were set up. For example:

A sledge of mass  $m = 10$  kg moves on a horizontal surface with a velocity of  $v_0 = 4.8$  m/s. The coefficient of friction between the runners of the sledge and the surface equals  $f = 0.12$ . After which distance  $r$  has the sledge's velocity reduced to  $v = 0$  m/s?

By making use of a simulation program of qualitative and quantitative problem solving in classical mechanics (Ploetzner, 1995), the problems were designed in such a way that – relative to the information presented in the instructional unit – their solutions demanded the coordinated application of knowledge about both qualitative and quantitative information. In order to design the problems, the simulation program was equipped with formal representations of the qualitative and quantitative information which was presented in the instructional unit. Afterwards, the simulation program was applied to formal representations of the four problems. When the simulation program was furnished with either qualitative or quantitative information, its problem solving attempts failed. The problem solving attempts succeed only when the simulation program was furnished with both qualitative and quantitative information.

**Strategies Applied by the Tutors** In the fourth section of the study, the students attempted to solve the four problems with support from tutors. While one group of students was supported in qualitative problem solving, the other group of students was supported in quantitative problem solving. Two physics students from the School of Education at Freiburg served as tutors. Both were trained to support the students in either qualitative or quantitative problem solving by means of two different problem solving strategies. The strategies are described in Table 1. The strategy to support qualitative problem solving focused on the construction and interpretation of free-body diagrams. The strategy to support quantitative problem solving addressed the systematic use of algebraic equations.

**Table 1:** The strategies applied by the tutors

---

**Strategy to support qualitative problem solving**

1. Drawing a sketch:
  - Identify the body!
  - Is the body in contact with the surface?
  - Draw a sketch!
2. Determining the resultant force:
  - Determine the forces on the body!
  - Draw an arrow for each force!
  - Determine the resultant force on the body!
  - Describe the resultant force algebraically!
  - Is it possible to simplify the algebraic description?
  - Draw a coordinate system!
  - Describe the magnitude of the resultant force relative to the coordinate system!
3. Relating the resultant force to the acceleration:
  - How is the resultant force related to the body's acceleration?
  - Determine the direction of the body's acceleration!
  - Determine the direction of the body's velocity!
  - How does the acceleration affect the velocity?

**Strategy to support quantitative problem solving**

1. Identifying the given and sought variables:
    - Identify the variables whose values are given!
    - Identify the variables whose values are sought!
  2. Selecting an algebraic equation:
    - Select an equation which includes a variable whose value is sought!
    - Attempt to apply Newton's second law  $\Sigma F = m \cdot a$ !
  3. Applying an algebraic equation:
    - Identify the variables whose values are known!
    - Identify the variables whose values are unknown!
    - If the values of all variables in an equation are known except the value which is sought, then substitute the variables for their values and compute the value which is sought!
    - Otherwise, select equations which include variables whose values are unknown and determine the unknown values!
    - After applying an equation, verify the units!
- 

Initially, the tutors explained and demonstrated the problem solving strategy they supported. Thereafter, the students attempted to solve the four problems. They worked on each problem in two phases. In the first phase, the students approached a problem on their own. To describe a problem's solution, the students constructed diagrams by paper and pencil as well as concept maps by taking advantage of the computerized concept mapping tool.

In the second problem solving phase, the students received

support from the tutors. The tutors assisted the students after they completed the first problem solving phase or when they did not show any further progress in their problem solving attempts. If the students raised questions which concerned problem solving steps addressed by the tutors' problem solving strategy, the tutors delineated the problem solving steps and encouraged the students to carry them out. If the students were not able to accomplish this, the tutors explained and demonstrated the problem solving steps. Afterwards, the students had to reproduce the tutors' explanation using their own words.

The tutors also encouraged the students to explain their partial or complete solution to a problem. Whenever a problem solving step addressed by the tutors' problem solving strategy was correct, the tutors provided affirmative feedback to the students. Whenever a problem solving step addressed by the tutors' problem solving strategy was incorrect or missing, the tutors indicated the error or omission to the students. Thereafter, the tutors encouraged the students to correct or add the problem solving step. Again, if the students were not able to accomplish this, the tutors explained and demonstrated the step. Afterwards, the students had to reproduce the tutors' explanation using their own words.

**Multi-Component Tests** In the third as well as in the fifth section of the study, the students worked on a multi-component test which assessed their knowledge about qualitative and quantitative information on classical mechanics. Each test was made up of three different components and comprised 16 problems in total. In order to design the problems, we again took advantage of the simulation program of qualitative and quantitative problem solving in classical mechanics (cf. Ploetzner, 1995).

The first component comprised four problems which assessed knowledge about qualitative information on classical mechanics. These problems were designed in such a way that – relative to the information presented in the instructional unit – their solutions only demanded the application of knowledge about qualitative information on classical mechanics. Correspondingly, the second component comprised four problems which only required the application of knowledge about quantitative information. The third component was made up of eight problems whose solutions demanded the coordinated application of knowledge about both qualitative and quantitative information.

Both tests comprised parallel problems. Each pair of parallel problems were designed in such a way that the same knowledge was applied by the simulation program of qualitative and quantitative problem solving to solve them. However, non-structural features such as the involved entities and numerical values varied across parallel problems. Within each test, the problems were arranged in random order.

The design of the tests allows one to hypothesize which problem solving performance should be observable in the three test components of the pre- and posttest.

With respect to the first test component on qualitative information, we predict that many problems can already be solved in the pretest after studying the instructional unit. While the qualitatively supported students should further improve from the pre- to the posttest, the quantitatively sup-

ported students should not do so.

With respect to the second test component on quantitative information, we also hypothesize that many problems can already be solved in the pretest. While the quantitatively supported students should further improve from the pre- to the posttest, the qualitatively supported students should not do so.

In contrast, with respect to the third test component on the coordination of qualitative and quantitative information, we predict that only few problems can already be solved in the pretest. Both qualitatively and quantitatively supported students should improve from the pre- to the posttest. We especially hypothesize, however, that qualitatively supported students improve considerably more than quantitatively supported students.

## Subjects

Twenty-four tenth graders, 11 girls and 13 boys, from three different high schools volunteered for the study. While the group of students which was supported in qualitative problem solving comprised 6 girls and 6 boys, the group of students which was supported in quantitative problem solving comprised 5 girls and 7 boys.

Before the study was conducted, the students' general ability was assessed by means of the Advanced Progressive Matrices Test (Raven, 1976). Subsequently, two students who had received the same or almost the same test scores were assigned to different groups. While the average test score of the students who received support in qualitative problem solving was 24.33 (SD = 3.60), the average test score of the students who received support in quantitative problem solving was 23.92 (SD = 3.85). Students from different schools also were equally distributed among the two groups. Furthermore, in each group of students, one half of the students received support from one tutor and the other half received support from the other tutor. The students were paid for their participation.

Because in German high schools Newtonian mechanics is commonly taught to eleventh graders, none of the students had attended classes on Newtonian mechanics as it was addressed in this study.

## Procedure

The students were investigated individually for four days running. On the first day, they worked on the introduction to concept maps, on the introduction to the computerized concept mapping tool, and on the first part of the instructional unit. On the second day, the students worked on the remaining parts of the instructional unit and on the pretest. On the third day, the students attempted to solve the first two problems with support from tutors. Finally, on the fourth day, the students attempted to solve the remaining two problems with support from tutors and worked on the posttest.

## Results

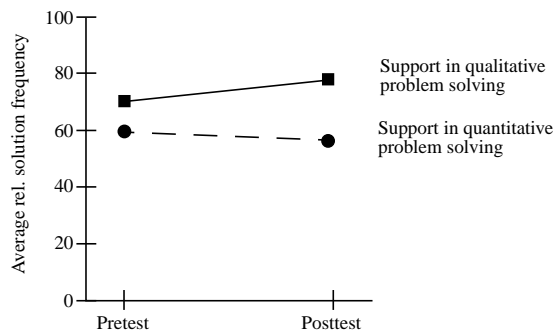
### Times Spent

On average, both groups spent virtually the same amount of

time on the different sections of the study ( $M = 73$  vs.  $M = 75$  minutes on the introduction,  $M = 221$  vs.  $M = 219$  minutes on the instructional unit,  $M = 78$  vs.  $M = 85$  minutes on the pretest,  $M = 154$  vs.  $M = 159$  minutes on problem solving and  $M = 88$  vs.  $M = 86$  minutes on the posttest).

### Problem Solving Performance

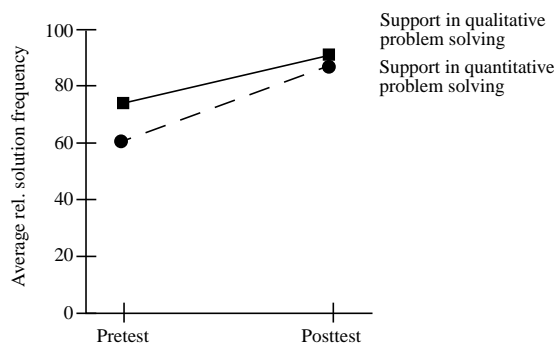
The average relative solution frequencies in the first test component, which assessed knowledge about qualitative information on classical mechanics, are displayed in Figure 1. In accordance with our expectations, the students had acquired considerable knowledge about qualitative information by studying the instructional unit. With respect to the first test component, although statistically not significant, only the qualitatively supported group improved a little from the pretest to the posttest.



**Figure 1:** Problem solving performance in the test component on qualitative information.

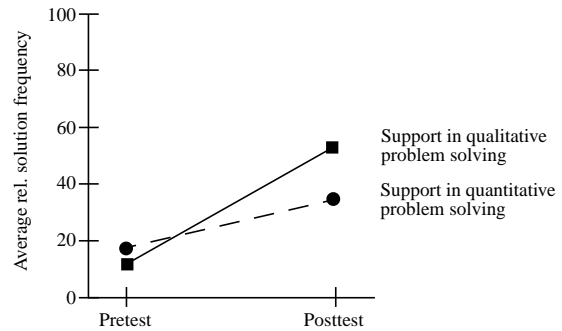
The average relative solution frequencies in the second test component, which assessed knowledge about quantitative information on classical mechanics, are shown in Figure 2. Again, as expected, the students had acquired substantial knowledge about quantitative information by studying the instructional unit. Furthermore, on average, the qualitatively as well as the quantitatively supported group improved significantly from the pretest to the posttest ( $F(1, 22) = 27.72$ ,  $p < .001$ ).

Figure 3 displays the average relative solution frequencies in the third test component which assessed the coordinated use of knowledge about qualitative and quantitative information on classical mechanics. In accord with our expectations,



**Figure 2:** Problem solving performance in the test component on quantitative information.

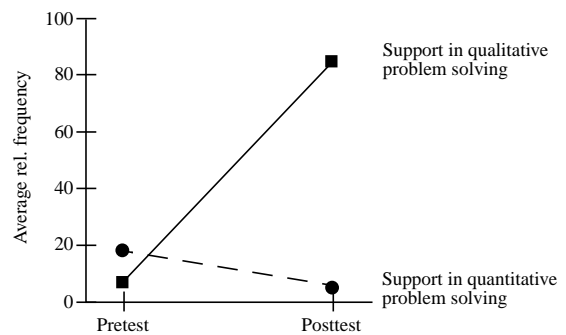
with respect to this test component, the students exhibited rather poor performance after studying the instructional unit. On average, both groups improved significantly from the pretest to the posttest ( $F(1, 22) = 46.48$ ,  $p < .01$ ). Furthermore, the interaction *Test* x *Group* indicates that the qualitatively supported group improved significantly more from the pretest to the posttest than the quantitatively supported group ( $F(1, 22) = 4.47$ ,  $p < .05$ ).



**Figure 3:** Problem solving performance in the test component on qualitative and quantitative information.

### Problem Solving Approach

With respect to the third test component, which assessed the coordinated use of knowledge about qualitative and quantitative information on classical mechanics, it was also analyzed how frequently the students approached these problems qualitatively and quantitatively.



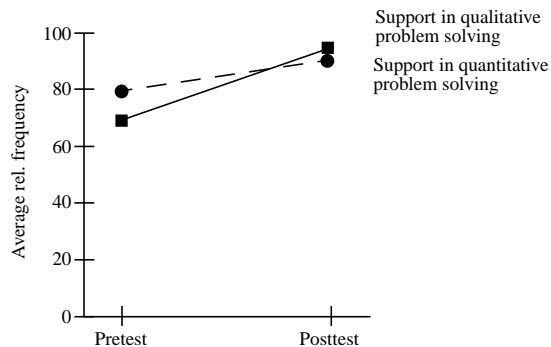
**Figure 4:** Qualitative problem solving approaches.

The average relative frequencies of qualitative and quantitative problem solving approaches are shown in Figure 4 and 5. The average relative frequency of qualitative problem solving approaches increased significantly from the pretest to the posttest ( $F(1, 22) = 54.68$ ,  $p < .01$ ). Due to the support from tutors, the students who were supported in qualitative problem solving drew more frequently a free-body diagram than the students who were supported in quantitative problem solving ( $F(1, 22) = 28.73$ ,  $p < .01$ ). The interaction *Test* x *Group* further demonstrates the consequences of the support from tutors. While the qualitatively supported group largely increased the number of qualitative problem solving attempts from the pretest to the posttest, the quantitatively supported group even decreased the number of qualitative



problem solving attempts ( $F(1, 22) = 103.38, p < .01$ ).

The average relative frequency of quantitative problem solving approaches also increased significantly from the pretest to the posttest ( $F(1, 22) = 17.75, p < .01$ ). As expected, however, with respect to the use of algebraic equations the qualitatively supported group did not differ significantly from the quantitatively supported group. There is also no statistically significant interaction *Test x Group*.



**Figure 5:** Quantitative problem solving approaches.

## Discussion

We presented an experimental study which started from the hypothesis that physics instruction with an emphasis on qualitative problem solving is more effective than physics instruction with an emphasis on quantitative problem solving. The focus of our analysis was on the solution of problems which demand the coordinated application of knowledge about qualitative and quantitative information on classical mechanics.

In such a context, the support of qualitative reasoning as well as the support of quantitative reasoning should enhance the students' problem solving performance. However, while quantitative information frequently helps to guide the use of qualitative information, the appropriate use of quantitative information very often seems to presuppose qualitative understanding (e.g., Chi, Feltovich & Glaser, 1981; de Kleer, 1977; Ploetzner, 1995). Without qualitative understanding, the duality of the physical situation under scrutiny and the quantitative structure set up gets easily lost. Therefore, we expected that the support of qualitative reasoning improves the students' problem solving performance more than the support of quantitative reasoning.

The results are in accord with our expectations. Both the support of qualitative reasoning and the support of quantitative reasoning significantly improved the students' problem solving performance. Especially, students who were supported in qualitative problem solving improved significantly more than students who were supported in quantitative problem solving.

Our results also underline an observation repeatedly made in psychological and educational research on problem solving in formal sciences such as physics. When problems have to be solved which ask for a precise quantitative solution, students strongly tend to focus on the use of quantitative-numerical information and to neglect the use of qualitative-conceptual information. While in the presence of quantita-

tive problems the necessity to make use of quantitative-numerical information seems to be obvious to the students, the necessity of applying qualitative-conceptual information needs again and again to be pointed out to the students as well as its use needs to be encouraged and supported.

## Acknowledgements

This research was supported by the German National Research Foundation (DFG) under contract PL 224/2-2. We thank Frank Tewissen, Andreas Loesch and Ulrich Hoppe from the research group COLLIDE at the University of Duisburg (Germany) for making their computerized concept mapping tool available to us.

## References

- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152.
- de Kleer, J. (1977). Multiple representations of knowledge in a mechanics problem-solver. *Proceedings of the Sixth International Joint Conference on Artificial Intelligence* (pp. 299-304). San Mateo, CA: Morgan Kaufmann.
- Hestenes, D. (1987). Toward a modeling theory of physics instruction. *American Journal of Physics*, 55, 440-454.
- Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge - Techniques for representing, conveying, and acquiring structural knowledge*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Larkin, J. H. (1983). The role of problem representation in physics. In D. Gentner, & A. L. Stevens (Eds.), *Mental models* (pp. 75-98). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Pfundt, H., & Duit, R. (1994). *Bibliography: Students' alternative frameworks and science education* (4<sup>th</sup> ed.). Kiel: Institute for Science Education.
- Ploetzner, R. (1995). The construction and coordination of complementary problem representations in physics. *Journal of Artificial Intelligence in Education*, 6, 203-238.
- Ploetzner, R., Fehse, E., Kneser, C., & Spada, H. (1999). Learning to relate qualitative and quantitative problem representations in a model-based setting for collaborative problem solving. *The Journal of the Learning Sciences*, 8, 177-214.
- Ploetzner, R., Hoppe, H. U., Fehse, E., Nolte, C., & Tewissen, F. (1996). Model-based design of activity spaces for collaborative problem solving and learning. In P. Brna, A. Paiva, & J. Self (Eds.), *Proceedings of the European Conference on Artificial Intelligence in Education* (pp. 372-378). Lisbon: Colibri.
- Raven, J. C. (1976). *Advanced Progressive Matrices, Sets I and II*. London: Lewis.
- Reif, F., & Heller, J. I. (1982). Knowledge structures and problem solving in physics. *Educational Psychologist*, 17, 102-127.
- VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology*, 47, 513-539.
- White, B. Y. (1993). ThinkerTools: Causal models, conceptual change, and science education. *Cognition and Instruction*, 10, 1-100.

# The Implications of Cognitive Science for the Significance of Experimentation in Science Teaching

Athanasios Raftopoulos (raftop@ucy.ac.cy)  
Constantinos P. Constantinou (c.p.constantinou@ucy.ac.cy)  
Department of Educational Sciences  
University of Cyprus  
P.O. Box 20537  
Nicosia 1678, Cyprus

## Abstract

In this paper we argue for a new role for experiment in science teaching and learning. Our proposition is based on the conception of experiment as an active ingredient of theory construction and not as a mere tool for theory testing. This latter view is based on the classical conception of the mind-world interaction, according to which human action purports to test the validity of a tentative solution to a problem and follows after mental processing. We present the new framework that views the interactions with the environment as active ingredients of the mind's problem solving activity. We also adduce evidence for this new role of experiment from the history of science. Finally, we discuss the repercussions of this view of cognition, as the activity of a mind-environment inseparable whole for the role of experiment in knowledge construction.

## Introduction

Experimentation was traditionally deemed to be the main prerequisite for the successful teaching of physical sciences in school, mainly because the experiment was construed as a means of confirmation of theories. As such, it could persuade the student about the adequacy of the theory presented in class and lead her to embrace it. This construal of experiment was based on the thesis that experiment follows theory with a view to testing it empirically (which was the prevalent view in philosophy of science until the 1960's).

This conception about the role of experiment in science, first, and education, later, was subsequently criticized on philosophical, psychological, and educational grounds. The main argument against the standard conception of the experiment was that:

(a) the student has formed a well established body of beliefs, intuitive theories, or phenomenological primitives (diSessa, 1993) about the world before she attends school, which constitutes an alternative, well entrenched, theory to

those taught in class (psychological critique);

(b) the knowledge that the student brings to a given learning situation influences the meaning that she constructs in that situation;

(c) experiments are not sufficient to establish the adequacy of theoretical ideas, since by themselves they do not constitute the basic criterion of choice among alternative theories (philosophical critique).

Therefore, experiments should abdicate their decisive role in science education, since even if the student actively participated in their making, they do not suffice to allow her to build the required concepts and to persuade her to abandon her intuitive theories. To demonstrate this point further, one could cite extensive research showing the failure of instruction with regard to classical physics.

In this paper we will briefly present the theoretical framework that led to the dispute about the role of experiment in education. We will argue that this framework is based on an erroneous conception of the role of experiment in problem solving, and *a fortiori* in science. We will claim that this error is based on the classical conception of the mind-world interaction in cognitive science and we will present the new framework emerging in cognitive science that views the interactions with the environment and experimentation not as a follow up of the mind's output purporting to test the validity of a tentative solution to a problem, but as an active ingredient of the mind's problem solving activity, that extend mind beyond its biological boundaries to the world (Clark, 1997). We will also adduce evidence for this new role of experiment from the history of science. Finally, we will discuss the repercussions of this view of cognition as the activity of a mind-environment inseparable whole for the role of experiment in knowledge construction and we will argue for the importance of experiment not as a test of theory only but as an integral part of theory construction.

## 1. Undermining the Role of the Experiment:

### An Overview

Logical positivism, the main philosophical paradigm during the first half of the 20th century conceived of experiment as a scientific activity that follows the theoretical, or mental, processing of raw data aiming to provide empirical testing of scientific theories. Hanson (1958), Kuhn (1962), Gregory (1973), Lakatos (1978) and others criticized this classical conception of science. They outlined the non linear and non cumulative character of the scientific enterprise, a conception that undermined the role of experience and of the experiment in the rationalistic choice among competing theories. In this context, the realization that experience is always interpreted through the lenses of a theoretical framework, led, on one hand, to diminishing the importance of the experiment as a means of theory testing, and on the other hand, to the marking out of the role of theory as the framework within which empirical data are interpreted.

This crisis regarding the role of experiment could not bypass the experiment as an instructional medium. The tendency towards criticizing experiment was strengthened in the early 1970's by the findings of psychologists (Carey, 1985, 1992; Chi, 1992; Karmiloff-Smith, 1992; Medin, 1983; Nersessian and Resnick, 1989; Rosch, 1978; Spelke, 1990) that a child's mind is no *tabula rasa* upon which the educator is called to imprint the acceptable scientific theories included in the curriculum by proving them experimentally. On the contrary, children have innately acquired, or constructed very early on the basis of some innate constraints, a set of persistent beliefs about the world (intuitive or naive theories).

Seen in this context, the experiment loses its function as the means par excellence of testing and proving theories, since the child has already an intuitive conceptual background from which she can formulate various interpretations of the experimental results that are not compatible with those interpretations that the instructor seemingly wishes the student to acquire. Thus, the mere presentation of, or participation in conducting, experiments does not suffice to prompt the student to accept the intended interpretations.

Leaving aside the issue of whether this set of beliefs constitute a theory, or merely a body of incompatible principles, one notes that even though they contain principles that allow children to make personal sense of their world-experience, they generally deviate from established scientific theories (Carey, 1985; Clement, 1982, 1983; Halloun and Hestenes, 1985; McCloskey, 1983; Nersessian and Resnick 1989; Viennot, 1979). These persistent ideas are epistemological obstacles that instruction must guide students to overcome, if it is to be effective. These obstacles are not merely erroneous pieces of knowledge about the world that the child could easily be persuaded to reject. Since they constitute the schemata on the basis of which she has come to interpret the world, they function as organizing principles. All experiences are made meaningful on the basis of these principles, and as a result, they are the least likely items to be put under experimental inquiry (Quine, 1961).

The criticism of experiment and of its role in science education was reinforced by the well established failure of traditional instruction of, say, Newtonian physics.

This intense criticism of experiment was accompanied by the realization of the need to complement instruction by exposing students in a systematic way to their own intuitive theories and guiding them to compare them explicitly and in detail with current scientific theories. The aim was to make students conscious of the implicit principles that they use to organize and understand the world, to render clear the points at which their intuitive theories are in conflict with the acceptable scientific theories, to make them realize that the latter are more adequate in explaining the world, and finally to lead them to construct the salient concepts of established scientific theories.

Theory becomes, thus, predominant in science education, while experiment loses some of its shine. This criticism of experiment does not imply of course the abandonment of experiments as educational means. It simply points out that experiments by themselves, without the presentation and discussion of the appropriate theoretical background, are no panacea for proper instruction in the natural sciences.

The educational paradigm that emerged in this new framework continued to conceive of the experiment as a follow-up to theory, a discrete step in the scientific enterprise of theory construction, whose role is the empirical testing of theory. Once this empirical testing is put in doubt, experiment automatically loses its appeal. In that regard, this new paradigm does not differ from the one it superseded.

This classical conception of the experiment is based on the belief that all cognizers (and *a fortiori* scientists) when engaged in a problem solving activity function according to the scheme: reception of external input (the data of the problem), mental processing of the internal representations of these data to figure out a solution (problem processing that consists in a search of the problem space) and, finally, output of a tentative solution to the problem that is tested for empirical adequacy and for compatibility with a body of accepted knowledge. This is the well known <input → mental processing → output> scheme of classical cognitivism, which in the case of scientific problem solving becomes the well known positivist scheme <experience → mental formal processing → experiment>.

## 2. A new Role for Experiment: Its Cognitive Background

The classical view of the interaction between cognizers and their environment, and therefore the classical view of experiment in everyday problem-solving and the scientific enterprise is severely questioned by some new tendencies in cognitive science and by the research findings in the history of science. Some cognitive scientists (Bickhard, 1993, 1998; Clark, 1993, 1997; Clark and Thornton, 1997; Elman, 1991; Hutchins, 1995; Rutkowska, 1993, Varela, et. al., 1993) on the one hand hold a different view for the cognizer-environment, and thus the theory-experiment,

interaction, which radically revises the relation between the mind and the world. Research in the history of science (Franklin, 1986; Gooding et. al. 1989; Gooding, 1990; Hacking, 1983; Nersessian, 1984), on the other hand, reveals that the experiment plays a much richer role than merely being a test of empirical adequacy of scientific theories.

According to the new conception of cognition, the mind does not function autonomously from the environment, in the sense that its relation with it does not consist simply in passively receiving input from it, and eventually processing it in its effort to find a solution to a certain problem. Instead, the strategies of mind include actions upon the world as an integral part of its problem-solving activity (Clark 1997), and, one might add, as a part of theory construction. This active intervention of the mind may transform the problem space, affecting the problem-solving process itself.

This can be done in various ways. First, the intervention upon the world may bring into light new data that could transform the problem-space, rendering its search more effective or even possible. Our action, for instance, might reveal some regularities that shed new light upon the existing data allowing perhaps their re-categorization, opening thereby new research avenues. Or, this same action may reveal some structural similarities, that were not lying in the surface structure of the problem-space, which allow the conceptual redeployment from another different field on to the given problem.

Second, the active intervention upon the environment may scaffold it so that the problem-space is structured in such a way that its effective search could be conducted, even allowing for the limited cognitive, perceptual and motor resources of the cognizer (Clark, 1997; Elman, 1991; Raftopoulos, 1997).

All these are ways of reducing, what Clark and Thornton (1997) have called, type-2 difficult problems, (i.e., problems whose statistical regularities do not lie on their surface structures but in their deep structure), to type-1 problems that wear their statistical regularities on their sleeves, and thus can be effectively solved by means of inductive heuristics.

The intervention on the environment, viewed as a part of theory construction, allows one to make sense of what it means to say that the learning process itself induces changes in the structures involved in learning. One way to understand this claim is to say that the neural substrate undergoes changes while it learns, as a result of this learning (Quartz and Sejnowski, 1997). Another way, is to interpret this statement to mean that the learning process changes the representational basis in which the search of the problem space takes place and this change influences this learning.

The active role of the mind and its action upon the environment results in the construction of new representations (either external or internal). This offers an alternative to the classical picture of learning as a search within a defined representational problem space (the problem of selective induction.) The cognizer builds representations as she learns, and thus shapes the

hypothesis-space. Since learning depends crucially on the statistical regularities of the problem input and the structural characteristics of the learner, the structure of the training data (and thereby the structure of the problem domain from which these data are drawn) and the processing characteristics of the learner shape the hypothesis space to their constraints and requirements.

Learning, thus, need not be an inductive search through a hypothesis space delineated by fixed representations that restricts search to solutions that can be expressed only by means of the pre-existing representations, in so far as new representations can be built during learning. The result is that processing strategies and representations co-evolve (Clark, 1993; Horgan and Tienson, 1996).

In this sense the result of our action upon the environment does not consist simply in testing a tentative solution to a given problem, and herein lies the fallacy of cognitivism's view of the mind, but in an active intervention upon the environment with a view to discovering new data and building new representations that might help the mind in solving the problem. This action becomes an active ingredient of the problem-solving process, and in the case of the scientific activity, an active ingredient of theory construction.

Schunn and Klahr (1995) offer a computational account of problem solving seen as a search in four problem spaces. These are the *data representation space* (from which representations of the salient data are chosen), the *hypothesis space* (in which hypotheses about causal relations amongst the data are drawn), the *experimental paradigm space* (the classes of experiments relevant to the problem at hand are chosen), and the *experiment space* (in which the values of the parameters within the selected paradigm are chosen). Though we do not have the space here to discuss this model in detail here, one can safely say that the upshot of the model is that the solution of a problem involves a constant flow of information among the four spaces. As a result, the processing within each space depends crucially upon the state of the research in the other spaces. This model shows clearly what it means to say that the learning process itself induces changes in the structures involved in learning, that new kinds of representations are developed which affect the search of the hypothesis space and so forth.

Our world is not merely a place in which we can store information and the testing ground of our theories and tentative solutions, although it certainly functions this way as well. It is also, and perhaps predominantly, the space upon which we act by transforming it and by building external representations so that it becomes an aid to the mind. Understanding the mind presupposes the rejection of the conception of the mind as isolated from the world building internal representations and models and processing them to discover solutions to problems (Rutowska, 1993). This view must be replaced by a mind situated in the world that uses it to facilitate its work and which shows that "the real power of human cognition lies in our ability to construct functional systems that accomplish our goals." (Hutchins 1995, 316).

This movement in cognitive space helps explain, and thus is being strengthened by, findings in the history of science that show that the experiment plays a much richer role in the scientific enterprise than being a mere test of empirical adequacy. The study of the actual processes of theory construction, based either on the notebooks and letters of scientists (Newton or Faraday, for instance) or on the in-situ observation of the workings of a research team renders clear that the experiments transcend the theory in the context of which they are first conceived. They acquire their own autonomy, they become themselves objects of inquiry independently of the theory and they are used not just to test the theory but also to discover new evidence that would facilitate the theoretical enterprise. They accomplish this either by revealing structural similarities with other domains, allowing thus conceptual redeployment, or by bringing forth certain basic regularities that reorganize the existing data, transform the problem space and, thereby, allow the discovery of the hidden structure. They also actively participate in the construction of the (partial) meaning of the theoretical terms of the theory.

### **Discussion: A New Educational Role for the Experiment**

We have seen that, according to cognitivism, the cognizer receives environmental input, builds internal representations and models of the world-situation pertaining to the problem, processes these representations and produces, as output, a tentative solution to the problem. The view of cognition that emerges from our discussion is entirely different. The cognizer is not a passive processor of information from the environment. She acts upon the environment, discovers new data that transforms the problem-space and diminishes the cognitive load of the problem. Hence, the problem space and the opportunities it offers for exploitation become an inseparable part of the problem-solving activity. Thus, the mind transcends its biological boundaries and extends itself to the world. This means, in return, that the well arranged triplet <input-processing-output> cedes its place to an action loop, that is, an interaction in which thought leads to actions which in turn change or simplify the problems confronting thought (Clark 1997). The continuous interaction between mind and environment becomes so intricate and complex that it is difficult to talk of two distinct factors that interact and is better to conceive them as forming an inseparable whole, which gives rise to cognition.

Learning in the physical sciences constitutes the development of a coherent conceptual framework that consists of a network of conceptual models within which mental models are constantly re-negotiated in dynamic interaction with the framework. Conceptual models are robust mental constructs that can be developed through appropriate instructional intervention. In effective learning environments, both mental and conceptual models are processed and manipulated consciously and explicitly by the learner.

In the context of our discussion, learning in science emerges as a process of elaboration of mental models through dynamic interaction between mind and environment. In this interaction, experiment as well as logical argumentation and syllogism both contribute in a dynamic and integral manner to the constructive process.

This view of the experiment has important implications for current classroom interpretations of the constructivist paradigm. Constructivism has attracted a lot of attention in science over recent years partly as the overarching framework underpinning active and collaborative learning. Constructivist classroom strategies invariably seek to facilitate learning outcomes by taking the students through a cycle of stages including formulation of ideas, cognitive conflict, knowledge re-organisation and extension. However, the assumptions that underlie the development and implementation of such constructivist strategies are at odds with the framework that we have presented here.

Before we go on to discuss this, we need to elaborate on two issues. Firstly, the conceptual models whose construction is the objective of effective science learning environments are not necessarily identical with established scientific theory. Learning is the outcome of individual construction of meaning even when that happens in a collaborative environment. Research in science education has repeatedly demonstrated that instructional approaches that rely on a knowledge transmissive model of teaching lead to rote memorization rather than real learning. Examples include rote applications of Ohm's law without fundamental understanding of the current model for electric circuits (McDermott and Shaffer, 1992), calculation of image magnification without basic understanding of geometrical optics (Wosilait et al., 1998) and rote application of the work-energy theorem without the basic realization that work is done by one body on another (O'Brien et al., 1998). It would appear that any effort to transmit knowledge to a group of learners does not usually result in effective construction of meaning.

Second, conceptual models are not in one to one correspondence with the phenomena they seek to model. The learning outcome in science is a series of mental constructs that seek to code and process specific aspects of the behavior of physical phenomena. For the individual learner, both the science discipline and the conceptual models should be aligned with the physical world in the way that this is observed and coded by the learner's mental processes. In other words, both the outcome of research (a socio-cultural construct by a community of researchers) and the outcome of learning (a cognitive construct of the learner) cannot be conceived as a mental reflection of physical reality but only as mental constructs that aid us in systematically pursuing this interaction between mind and matter. For instance, the theory of the Big Bang does not describe the birth of the Universe as we currently know it; rather it seeks to describe the birth of the Universe had an observer been there to observe what happened, for Physics and other sciences are constrained to formulate questions, hypotheses and theories that are epistemologically compatible with the mind-environment

interaction that is inherent in their development.

In view of these, the experiment in natural sciences plays a more fundamental and complex role than was traditionally thought. It is not just a means for testing and confirming a theory (as was conceived and implemented by traditional instruction with the laboratory as a supplement to the theoretical lecture-based transmission of the knowledge to be learned), or a means of choosing among conflicting predictions and alternative theories (as is conceived by the modern proponents of the constructivist model in education). In view of the fact that learning is a process of mental construction, and as such the product of the interaction of the mind with the physical world, the experiment provides the means of this interaction and implements it by enabling the construction of meaning.

Inquiry-based approaches to teaching science are closer to this reconceived formulation of the role of experiment in the construction of meaning. In particular, the implementation of inquiry developed by the Physics Education Group at the University of Washington (McDermott, 1996) seeks to familiarize students with the process of using experimental evidence as a medium for recognizing the need of new concepts, constructing operational definitions of useful quantities and using those and the experimental evidence to synthesize models that are, in their turn, continuously open to validation and constant reformulation in the light of new evidence.

To demonstrate the way experimenting can influence the mental representation, it would be useful to present an example from electric circuits. In *Physics by Inquiry*, students initially explore how they might be able to light a bulb with a single wire and a battery. At this stage, post-test data indicate that students have one of several models concerning the underlying cause. Although most often they give the name current flow to their models, they tend to describe flow models that begin at one point and end at another, or alternatively are unidirectional, always running from the battery to the light bulb, or even more commonly involving current consumption along the way. In subsequent experiments they short-circuit a battery with a bare wire and make the observation that both the battery and the wire get warm, and that all points of the wire at some distance from the battery get equally warm simultaneously. All three of these observations contradict different aspects of their initial models. When the issue is raised of what flow model might account for these observations, students have to tackle specific aspects of their initial model one by one until they arrive at a more valid representation of current flow. In the process they have to go back and forth between their observations and their model every time improving on both. The emerging representation is aligned with continuous flow that upon closure of the switch starts instantaneously at all points of the circuit and uniformly cover all parts of the circuit. Once students have developed a model for electric current, they can then use it to make predictions of the relative brightness of light bulbs in fairly complex circuits.

In the context of the interaction between mind and environment, the experiment accomplishes various

essential functions. Firstly, the experiment determines which aspects of a hypothesis or working theory are valid or in need of reformulation. The experiment also enables us to identify interacting variables and, via the confirmation of hypotheses, plays a substantial role in the construction of theory. Second, the degree to which the learning outcome is correct is not determined by the extent to which the outcome and hence the student ideas overlap with current scientific thinking. Rather this is determined by the experiments that are accessible to the learner up to the time that instruction takes place. The degree of correctness and of the validity of the learning outcome is determined by the epistemological basis of the experimental process that led to the construction of meaning. Real learning is a result of logical argumentation that feeds on experimental data.

This last point is in stark contrast to current innovative approaches that seek to implement the constructivist paradigm by shifting the student conceptions from the naïve to the established through cognitive conflict and knowledge reconstruction events. The experiment cannot be conceived as an instructional means of shifting student conceptions or as a means of embedding theoretical knowledge. The experiment is a viable tool in the science classroom, a tool that is continuously used in the construction of a coherent conceptual framework and guides subsequent theory development and evaluation by mediating the interaction between mind and matter that extends the boundaries of our cognition beyond the biological confines of the brain.

## References

- Bickhard, M. H. (1993). Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence*, 5, 285-333.
- Bickhard, M. H. (1998). Constraints on the Architecture of Mind. *New Ideas in Psychology*, 16, 97-105.
- Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: The MIT Press.
- Carey, S. (1992). The Origin and Evolution of Everyday Concepts. In R. Giere (ed.), *Cognitive Models of Science*, Minneapolis: Minnesota University Press, 89-128.
- Chi, M. T. H. (1992). Conceptual Change within and across Ontological Categories. In R. Giere (Ed.), *Cognitive Models of Science*, Minneapolis: Minnesota University Press, 129-86.
- Clark, A. (1993). *Associative Engines: Connectionism, Concepts, and Representational Change*. Cambridge, MA: The MIT Press.
- Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: The MIT Press.
- Clark, A., and Thornton, C. (1997). Trading spaces: Computation, representation, and the limits of uninformed learning. *Behavioral and Brain Sciences*, 20, 57-66.
- Clement, J. (1982). Student's Preconceptions in Introductory Mechanics. *American Journal of Physics*, 50, 66-71.

- Clement, J. (1983). A Conceptual Model discussed by Galileo and used Intuitively by Physics Students. In D. Gentner and A. L. Stevens (Eds.), *Mental Models*, Hillsdale, NJ: Lawrence Erlbaum, 325-339.
- diSessa, A. A. (1993). Toward an epistemology of physics. *Cognitive Science*, 12, 1-55.
- Elman, J. (1991). Learning and development in neural networks: the Importance of Starting Small. *Cognition*, 48, 71-99.
- Franklin, A. (1986). *The Neglect of Experiment*. Cambridge: Cambridge University Press.
- Gooding, D. (1990). *Experiment and the Making of Meaning*, Kluwer Academic, Dordrecht.
- Gooding, D., Pinch, T., and Schaffer, S. (Eds.) (1989). *The Uses of Experiment: Studies in the Natural Sciences*. Cambridge: Cambridge University Press.
- Gregory, R. (1974). *Concepts and Mechanisms of Perception*. New York: Charles Scribners and Sons.
- Hacking, I. (1983). *Representing and Intervening*. Cambridge: Cambridge University Press.
- Halloun, I. A., and Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Psychology*, 53, 1056-1065.
- Hanson, N. R. (1958). *Patterns of Discovery*. Cambridge: Cambridge University Press.
- Horgan, T., and Tienson, J. (1996). *Connectionism and the Philosophy of Psychology*. Cambridge, MA: The MIT Press.
- Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: The MIT Press.
- Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: The MIT Press.
- Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: Chicago University Press.
- Lakatos, I. (1978). *Philosophical Papers*. Cambridge: Cambridge University Press.
- McCloskey, M. (1983). Naive Theories of Motion. In D. Gentner and A. L. Stevens (Eds.), *Mental Models*, Hillsdale, NJ: Lawrence Erlbaum, 299-324.
- McDermott, L. C. and Shaffer, P. S. (1992) Research as a guide for curriculum development: An example from introductory electricity, Part I: Investigation of student understanding. *American Jour. Phys.* 61, 81
- McDermott, L. C. and The Physics Education Group (1996) *Physics by Inquiry*, Vol I & II, J. Wiley, NY.
- Medin, D. L. (1983). Structural Principles of Categorization. In Th. J. Tighe and Br. E. Shepp (Eds.), *Perception, Cognition, and Development: Interactional Analyses*. Hillsdale NJ: Lawrence Erlbaum Associates, 203-30.
- Nersessian, N. J. (1984). *Faraday to Einstein: Constructing Meaning in Scientific Theories*. Hingham, MA: Martinus Nijhoff Publishers.
- Nersessian, N. J., and Resnick, L. B. (1989). Comparing Historical and Intuitive Explanations of Motion: Does 'Naive Physics' have a Structure? *Proceedings of the Twelfth Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum, 412-17.
- Quartz, S. R., and Sejnowski, T. J. (1997). The neural basis of cognitive development: a constructivist manifesto. *Behavioral and Brain Sciences*, 20, 537-556.
- Quine, W. O. (1961). Two dogmas of empiricism. In W. O. Quine *From a Logical Point of View*. Cambridge, MA: Harvard University Press, 2nd ed., 20-47.
- O'Brien, Pride, T., Vokos, S., and McDermott, L. C. (1998) The challenge of matching learning assessments to teaching goals: an example from the work-energy and impulse-momentum theorems. *American Journal of Physics*, 66, 906.
- Raftopoulos, A. (1997). Resource limitations in early infancy and its role in successful learning: A connectionist approach". *Human Development*, 40 : 5, 293-319.
- Rosch, E. (1978). Principles of Categorization. In E. Rosch and B. B. Lloyd (Eds.), *Cognition and Categorization*. Hillsdale, NJ: John Willeys and Sons.
- Rutkowska, J. C. (1993). *The Computational Infant: Looking for Developmental Cognitive Science*. Hertfordshire: Harvester Wheatsheaf.
- Schunn, C. D., and Klahr, D. (1995). A 4-space model of scientific discovery. In J. D. Moore and J. F. Lehman (Eds.) *Proceedings of the Seventeen Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum, 196-111.
- Spelke, E. S. (1990). Principles of Object Perception. In *Cognitive Science*, 14, 29-56.
- Varela, F. J., Thompson, E., and Rosch, E. (1993). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: The MIT Press.
- Viennot, L. (1979). Spontaneous reasoning in elementary Dynamics. In *European Journal for Science Education*, 1:2, 205-221.
- Wosilait, K., Heron, P. R. L., Vokos, S. and McDermott, L. C. (1998) Development and assessment of a research-based tutorial on light and shadow. *American Journal of Physics*, 66, 906.

# Evidence for the processing of re-representations during the mapping of externally represented analogies

Michael Ramscar

michael@cogsci.ed.ac.uk  
Division of Informatics, University of Edinburgh  
2 Buccleuch Place, Edinburgh EH8 9LW, Scotland

## Abstract

High level descriptions of the analogical reasoning process in cognitive science have now converged to present a relatively unified account (Hummel and Holyoak, 1997). However, the broad, consensual account of analogy is still far from complete: whilst it is possible to give a good explanation of the mapping of larger, structured representations in analogy, accounts of the mappings of individual sub-elements in these representations are still under-specified. Here, we review some possible approaches to this problem, describe an experiment that provides some empirical support for the 're-representation' approach to sub-mapping, and then identify some shortcomings in the 're-representation' approach as it is currently conceived.

## Introduction

Cognitive science has made great strides towards answering the important question of 'How do humans reason by analogy?' If we take a familiar example, the analogy between the solar system and Rutherford's model of the atom, then it is possible to explain – in broad terms – exactly how it is that two seemingly disparate objects can both remind us of one another in the first place, and then how it is that we can make meaningful correspondences between them.

Studies have shown that reminding (or retrieval) is driven by a computationally inexpensive process that initially matches surface (or semantic) elements in representations (witness the frequency – and mundanity – of most similarity based reminders, such as a lamp-shade reminding a party joker of a hat; see Gentner, Ratterman and Forbus, 1993).

Analogical mappings, on the other hand, are determined by a relatively more computationally expensive process. Global, systematic structural similarities between items to be matched need to be detected in order to make the kind of 'deeper', inference supporting correspondences that characterise analogy (Gentner, 1983; Goswami, 1992; Holyoak and Thagard, 1995; Hummel and Holyak, 1997).

Whilst theories and models of analogy are very compelling at one level of abstraction, there are certain assumptions made by *all* analogical theories that beg interesting questions if one seeks a more detailed explanation. As one increases the resolution of the question 'How do humans reason by analogy?' it appears that there are important gaps in current theories and

process models. Here, we wish to consider just one aspect of one of these gaps: the problem we focus on is that of matching the 'semantics' of elements during the analogical mapping process. This problem can be summarised as follows: suppose that in your representation of the atom, you describe the motion of an electron in relation to the nucleus in terms of it "revolving around" the nucleus (perhaps this is how you ordinarily think about this motion). On the other hand, suppose that in your representation of the solar system you conceive the motion of the planets in terms of their "orbiting" the sun.

At one level of abstraction, it may be sufficient to say that similarities in the meanings – or usage – of these words determine these mappings. However, in a more detailed account – and model – of analogy we might wish to do more than appeal to humanistic intuitions about similarities of meaning. We might wish to account for the way in which these sub-elements of our representations of the atom and the solar-system are mapped onto one another with the same level of detail with which we account for the mappings between the representations themselves.

If we are to fully explain high-level mapping in analogy, we must also account for the way lexically distinct but 'semantically' similar items in representations are reconciled with each other in a way that allows high-level mappings to be made. Here, we review some possible approaches to this problem, and present some evidence that offers some support to a popular proposal in the literature: the *re-representation* hypothesis.

## Semantic reconciliation and the re-representation hypothesis

Perhaps the most straightforward way to explain the mapping between "revolving around" and "orbiting"<sup>1</sup> would be in conceptual terms. If "revolving around" and "orbiting" could be shown to decompose into some canonical conceptual representation (say "circumnavigating"), then the link between them could be explained by reference to that concept, and the process by which it is made. This proposal is put forward by Gentner, Ratterman and Forbus (1993):

"[the...] constraint of matching identical predicates assumes canonical *conceptual* representations, not lexical strings. Two concepts that are similar but not identical (such as "bestow" and "bequeath") are assumed to be decomposed into a canonical

<sup>1</sup> We shall refer to this as the problem of semantic reconciliation in analogy.



representation language so that their similarity is expressed as a partial identity (... “give”)” Gentner, Ratterman and Forbus (1993, p 553)

The main drawback to this proposal is the lack of any specification of what a canonical conceptual representation (or a canonical representation language) is. Research into the mental representation of concepts suggests that human conceptual representations are anything but canonical; the proposals for generalised theories of representation that exist in the concepts literature fall well short of providing the kind of ‘neat’ account of concepts that canonical conceptual representation assumes (see Komatsu, 1992; Ramscar and Hahn, 1998 for reviews).

This problem has not gone unrecognised. In conjunction with other factors, such as evidence of the important role that structural commonalities (the ‘what’ of analogy) play in ‘ordinary’ conceptual tasks (e.g. Ahn, 1998), and the sheer difficulty of distinguishing analogy from ‘ordinary’ conceptual tasks (Ramscar and Pain, 1996), a widespread view has emerged that suggests that analogy *itself* may play an important role in semantic reconciliation (Forbus, Gentner, Markman and Ferguson, 1997, Hummel and Holyoak, 1997).

The basic idea behind this is outlined by Forbus, Gentner, Markman and Ferguson (1997) who propose that semantic terms might be decomposed into sub-predicate re-representations, with mapping between these being determined using the same process as similarity based transfer:

“re-representation allows relational identity to arise out of... analogical alignment, rather than as a strict constraint on the input descriptions” Forbus, Gentner, Markman and Ferguson (1997, p 246).

A similar re-representation proposal is advanced by Hummel and Holyoak (1997):

“With the notion of chunked predicates and objects, LISA hints at a kind of recursive representation for meaning that may ultimately ground itself in basic perceptual primitives. In its current implementation, LISA can represent and map hierarchical propositions of arbitrary. Analogously, it is possible to imagine structures for roles and objects that are, themselves, deeply embedded recursive structures. The depth to which a role or object would need to be decomposed for the purposes of mapping would depend on the task at hand. For example, mapping ‘John lifted the hammer’ onto ‘Bill raised the book’ may require little or no decomposition of the predicates ‘lift’ and ‘raise’, which will have substantial overlap in their *semantic features*.<sup>2</sup> On the other hand, mapping ‘John lifted the hammer’ onto ‘Bill pushed the cart’ where the predicates have less feature overlap, may be more likely to depend on decomposition of ‘lift’ into ‘caused to rise’ and ‘push’ into ‘cause to move laterally’, thereby making explicit the parallelism of their internal structures. Recursively ‘rise’ and ‘move laterally’ might be decomposed into structures relating simpler predicates with basic perceptual primitives representing motion and locations in space residing at the very bottom.” Hummel and Holyoak (1997, p.457).

Whilst re-representation is a popular idea in the analogy literature, its current status is largely hypothetical: re-representation proposals are usually couched in terms that relate to computational models, and as yet no evidence has been offered to support the psychological validity of the proposal.

The following experiment was designed to formulate a concrete re-representation proposal, and explore it empirically. The problem of semantic reconciliation

revolves around supplying an account of what happens when two ‘semantically similar’ terms – “revolving around” and “orbiting” – are encountered during the mapping process. In ordinary usage, the representations of human category information involved in these processes are implicit; people know what – “revolving around” and “orbiting” mean, and they reconcile (or map between) the two terms accordingly. But the exact nature of *what* they know, and *how* such knowledge is represented appears to be inaccessible at the level of detail required to specify and model the underlying cognitive processes involved in the semantic reconciliation of the two terms.

Participants were asked to make inferences with the aid of two candidate bases (see figure 1). In both the target and each of the two candidate bases, the term that was crucial to determining the representation of higher order structure in the scenarios was a novel, artificial term. By supplying ‘definitions’ for that term, we hoped to be able to control the representations participants used for semantically reconciling particular terms during their analogising. By doing this, we hoped to test the prediction that participants would use the same *process* to match semantic items in their representations as they would in ultimately determining their analogies - i.e. that in these externally represented analogies, at least, participants would use and process re-representations to facilitate semantic reconciliation.

---

## SCENARIOS

### TARGET - The Guralaga

*can be found in Australia  
lives in Rainforests  
only eats gau-gau berries  
has a cronomus lucundus  
the cronomus lucundus enables the Guralaga to eat gau gau berries.*

### BASE 1 - The Mongret

*can be found in Australia  
lives in Rainforests  
only eats gau-gau berries  
has a probus razoris  
the probus razoris enables the Mongret to eat the gau gau berries.  
Thanks to the way they eat, Mongrets live to a ripe old age and rarely suffer from cancer*

### BASE 2 - The Crany Dog

*can be found in Papua new Guinea  
lives in the grassy backlands  
eats vegetation  
has a remulum grandoso  
because of the remulum grandoso the Crany Dog can eat vegetation.  
Crany Dogs are particularly prone to cancer, which originates in their digestive system.*

---

**Figure 1:** A base and two targets. The surface similarities between the target and base 1 are highlighted. The target and base 2 share few surface similarities

<sup>2</sup> The emphasis is ours

	Surface Match Base (SMB)	Structurally Similar Base (SSB)
Analogy Level	Shares <b>surface</b> features with target  <i>Structural</i> overlap with target determined by dictionary mapping	Doesn't share surface features with target  <i>Structural</i> overlap with target determined by dictionary mapping
Dictionary Level	Shares <b>surface</b> features with target  Doesn't share structure with target	Doesn't share surface features with target  Shares <b>structure</b> with target
Inference	In <b>type A</b> sets, the <b>B</b> inference is only supported by <b>surface</b> matches between target and SMB the dictionary entries  In <b>type B</b> sets, the <b>A</b> inference is only supported by <b>surface</b> matches between target and SMB the dictionary entries	In <b>type A</b> sets, the <b>A</b> inference is only supported by <b>structural</b> matches between target and SSB the dictionary entries  In <b>type B</b> sets, the <b>B</b> inference is only supported by <b>structural</b> matches between target and SSB the dictionary entries

Figure 2: The relationships between the base, targets, dictionary entries and inferences in the main stimulus groups.

## Experiment

### Participants

The participants were 170 volunteers, a mixture of postgraduate and undergraduate students from the Department of Artificial Intelligence, Centre for Cognitive Science, Department of Psychology and the Faculty of Music at the University of Edinburgh.

### Materials, Design and Hypotheses

The materials comprised 5 groups of specially constructed scenarios (figure 1) with corresponding sets of novel dictionary entries (figure 3) and candidate inferences for each group (figure 4).

To control for biases towards particular inferences, each scenario group was further sub-divided into two versions of the scenario sets, and two versions of the dictionary entry sets, so that each scenario / dictionary sub-set supported one of the two different candidate inferences.

To classify the different structural / featural relation amongst the scenarios, we used Gentner, Ratterman and Forbus's (1993) taxonomy of similarity relationships:

- *Literal similarity* matches include both common relational structure and common object descriptions;
- *Surface matches*: based upon common object descriptions, plus some first order relations;
- *Structural similarity*, matches based upon common system of relations.

The relations between the various scenarios in a given scenario group can be summarised as follows (see also figure 2): In a group in which re-representation supported inference A, the target and one candidate base scenario (the SSB, or structurally supported base) shared only structural matches; mappings between the SSB's dictionary entry and the base dictionary entry also shared only structural matches.

There was a structural correspondence between the target structure supported by the target dictionary entry and the SSB's dictionary entry which in turn structurally supported the transfer of candidate inference A in the base.

Mappings between the target and the other candidate base scenario (the SMB, or surface match supported base) were supported by shared surface features, and mappings between the SMB's dictionary entry and the target dictionary entry also shared common object descriptions. However, there was a structural correspondence between the base structure supported by the base dictionary entry and the SMB's dictionary entry which supported candidate inference B. This allowed participants to use shared surface features to determine semantic reconciliation, but still use structural correspondences (c.f. Gentner, 1983) to determine their inferences (in this case, making a 'literally similar' match at the analogy level; see figure 2).

In a group where re-representation supported inference B, this pattern of correspondences was reversed.

### "DICTIONARY ENTRIES"

#### BASE DICTIONARY ENTRY

##### Cronomus lucundus

*are unique to certain types of bird*  
*are important to berry eaters*  
*is a long spleen-like organ*  
*keeping berries in the cronomus lucundus allows the berries to slowly ferment, allowing the goodness inside the bitter skins to be released*

#### SMB DICTIONARY ENTRY

##### Probus razoris

*are unique to certain types of bird*  
*are important to berry eaters*  
*is a long plier-like bill*  
*crushing berries in the probus razoris allows the goodness inside their bitter skins to be released without the skins having to be swallowed*

#### SSB DICTIONARY ENTRY

##### Remulum grandoso

*is unique to certain types of dog*  
*are important to dogs which eat a wide range of vegetation*  
*is a short intestine-like organ*  
*keeping vegetation in the remulum grandoso allows it to slowly ferment, allowing the goodness inside the outer skins to be released*

Figure 3: Dictionary entries for the Target and two Bases in figure 1. Surface similarities between the Target and SMB are in bold italic print; the structural match between the Target and the SSB is in normal italic

---

## INFERENCES

A. *Guralaga live to a ripe old age and rarely suffer from cancer.*

B. *Guralaga are particularly prone to cancer.*

---

**Figure 4:** The target inferences for the stimulus group shown of the following pages. In a type A set, structural commonalities would support the A inference; surface similarities would support the B inference. In a type B set, structural commonalities would support the A inference; surface similarities would support the B inference.

To try and simplify the above: in each group of stimuli, the target and candidate base scenario, and their corresponding dictionary entries, shared surface features, and a higher order structural correspondence that corresponded with one candidate inference, whilst the target and the other candidate base scenario, and their corresponding dictionary entries, shared structural correspondences, and a higher order structural correspondence that corresponded with the alternative candidate inference. Each stimulus set was divided into two subsets: in one, structural features in the bases and their novel term dictionary entries supported one set of inferences (Type A), whilst in the second sub-set, the same kind of matches supported the contrasting inference (Type B), so that biases towards a given inference could be eliminated (see figure 2).

### Experimental Hypothesis

In keeping with the analysis presented above, we expected that participants would use analogy to reconcile semantic terms in order to perform analogical mappings between the scenarios and generate support for one candidate inference. We predicted that in order to be able to carry out the top level analogy, participants would carry out another analogy in parallel - mapping structures only in the dictionary entries - reconciling semantic terms in a way that supported the top-level 'analogical' structure mapping over the top-level surface mapping, and favour the inference that corresponded to the structurally similar scenario over the scenario that shared only surface features.

### Additional Controls and Control Hypotheses

In addition to the basic stimuli, 3 sets of control stimuli were also created:

1 In the main control, the dictionary entries were eliminated, and participants were given only the target and the two candidate bases. In this control, in the absence of any structural support from the dictionary entries for the SSB inference, we expected participants to use the surface commonalities between the target and the SMB to determine their inference choice (i.e the prediction was that when subjects were asked to make an inference in a situation where neither of the base inferences benefitted from any structural bias, participants would prefer the inference which was additionally supported at the object level over the inference that received no such support; consistent with the findings of previous studies, such as Gentner, Ratterman and Forbus, 1993, we expected weak similarity to provide more support than no similarity).

2 In the second control, participants were given materials in which the novel terms were removed, and the structural information in the dictionary entries was added to the bases and target - in effect creating 'normal' analogy materials (see figure 5). In this control, where no re-representation was required, we expected the structural commonalities between the target and the SSB to determine the choice of inference, overriding the surface commonalities between the target and the SMB (this would be consistent with previous findings such as Gentner, 1983).

---

#### Chateau Bogusse:

**is a vineyard.**

**is in the southern French district of Pretence.**

**has sandy soils, with a lot of surface pebbles**

**has a warm microclimate which enables grapes to be produced.**

**the particular microclimate results in ripe grapes.**

*the ripeness causes the sugar level in the grapes to rise.*

*this makes the walls of the grapes weaken and collapse.*

#### Domaine Fraudulent:

grows plums.

has clay soils in which wildflowers grow

is in the western Departement of Maidoop.

its warm microclimate causes melons to grow

the particular microclimate yields extremely ripe plums.

*the extreme ripeness causes the plums to become very sweet*

*this super-sweetness makes the plums soft and squashy*

Because of their squashiness Domaine Fraudulent's plums are held in low esteem, and sell poorly.

#### Mas de la Fiction:

grows grapes.

**is in the southern Departement of Whaupper.**

**has sandy soils, with a lot of surface pebbles**

**its fine microclimate causes grapes to grow.**

**the particular microclimate results in ripe grapes.**

the ripeness causes some of the moisture in the grapes to evaporate

this evaporation leads to extremely concentrated flavours

Because of their concentrated flavours Mas de la Fiction's grapes are prized and sell for high prices.

#### Inferences

A. Chateau Bogusse's grapes are highly prized and sell for high prices.

B. Chateau Bogusse's grapes are held in low esteem, and sell poorly.

---

**Figure 5:** A control set in which the structural information in the dictionary entries has been included in the bases and target to create an 'ordinary' analogical problem. Surface similarities are illustrated in bold; structural similarities are italicised.

3 In the final control set the dictionary entries were altered so that surface and structural commonalities all supported the same mapping (the LSB, or literally similar base). In this final control, both structural and surface commonalities between the target and the LSB, and their dictionary entries were aligned in support of one inference. Since structure was predicted to be the key factor in deciding inferences (in line with the findings of previous studies), we did not expect the results from this control to differ significantly from the main experimental task.

In all of the controls, the inference supported by the various similarities was again randomised to control for any inherent biases towards particular inferences.

## Procedure

Participants were presented with 2 x 6-page questionnaires, each of which contained one scenario set, with its dictionary and candidate inferences, a diversionary task and a scenario set and pair of candidate inferences without a dictionary (the main control). The order in which the sets were presented ('with-dictionary' versus 'without-dictionary' control), was randomised, as was the presentation order of the targets within the sets. A second, smaller group of participants were given the other two controls in similar fashion.

Participants were asked to infer one candidate inference, and give a confidence rating (1=not at all confident; 5=very confident). They were told that the dictionary entries might be useful to them, but told explicitly that the use of them was left to participants' discretion.

## Results

Consistent with the initial hypothesis, in the main control condition where no dictionary entries were provided, the inference that received common surface-feature support was favoured by 67% of participants, with only 33% preferring the inference that was not supported by any commonalities,  $\chi^2(1, N=140) = 17.1, p < .001$ .

However, in the main experimental condition, where definitions – which offered the possibility of structural mappings – were provided, participants reversed their preferred inference for a given target / candidate bases set. Again consistent with the initial hypothesis, in this condition, if participants had preferred the A inference in the first control, when provided with scenario sets where structural commonalities in the dictionary supported the B inference, then participants now chose the B inference. Overall the inferences which received structural support were favoured by 71.2% of participants, with only 28.8% preferring the inference that was supported by surface commonalities alone,  $\chi^2(1, N=125) = 20.748, p < .001$ .

Also consistent with the initial hypothesis, in the control condition with no novel terms, where structure in the dictionary entries was included in the base and targets, inferences which received common surface-feature support were favoured by only 27.0% of participants, with 73.0% preferring the inference that was supported by structural commonalities,  $\chi^2(1, N=26) = 3.869, p < .05$ .

There was no deviation from this pattern in the final control condition, where the dictionary entries were altered so that surface and structural commonalities all supported

the same base - target (the LSB) mapping, the inference supported by the LSB was favoured by 75.0% of participants,  $\chi^2(1, N=28) = 5.17, p < .05$

Analysis of participants' confidence scores in the main control show significantly greater confidence for inferences based on surface commonalities when no structure was present,  $t=8.72, p < 0.001$ . However, this trend was reversed in the other controls and the main experiment - given the choice, participants seem to prefer structurally supported inferences. In the second control condition (analogies) inferences based upon structural commonalities received a significantly higher confidence rating than those based on surface features,  $t=3.982, p < 0.001$ . Similarly, in the main experimental condition, when definitions were provided, inferences based upon structural support received a significantly higher confidence rating than those based only on surface commonalities,  $t=2.9, p < .005$ . This trend was repeated in the third control, though mean differences were not significant,  $t=1.02, p = 0.33$ .

## Discussion

This experiment seems to show, consistent with the re-representation hypothesis, that participants can use the same process that they used to make analogical inferences to reconcile the semantic discrepancies they encounter in the representations of base and target analogs.

Participants made inferences with the aid of two targets. By controlling the structure of the information representing the 'semantics' of the term that was in turn crucial to the determination of the representation of higher order structure in the base and each of the targets, we were able to control the representations participants used for semantically reconciling particular terms during their analogising. The re-representation processing prediction – that participants would use the same mapping process to match semantic items in their representations as they would in ultimately determining their analogies – appears to be supported by the results of this experiment.

## General Discussion

Two very reasonable objections might be made to the results of this experiment:

1. Firstly, the 'dictionary entries' in the main task were artificial: there is a wealth of evidence that definitions are an inadequate basis for conceptual semantics (see Komatsu, 1992). Since the 'dictionary entries' are no more than definitions, it seems reasonable to question whether the use of definitions in exploring conceptual reconciliation affects the validity of our results.

2. A second obvious objection to the findings of the experiment is that participants were presented with the tasks on paper, and had unlimited time in which to solve the inferencing problems, and reconcile and map any 'semantics' in the various base and target specifications. It might be said in objection that since structure mapping is a computationally expensive process – especially in comparison to mapping surface features – this experiment has little relevance to the on-line demand characteristics of analogical processing 'in the wild'. Since participants in this experiment had unlimited time, and external representations of the problems, their behaviour is no predictor of the kind of processes used in making

analogical mappings in memory, where working memory limits will impose restrictions on processing.

Though we acknowledge our sympathy for these objections, neither of them should militate against our interpretation of these results: that the processing of re-representations is possible with externally represented problems. Obviously the second objection – which relates to internal representations – cannot apply to this interpretation. In respect of the first, we note that even though participants used what amounted to definitions in reconciling semantics in the main inferencing task, it is the *processing* that they used to map re-representations in semantic reconciliation (and not the particulars of the representations themselves) that is of interest here. To the extent that participants' processing matched our predictions (and what empirical findings there are in respect of natural representations in similar tasks, e.g. Ahn, 1998), it seems reasonable to assume that this processing was ecologically valid, even if the representations it worked with were not.

These objections do, however, highlight aspects of the re-representation hypothesis that are still seriously under-specified. In particular, the re-representation hypothesis lacks detail concerning the representations it supposes, and the processing demands it appears to make.

In the experiment above, we concentrated on the semantic reconciliation of *one* set of similar-but-not-identical terms, and followed this reconciliation process down through *one* level of recursion, where we saw – consistent with the re-representation hypothesis – that the same process was used to resolve semantic ambiguities as was used to determine analogical mappings.

However, it is unlikely that realistic representations of real-world analogies will contain such a small number of similar yet non-identical predicate matches to reconcile. These representations will contain many more such predicates, and the re-representations of these predicates – whose predicates will need to be matched during semantic reconciliation of the original predicates – may contain many more non-identical but semantically similar predicates, potentially as a factorial of the original number of predicates re-represented in semantic reconciliation.

Logically, at least, this seems to point to both a combinatorial explosion – in terms of the number of predicates to be reconciled, and hence individual semantic reconciliation sub-processes to be run – and a potential infinite regress: if an *identical* mapping process is to be run recursively, and if re-representation doesn't ultimately uncover *identical* predicate-decomposition representations at some level, then mapping may not terminate.

One solution to this problem might be the basic perceptual primitives posited by Hummel and Holyoak (1997; see above). We see two problems with this account: firstly, quite what 'perceptual primitives' are is unclear: at present, they offer no more explanatory clarity than 'concepts' when it comes to explaining semantic reconciliation; and secondly, and more worryingly, this proposal – like all re-representation hypotheses – assumes an almost unlimited capacity for structural mapping in memory. However, recent research (Halford, Wilson and Phillips, 1998) indicates that in reality this is far from the case: human capacity for representing and processing

structured information appears to be seriously constrained.<sup>3</sup> In the light of these considerations, we are cautious in inferring too much from the findings reported here. We have shown that re-representation is possible in externally represented tasks. Whether these results *can* be replicated in ecological analogy tasks in memory – and the extent to which re-representation is a viable psychological account of semantic reconciliation – remain open questions in need of further empirical investigation.

## Acknowledgements

Thanks to Lera Boroditsky, Ken Forbus, Dedre Gentner, Usha Goswami, Mark Keane, John Lee and Helen Pain for insightful discussions about this work. I'm grateful to Andrew Wishart, Dan Yarlett and Alex Heneveld for comments on an earlier draft of this paper.

## References

- Ahn W-K (1998) Why are different features central for natural kinds and artifacts? The role of causal status in determining feature centrality. *Cognition*, 69, 135-178.
- Forbus K, Gentner, D, & Law, K (1995) MAC/FAC: A model of similarity based retrieval *Cognitive Science* 19:141-205
- Forbus, K.D., Gentner, D., Markman, A.B., & Ferguson, R.W. (1998). Analogy just looks like high level perception. Why a domain-general approach to analogical mapping is right. *JETAI* 10, 231-257.
- Gentner, D (1983) Structure-mapping: a theoretical framework for analogy. *Cognitive Science*, 7: 155-170.
- Gentner, D, Ratterman, MJ & Forbus, K (1993) The roles of similarity in transfer. *Cognitive Psychology* 25: 524-575
- Goswami, U. (1992) *Analogical reasoning in children*. LEA, Mahwah, NJ
- Halford, G.S., Wilson, W.H. & Phillips, S. (1998) Processing capacity defined by relational complexity: implications for comparative, developmental and cognitive psychology. *Behavioral and Brain Sciences*, 21, 803-864
- Holyoak, KJ & Thagard, P (1995) *Mental Leaps*. MIT Press, Cambridge, Ma.
- Hummel, J.E. & Holyoak, K.J. (1997) Distributed Representations of Structure: A Theory of Analogical Access and Mapping. *Psychological Review*, 104, 427-466.
- Komatsu, L. K. (1992): Recent views of conceptual structure. *Psychological Bulletin*, 112, 500-526.
- Ramscar, M.J.A. and Hahn, U. (1998) What family resemblances are not. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, LEA, Mahwah, NJ 865-870
- Ramscar, M.J.A. and Pain, H. (1996) Can a real distinction be drawn between cognitive theories of analogy and categorisation? In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, LEA, Mahwah, NJ 346-351

---

<sup>3</sup> So far we have only been concerned with re-representation in relation to the mapping process: these problems will multiply massively in relation to the semantic reconciliations that must be made in order to facilitate retrieval.

# Searching for Alternatives in Spatial Reasoning: Local Transformations and Beyond

Reinhold Rauh<sup>1</sup> (REINHOLD@COGNITION.IIG.UNI-FREIBURG.DE)

Cornelius Hagen<sup>1</sup> (HAGEN@COGNITION.IIG.UNI-FREIBURG.DE)

Christoph Schlieder<sup>2</sup> (CS@INFORMATIK.UNI-BREMEN.DE)

Gerhard Strube<sup>1</sup> (STRUBE@COGNITION.IIG.UNI-FREIBURG.DE)

Markus Knauff<sup>1</sup> (KNAUFF@COGNITION.IIG.UNI-FREIBURG.DE)

<sup>1</sup> Center for Cognitive Science, Institute of Computer Science and Social Research, University of Freiburg  
Friedrichstr. 50, 79098 Freiburg i. Br., Germany

<sup>2</sup> Department of Mathematics and Computer Science, University of Bremen  
P. O. Box 330440, 28334 Bremen, Germany

## Abstract

Searching for alternative solutions of an indeterminate reasoning task is an important and necessary step in order to draw certain inferences as in the case of deduction. To elucidate the underlying mental representations and processes of the search for alternatives in spatial reasoning, an experiment was conducted that used specific material stemming from AI research of Qualitative Spatial Reasoning. The results showed that searching for alternative solutions can be best explained as a revision process starting with an initial mental model of the premises. Proceeding from one solution to an alternative is apparently achieved by local transformation. Interestingly, local transformations have a "logic of their own": They can lead to systematic errors of omission and to errors of commission.

## Spatial Reasoning and Mental Models

Dealing with spatial problems is a frequent and important challenge in everyday as well as in professional life. It occurs across various fields like spatial navigation or spatial configuration and design. In this paper, we will concentrate on a special sort of spatial problem solving, namely reasoning based on spatial relational descriptions. This type of reasoning can be investigated with recourse to several background theories of thinking developed in cognitive psychology. According to previous research in spatial reasoning (Byrne & Johnson-Laird, 1989; Evans, Newstead, & Byrne, 1993) and according to our own previous findings (Knauff, Rauh, & Schlieder, 1995; Knauff, Rauh, Schlieder, & Strube, 1998; Rauh & Schlieder, 1997) the most promising and most successful framework is the theory of mental models.

## Mental Model Theory as Framework

The core assumption of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) states that when we reason we build an integrated representation of the situation that the premises describe. This integrated representation—the mental model—is in certain aspects analogous to the state of affairs and, as a consequence, lacks the information whether relationships are explicitly mentioned in the

premises and or are implicitly determined by the representational format.

A further consequence of the assumption of integrated representation becomes evident when certain kinds of inferences have to be drawn. Take deductive inference for example: To test whether a contingent relationship in the initial mental model is necessarily true, the reasoner has to test all the alternative models of the premises. If a contradictory example is found, the putative conclusion will be rejected; if not it will be accepted as a valid conclusion.

The search for alternative models takes place during what we call the phase of model variation. It seems to be a deliberate mental process so fragile that it causes many systematic reasoning errors. There are errors of omission, i.e. inferences that could have been validly drawn, and there are errors of commission, i.e. inferences that are not justified by the premises.

Therefore, model variation has attracted much attention, but little is empirically known about how the mental search for alternative models is accomplished by the human process of reasoning. For a precise investigation of the model variation phase, there is the need for relational material with a rich inherent structure and unambiguous semantics.

## Spatial Reasoning with Interval Relations

Traditional investigations of spatial reasoning used relations like *left-of*, *right-of*, *in front of*, and *behind*. As argued elsewhere (Knauff et al., 1998), these spatial relations have no clear semantics. Therefore, studies of reasoning using these spatial relations are problematic because it is unclear whether the results obtained can be attributed to the inference processes, or are due to the ambiguity of these relations. To remedy this situation, we use Allen's (1983) set of 13 qualitative interval relations that enables one-dimensional spatial reasoning. These relations have clear geometric semantics based on the bounding points of the intervals, i.e. their starting points and ending points. They also have the property of being jointly exhaustive and pairwise disjoint (JEPD)—a property that also reduces the risk of misinterpre-

tations. In Table 1, we shortly introduce these relations together with verbalizations that we use in our experiments.

Table 1: The 13 qualitative interval relations, associated natural language expressions, and a graphical example (adapted and augmented according to Allen, 1983).

Relation symbol	Natural language description	Graphical example
$X < Y$	X lies to the left of Y	
$X m Y$	X touches Y at the left	
$X o Y$	X overlaps Y from the left	
$X s Y$	X lies left-justified in Y	
$X d Y$	X is completely in Y	
$X f Y$	X lies right-justified in Y	
$X = Y$	X equals Y	
$X fi Y$	X contains Y right-justified	
$X di Y$	X surrounds Y	
$X si Y$	X contains Y left-justified	
$X oi Y$	X overlaps Y from the right	
$X mi Y$	X touches Y at the right	
$X > Y$	X lies to the right of Y	

With these relations, reasoning tasks known as three-term series problems can be constructed. One example is "X overlaps Y from the left. Y surrounds Z." The example also shows that there are many three-term series problems generated from these relations that have more than one solution. To be precise, there are 42 three-term series problems that have three solutions, 24 that have five solutions, 3 that have nine, and another 3 that have thirteen solutions. We utilize this property in order to construct indeterminate three-term series problems to investigate precisely the phase of model variation. In the next section, we will present a more formal analysis of these tasks. From this analysis and the revealed properties of the different tasks, hypotheses can be derived that we will test in a model variation experiment.

## A Formal Framework for Model Variation

In principle, there are two ways to construct alternative models of the premises. The first consists of repeating the complete construction of alternative models one after another (*model iteration*). We will examine the more plausible varia-

tion strategy that consists of generating alternative models by *locally* transforming the *initial model* (see also Schlieder, 1998), i.e. the first model constructed during model variation (*model revision*).

In this view, any sequence of models  $M_0, M_1, \dots, M_n$  corresponds to a *sequence of transformations*  $T_1, T_2, \dots, T_n$ , where the output model  $M_i$  of  $T_i$  is the input model of  $T_{i+1}$ . The set  $\{M_0, \dots, M_n\}$  is *ordered* by the sequence  $T_1, T_2, \dots, T_n$ .

Since models of a three-term-series problem are completely determined by only one relation, namely the one between X and Z, we can treat models and relations equivalently. Seen this way, a transformation is a transition from one relation  $r_1$  to another relation  $r_2$ , or, in short,  $r_1 \rightarrow r_2$ .

## Conceptual Neighborhoods

Freksa (1992) introduced the notion of conceptual neighborhood between interval relations. Formally, the three conceptual neighborhoods are defined by the graphs in Figure 1. Two relations are *neighbors* iff they are connected by an edge of the corresponding graph.

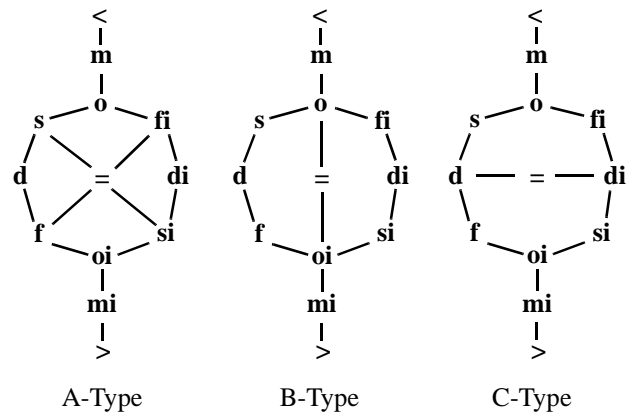


Figure 1: Freksa's (1992) conceptual neighborhoods.

The common generic principle underlying the three types of neighborhood reads as follows: Interval relations  $r_1$  and  $r_2$  are said to be *conceptual neighbors* if a model of intervals X and Y satisfying  $X r_1 Y$  can be continuously transformed into a model of intervals X' and Y' satisfying  $X' r_2 Y'$  such that during the transformation no model arises in which a relation different from  $r_1$  and  $r_2$  holds (see Schlieder & Hagen, in press). Their peculiarities arise from different transformation processes. The A-neighborhood is based on a transformation that can be described as the movement of one single bounding point of one interval whereas the B-neighborhood relies on the movement of a complete interval of fixed length. The transformation defining the C-neighborhood consists of keeping the center of the changing interval fixed and varying its length. The types of transformations defining the A(B,C)-neighborhoods will be called *A(B,C)-transformations*.

## Local Transformations: Steps between A-Neighbors

An examination of sequences of A-transformations revealed a need to formally refine the conceptual framework. In order to describe the model revision process adequately the definition has to include the movement of bounding points and its direction. An A-transformation between intervals X and Y does not specify the moving bounding point since it can always be accomplished in two ways by movements of a suitable bounding point: Either by moving one bounding point of interval X in one direction or one of Y in the opposite direction (see Table 2). An A-transformation with specified moving point  $p$  will be called a *step (of bounding point  $p$  in direction  $d$ )*.

Table 2: The relation of A-transformations and steps.

A-transformation	step right	step left
$< \rightarrow m$	$E_X$	$S_Y$
$m \rightarrow o$	$E_X$	$S_Y$
$o \rightarrow fi$	$E_X$	$E_Y$
$fi \rightarrow di$	$E_X$	$E_Y$
$di \rightarrow si$	$S_X$	$S_Y$
$si \rightarrow oi$	$S_X$	$S_Y$
$oi \rightarrow mi$	$S_X$	$E_Y$
$mi \rightarrow >$	$S_X$	$E_Y$
$o \rightarrow s$	$S_X$	$S_Y$
$s \rightarrow d$	$S_X$	$S_Y$
$d \rightarrow f$	$E_X$	$E_Y$
$f \rightarrow oi$	$E_X$	$E_Y$
$s \rightarrow =$	$E_X$	$E_Y$
$= \rightarrow f$	$S_X$	$S_Y$
$= \rightarrow si$	$E_X$	$E_Y$
$fi \rightarrow =$	$S_X$	$S_Y$

Note that tracking sequences of interval relations does not permit the direct observation of steps. *Step-sequences*, i.e. sequences of steps that refer to the same point  $p$  moving in constant direction  $d$ , can explain errors of omission or commission that cannot be explained on the level of A-transformations. In order to show this, we need one more definition. A step-sequence  $S_1, \dots, S_n$  is *extendible at the beginning (or the end)* iff there exists a step  $S_0$  (or  $S_{n+1}$ ) such that  $S_0, S_1, \dots, S_n$  (or  $S_1, \dots, S_n, S_{n+1}$ ) is a step-sequence. If it is extendible at the beginning or at the end it is (*totally*) *extendible*.

**Errors of Omission and Errors of Commission.** Our general assumption about the implications of this formalism for the traversal of solution sets is as follows: Moving along a step-sequence, i.e. keeping the moving point and its direction constant, is easier to process than changing them or even performing a non-A-transformation.

Therefore, errors of omission should be observed more frequently if the end of a step-sequence is reached but the solution set is not completely traversed. Errors of commission, in turn, should occur more frequently with non-solutions which are a continuation of a step-sequence.

## Hypotheses

In the following we present hypotheses specifying the implications of the above considerations in more detail. They can easily be verified consulting Table 2 and Figure 2, which displays solution sets of all three-term series problems with multiple models.

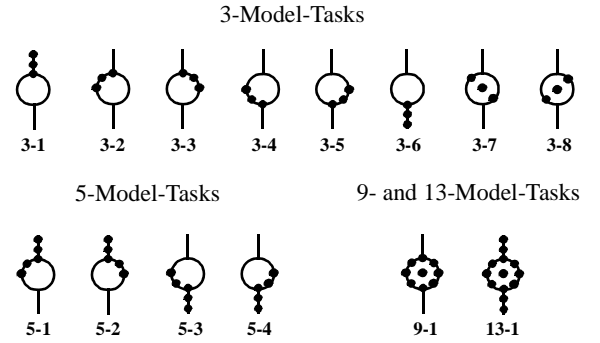


Figure 2: The solution sets of three-term-series problems with multiple models. The valid relations are represented as points at corresponding positions of Figure 1.

**3-Model-Tasks.** The relations determining the solution set of a 3-model-task can be ordered in two ways by sequences of A-transformations (e.g. for (3-1):  $< \rightarrow m \rightarrow o$  or  $o \rightarrow m \rightarrow <$ ). Each of these sequences can be accomplished in two ways as step-sequence (e.g.  $< \rightarrow m \rightarrow o$  by steps to the right of the ending point  $E_X$  of interval X or by steps to the left of the starting point  $S_Y$  of Y). One of these sequences is extendible except for the solution sets (3-7) and (3-8) where all sequences are non-extendible. There are two interesting hypotheses concerning 3-model tasks: (1) 3-model-tasks having extendible solution sequences are prone to errors of commission, and (2) 3-model-tasks with solution sets (3-7) and (3-8) have significantly less errors of commission than the other 3-model-tasks.

**5-Model-Tasks.** The solution set of a 5-model-task can be ordered in two ways by sequences of A-transformations. Each of these sequences can be accomplished in two ways, as step-sequence that is non-extendible, or as a sequence  $S_1, S_2, S_3, S_4$ , where  $S_1, S_2$  and  $S_3, S_4$  are non-extendible step-sequences, having the same direction but referring to different



bounding points of the same interval. Accordingly, we can formulate the hypothesis, that errors of omission will most frequently occur between step 2 and step 3.

**9-Model-Tasks and 13-Model-Tasks.** The solution set of a 9-model-task or of a 13-model-task can be ordered in multiple ways by sequences of A-transformations. Each of them fall into several step-sequences, including necessary changes of direction between them. So we expect a decreased number of correct and complete solution sequences for these tasks.

## Experiment on Model Variation

### Participants

24 students (12 female, 12 male) of the University of Freiburg were paid for participation.

### Materials

The material consisted of the 72 indeterminate three-term series problems that can be constructed by the 12 interval relations, if the trivial "=" relation is omitted. In each three-term series problem the spatial relationship between a red and a green interval is described in the first premise, and the relationship between the green interval and a blue one is given in the second premise.

### Procedure

The computer-assisted experiment was divided into three phases. During the *definition phase* participants were given the verbalizations of the interval relations together with an explanation of the semantics with respect to the ordering of starting points and ending points. Additionally, a pictorial example was displayed.

During the *learning phase*, participants read sentences describing the relation between a red and a blue interval. For each sentence they had to specify the relationship of the two intervals graphically by clicking the mouse in rectangular regions on the screen. After having confirmed the final choices, the participant got feedback about the accuracy of the configuration. If the configuration did not match the relation, additional information about the correct answer was given, i.e. a verbal description of the ordering of start points and end points. Learning trials were blocked with 13 sentences using the interval relations. If one relation was answered correctly in three consecutive blocks, the learning criterion for this relation was accomplished. As soon as the learning criterion was reached for all relations, the learning phase stopped.

During the *inference phase*, participants were given 3 practice trials, and then received the 72 indeterminate three-term series problems. After self-paced reading of the premises, the premises vanished, and the participants were asked to generate all possible relationships between the red and the blue interval. By clicking the mouse they specified

the spatial relationships analogous to the interval-specifying procedure in the learning phase. After finishing the configuration, participants could either continue specifying other solutions, or stop working on the present task and go to the next three-term series problem.

We recorded premise processing times, drawing times, and, of course, the sequence of solutions by pixel coordinates and by interval relations.

### Results

In the following, data analyses are applied to the constructed solution sequences. Since all participants passed the learning phase successfully, all data collected in the inference phase were included in the statistical analyses.

First, we tested the hypothesis that solution sequences followed the principles of conceptual neighborhood. All transitions in the solution sequences were analyzed for the existence of A-, B-, and C-transformations. We found that the significant majority of the transitions (3145 of 4462 [= 70.48%]) conformed to A-transformations. Transitions conformed to B- or C-transformations in 64.95% or 64.34% of all the cases, respectively. The three values are rather similar, since most transitions are consistent with all three types of conceptual neighborhood. Only transitions involving the "=" relation discriminate between different types of conceptual neighborhood (see Figure 1). Therefore, we performed an analysis for these transitions and found the frequencies listed in Table 3.

Table 3: Number of "="-transitions conforming to different types of conceptual neighborhood.

	Absolute	Percent
A-transformation	296	75.13%
B-transformation	49	12.44%
C-transformation	22	5.58%
Other	27	6.85%
Total	394	100%

We obtained the results in Table 4 by exclusively analyzing correct and complete solution sequences of 3-, 5-, 9-, and 13-model tasks.

The interesting fact is the nearly monotonic decrease of the number of correct and complete solution sequences in dependence of the number of models. Besides, it is noteworthy that correct and complete sequences of the 9- and 13-model problems (i) are rarely observed (as predicted by our hypothesis), and (ii) that none of these sequences conformed perfectly to any of the neighborhood transformations. We will return to the latter point below.

Table 4: Number of correct and complete solution sequences

	Percent	A-Transf.
3-model-tasks	52.88% (533 of 1008)	75.61% (403 of 533)
5-model-tasks	34.20% (197 of 576)	86.29% (170 of 197)
9-model-tasks	13.89% (10 of 72)	0% (0 of 10)
13-model-tasks	16.67% (12 of 72)	0% (0 of 12)
Total	43.52%	73.27%

**Errors of omission.** To test for the hypothesis of systematic errors of omission between step 2 and step 3, we looked at the solution generated last in the whole solution sequence for all 5-model-tasks. In Table 5 the results for the six 5-model-tasks with solution set (5-2) (see Figure 2) are listed. As stated above, we expected an increasing number of solution sequences terminating after the second step, i.e. for relation  $o$ .

Table 5: Frequencies of relations as last solution for 5-model tasks with solution set (5-2).

di	fi	o	m	<
10	7	22	8	87
7.46%	5.22%	16.42%	5.97%	64.93%

As Table 5 shows, there are indeed many solution sequences terminating with the relation  $o$  (22 of 134). This pattern of results was also obtained for the 5-model-tasks with the other three solution sets. The result confirms our predictions of systematic errors of omission between steps 2 and 3.

**Errors of commission.** According to our predictions of systematic errors of commission, the 3-model-tasks with solution sets (3-1) to (3-6) were analyzed for transitions from relation  $o$  ( $oi$ ) followed by an erroneous one. The number of such transitions was 57. It turned out that 26 of them were steps with the  $o$  ( $oi$ ) relation as precursor. Given that there are at least 8 other erroneous relations that are not A-transformations of  $o$  ( $oi$ ), this shows that the transition from a correct solution to an erroneous one is about three times more probable if the erroneous solution is the next step in the step-sequence. The result corroborates our hypothesis of systematic errors of commission. Additionally, the 3-model-tasks with solution sets (3-7) and (3-8) had 13.5 commission errors on the average, much less than the 72.0 commission errors that could be observed on the average for the 3-model tasks with solution sets (3-1) to (3-6).

**Strategies for 9- and 13-Model-Tasks.** As shown in Table 4 none of the correct and complete solution sequences of the 9-model-tasks and the 13-model-tasks conformed perfectly to any of the conceptual neighborhood transformations. In an exploratory data analysis, we identified two classes of strategies for navigating through the solution set that guided the successful search for alternatives in solving 9- and 13-model-tasks.

*Constant-Direction-Strategies.* The first class of strategies consists of three sequences of A-transformations following one after another. The two transformations joining them are not A-transformations, but *jumps* in the graph of the A-neighborhood. (see the diagram in Figure 3)

As the pseudo code description in Figure 3 shows this strategy can be accomplished in a simple way: All steps refer to points of the same interval and proceed with the same direction. For each step the other bounding point of the interval is tested if a step leads to a valid model, and the information determining this model is stored if necessary. The jumps occur only if proceeding within a step-sequence is not possible. Then the stored information is retrieved again to construct the corresponding model to begin the next step-sequence.

The success of this kind of strategy depends highly on the choice of the initial model since the moving direction is constant and an omitted model will never be reached.

Choose an initial model;  
 Choose an interval (with bounding points  $p$  and  $q$ ) that is part of the relation between the first and the third interval;  
 Choose  $p$  and direction  $d$  such that  $step(p, d)$  possible;

```

while step(p, d) or step(q, d) possible
begin
  if step(p, d) possible then
    begin
      if M empty and step(q, d) possible then
        Store info identifying the result of step(q, d) in M;
        step(p, d);
      end
    else
      begin
        if M not empty then
          Continue with the model
            identified by M;
        else if step(q, d) possible then
          step(q, d);
        end
      end
    end
end
  
```

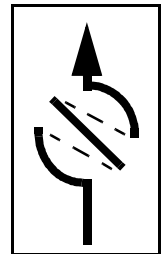


Figure 3: Constant-Direction strategies as pseudo code and a diagram of a possible path in the Freksa-graph. Details of the algorithm are specified only as far as necessary;  $step(p, d)$  represents a step-transformation of  $p$  in direction  $d$ ,  $M$  information identifying a model.

*Symmetry-Strategies.* The second class of strategies is based on the use of symmetric transformations mapping relations to their inverses (*transposition-symmetry*). Their limitations and strengths concerning the traversal of the solution set arise from the fact that the solution sets of 9- and 13-model tasks fall into several disjointed subsets that are closed in relation to symmetry-transformations. An extended version involves additional *reorientation-symmetry*. This type of symmetry can be described as reflection of the graphical example in Table 1 at the vertical axis. All relations are symmetrical to themselves with respect to reorientation except the pairs  $f$ - $s$  and  $f$ - $si$ . In place of the closed subsets  $\{f, fi\}$  and  $\{s, si\}$  their union now forms a closed subset.

For the traversal of the solution set of a 13-model-task following the extended type of strategy this implies that at least 5 non-symmetric transformations (out of a total of 12 necessary transformations) are needed to traverse all relations. A 9-model-task needs at least 3 non-symmetric transformations (out of a total of 8). The type of strategy that relies only on transposition requires one more non-symmetric transformation. Especially for 13-model-tasks we cannot expect complete solutions without an additional guiding principle. Furthermore, errors in finding a closed subset will lead to omitting it completely. On the other hand due to the cyclic structure of a closed subset, its traversal is insensitive to the first relation established.

## General Discussion

In summary, the presented results corroborate the assumption that searching for alternatives is based on a model revision process proceeding from an initial model to alternatives by local transformations. We demonstrated and specified this for one-dimensional spatial reasoning, where local transformations appear as movements of a point along a step-sequence. Additionally, we were able to show that local transformations have a logic of their own: They can systematically suppress certain inferences on the one hand, but, on the other hand, lead to false ones. Again, we specified these conditions with the help of our relational material, and thus were able to predict errors of omission and errors of commission precisely. This point is also very important for augmenting our existing cognitive modeling of mental model construction with an empirically adequate revision process.

With respect to psychological theories of reasoning, our results are pretty much in accordance with the mental model theory. In particular, the decline of number of correct and complete solution sequences with the number of models corresponds well with mental model theory assumption that the difficulty of a reasoning task is dependent on the number of models. Likewise, the notion of local transformation only makes sense with recourse to analog representations, e.g. mental models. Therefore, our data also present a new challenge for other theories of reasoning.

## Acknowledgements

This research was supported by the German National Research Foundation (Deutsche Forschungsgemeinschaft; DFG project *MeMoSpace* under contract no. *Str 301/5-1* to Gerhard Strube, Christoph Schlieder, and Reinhold Rauh). We would like to thank Goran Sunjka for his extensive help in implementing the computer-aided experiment, Katrin Balke for running the experiment, and Rebecca Ellis and Patrick Mueller for proof reading an earlier draft.

## References

- Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26, 832-843.
- Byrne, R. M. J., & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory and Language*, 28, 564-575.
- Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning. The psychology of deduction*. Hove (UK): Lawrence Erlbaum Associates.
- Freksa, C. (1992). Temporal reasoning based on semi-intervals. *Artificial Intelligence*, 54, 199-227.
- Johnson-Laird, P. N. (1983). *Mental models. Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. N. & Byrne, R. M. J. (1991). *Deduction*. Hove(UK): Lawrence Erlbaum Associates.
- Knauff, M., Rauh, R., & Schlieder, C. (1995). Preferred mental models in qualitative spatial reasoning: A cognitive assessment of Allen's calculus. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 200-205). Mahwah, NJ: Lawrence Erlbaum Associates.
- Knauff, M., Rauh, R., Schlieder, C., & Strube, G. (1998). Mental models in spatial reasoning. In C. Freksa, C. Habel, & K. F. Wender (Eds.), *Spatial cognition. An interdisciplinary approach to representing and processing spatial knowledge* (pp. 267-291). Berlin: Springer.
- Rauh, R., & Schlieder, C. (1997). Symmetries of model construction in spatial relational inference. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 638-643). Mahwah, NJ: Lawrence Erlbaum Associates.
- Schlieder, C. (1998). Diagrammatic transformation processes on two-dimensional relational maps. *Journal of Visual Languages and Computing*, 9, 45-59.
- Schlieder, C., & Hagen, C. (in press). Interactive layout generation with a diagrammatic constraint language. In C. Freksa, C. Habel, & K. F. Wender (Eds.), *Spatial cognition II*. Berlin: Springer.

# Memory for Continually Changing Information: A Task Analysis and Model of the Keeping Track Task

Wolfgang Schoppek (wschoppe@gmu.edu)  
Human Factors and Applied Cognition Program  
George-Mason University  
Fairfax, VA 22030-4444, USA

## Abstract

Keeping track of continually changing information has been investigated since Yntema & Mueser's (1960) seminal work. The fact that types of mappings between objects and values and of memory load affect performance are well established, but have never been integrated in a theory. As a step toward such a theory, this paper describes a mathematical model that combines a task analysis with a set of assumptions derived from the ACT-R theory about the dynamics of memory traces. The model's remarkable reproduction of data published by Venturino (1997) demonstrates that standard memory concepts are sufficient to explain the results related to this paradigm. The model yields a clear implication about what causes interference and helps specify open questions.

In many areas of supervisory control, operators have to keep track of the changing values of a number of variables. Knowing the current state of a dynamic system is an important component of situational awareness (Endsley, 1995). For example, a pilot flying a modern automated aircraft needs to know the current altitude, speed, and course of the aircraft, the current settings and modes of the flight management system, just to mention a few of the variables.

In the experimental paradigm for keeping track of continually changing information, introduced by Yntema and Mueser (1960), object-value pairs are presented successively, interrupted by queries about the value associated with a certain object. The most common variables manipulated are the number of objects and the number of attributes from which the values are selected.

In Yntema and Mueser's (1960) experiment, subjects either had to keep track of changing values of many attributes for one object or changing values of the same attribute for many objects. Memory performance was worse in the latter condition. This was attributed to a high degree of interference when only one attribute is used.

Venturino (1997) argued that Yntema and Mueser (1960) confounded attribute similarity and information organization. Figure 1 illustrates how the former factor is defined by the number of attributes, the latter defined by the number of objects. In order to investigate the relative influence of the two factors on memory performance,

Venturino (1997) completely crossed these two factors, such that all four possible combinations between high and low attribute similarity and high and low information organization were included. A third factor was memory load. Attribute similarity had a large effect on memory performance, which confirmed Yntema and Mueser's (1960) findings. Information organization also had a significant effect, but this effect was much weaker. As expected, performance declined with memory load.

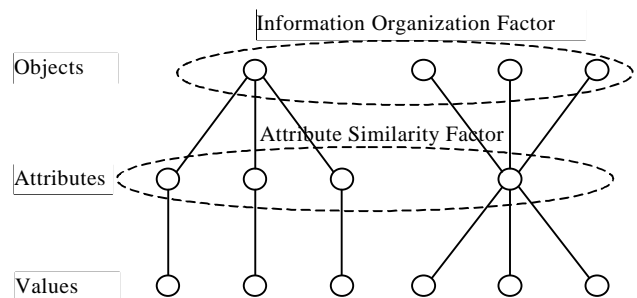


Figure 1: Illustration of the relations between objects, attributes, and values in the paradigm of continually changing information

This same paradigm was used by Hess, Detweiler and Ellis (1999) to prove the superiority of spatially rich displays over displays that show values of different attributes in the same location. Although their research goal was different from Yntema and Mueser's, the basic findings of the paradigm were confirmed in these experiments.

To summarize, the effects of attribute similarity and of memory load are well established. Although the main effects can be explained through the interference that occurs between values of the same attribute, the interactions between attribute similarity and memory load are understood less well. There is no integrative theory that accounts for all the effects. Venturino (1997) interpreted his results as suggesting a distinction between memory capacity for static information and memory capacity for dynamic information, because memory performance in

the same-attribute condition was worse than what would be expected in a comparable static memory task.

The goal of this work is to explore if the results about keeping track of dynamically changing information can be explained more parsimoniously with standard assumptions about memory. As a means for this exploration, I developed a mathematical model of the experiment by Venturino (1997). The model combines a task analysis with a set of assumptions about the dynamics of memory traces that are derived from the ACT-R theory (Anderson & Lebiere, 1998). The model may also contribute to an integrated understanding of all the effects related to the paradigm.

In the following sections, I first describe Venturino's experiment in more detail before I present and discuss the model.

### Venturino's Experiment

The material used in the experiment consisted of the names of six different fire engines and six different attributes with six values each. Continually changing attribute values were assigned to the fire engines. The task was to memorize these values. After a series of five to seven updates, the subject was asked for the current attribute value of a certain fire engine. For example, in keeping track of the current values of two fire engines, a subject might have to keep track of the number of firefighters for a pumper engine and the location of a tanker engine.

This continual updating is shown with a detailed example in Table 1. Time is represented in discrete steps, where 1 denotes the time of the most recent update, 2 the time step before, and so on. I will refer to these steps as lag, indexed by the variable *i*.

Table 1: Illustration of the continual updating of values (asterisks indicate an updating event)

lag time	stimuli		current value of		
	fire engine	n fire-fighters	tanker	ladder	pumper
...	...	...	...	...	...
4	tanker	4	*4	...	...
3	ladder	7	4	*7	...
2	tanker	5	*5	7	...
▼1	pumper	4	5	7	*4
now					

The example shows tanker being updated with the value four at lag 4, ladder being assigned the value seven at lag 3, tanker updated with the value five at lag 2, and pumper being assigned the value four at lag 1. Every fire engine keeps its value until it is updated. These update events are indicated by asterisks in Table 1. The three current values "now" are five firefighters for tanker, seven firefighters for ladder, and four firefighters for pumper. Note that the values differ in "age".

Three independent variables were manipulated in the experiment: number of objects (one vs. many fire en-

gines), attribute similarity (same vs. different attribute), and memory load (two, four or six values to keep track of). The first two factors were varied between subjects; the last factor was varied within subjects. In the many-object/different-attributes condition, unique mappings between objects and attributes were used, such that each of the two, four, or six engines had a value of a different attribute. In the many-objects/same-attribute condition, two, four, or six fire engines had multiple values of the same attribute. In the one-object/different-attributes condition, one engine had values of two, four, or six attributes. In the one-object/same-attribute condition, one fire engine had a value of one attribute. In order to manipulate memory load in this condition, subjects had to memorize the history of the last two, four, or six values. Despite the different mappings, the same number of values had to be remembered in each memory load condition.

Each block began with an initialization of values, followed by 75 to 105 updates, presented at a rate of one update each seven seconds. The updates were randomly interrupted by 15 queries. There were 100 subjects total, randomly assigned to one of the four conditions. In a first session, subjects studied the experimental material and, after a few practice trials, worked on the block with memory load 2. Two days later, the blocks with memory load 4 and 6 were administered.

Performance was measured as the proportion of correct answers. The outlined markers of Figure 3 illustrate the main results. All three independent variables had significant main effects on performance, but they were differently strong. Attribute similarity accounted for 15% of the variance, information organization (number of objects) for only 1%. The main effects were qualified by a significant three-way interaction of all factors. Separate analyses revealed significant interactions between attribute similarity and memory load in both object conditions: Memory load affects performance much more when the same attribute is used than when different attributes are used.

In the same-attribute condition, there was a significant interaction between memory load and number of objects: In the many-object condition performance decreased more sharply as memory load increased than in the one-object condition. In the different-attribute condition, the number of objects had no significant effect on performance.

An error analysis revealed that 44% of the errors were previous state errors, i.e. a subject responded with the previous value of an attribute rather than its current value. Interestingly, subjects responded significantly faster ( $M = 4.58$  s) when making a previous state error than when making any other type of error ( $M = 5.30$  s).

### Model

In this section, a model will be described that is able to reproduce the results of Venturino's experiment. The predictions of the model are not derived from simulation, but from a mathematical combination of the probabilistic

Table 2: Task analysis of the keeping track task

lag $i$	fire-engine	value $v_i$	$v_i$ current now?	$p$ ( $v_4$ current after lag $i$ )	$p$ ( $v_3$ current after lag $i$ )	$p$ ( $v_i$ current now) = $q_i$
4	tanker	4	no	1	...	$0.75^3$
3	ladder	7	yes	0.75	1	$0.75^2$
2	tanker	5	yes	$0.75^2$	0.75	0.75
1	pumper	4	yes	$0.75^3$	$0.75^2$	1
<b>now</b>						

structure of the material and basic assumptions about the dynamics of memory elements. The psychological assumptions originate from the ACT-R theory (Anderson & Lebiere, 1998).

Suppose that each update event is stored as a unique memory trace. The probability that this trace contributes to a correct answer equals the probability that the trace represents a current value times the probability that it is retrieved from memory. The first factor is given by the task analysis described below, the second factor is derived from a cognitive model. Summing up the probabilities of contributing to a correct answer for all memory traces gives an estimate of the number of correct answers for all possible probes.

### Task Analysis

The first component of the model is an analysis of the probabilistic structure of the material used in the experiment. This task analysis allows us to determine the probability that a value is current as a function of the update time and the memory load condition.

Table 2 is built on the example given in Table 1 and contains information that is relevant to understanding the task analysis. Time is again indicated by lag. The values that were presented at each time step are referred to as  $v_i$ . Column 4 contains the "currency" of the respective values  $v_i$  at present time (now), i.e. immediately after lag 1. The values  $v_1$ ,  $v_2$ , and  $v_3$  are still current, but  $v_4$  is not, because it was overwritten with  $v_2$ . Column 5 shows the probability of  $v_4$  being current at the end of each time step. At the end of lag 4,  $v_4$  is current (probability equals 1.0), because it has just been updated. At lag 3, one of the four vehicles is randomly chosen for an update. Thus, the probability of  $v_4$  being updated at lag 3 is 0.25. Put another way, the probability of  $v_4$  being current at the end of lag 3 is  $1 - 0.25 = 0.75$ . The same considerations hold for the following steps.

Because the updates are independent events, the probabilities for each time step must be multiplied to obtain the overall probability that a value is still current. Thus, the probability of  $v_4$  being current after lag 1 ("now") is  $0.75^3$ . Column 6 exemplifies that for the update of "ladder" at lag 3. The last column of Table 2 contains the resulting probabilities of being still current for  $v_1$  through  $v_4$ . Equation 1 is the generalized form of the probability  $q_i$  of value  $i$  still being current.

$$q_i = p_s^{i-1} \quad (1)$$

In Equation 1,  $p_s$  is the probability of not being updated in the following step. This variable depends on the memory load  $n_c$  (i.e. number of current values given by the number of vehicles and/or attributes), according to Equation 2.

$$p_s = 1 - 1/n_c \quad (2)$$

Applying Equations 1 and 2 to Venturino's experimental materials results in the probabilities depicted in Figure 2. Each memory load condition results in one curve. Memory load condition 6 involves six current values, distinguished by the type of vehicle, the attribute, or a unique mapping between vehicle and attribute. Similarly, memory load conditions 4 and 2 involve four and two values, respectively. It is obvious that the probabilities of being current diminish much faster the fewer current values there are, because the probability for each value being updated is higher when there are fewer dimensions (attributes and/or objects).

The task analysis also reveals that the probabilistic structure of the one-object/same-attribute condition deviates considerably from this scheme. Because in this condition, the last two, four, or six values of the same attribute have to be remembered for only one object, the probabilities of these values being current are one, the probabilities of all other values are zero. This different structure was entered at the appropriate places in order to calculate the model's prediction.

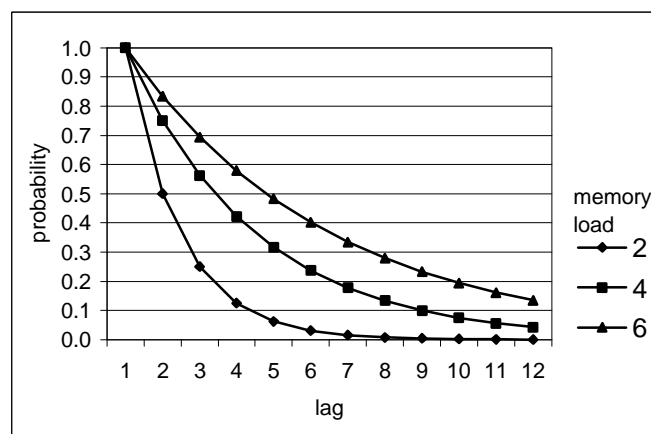


Figure 2: Probabilities  $q_i$  that a value that was updated in a certain time step is still current.

## Cognitive Model

The second component of the model is a set of assumptions about the dynamics of the memory representations that are formed from the update events. The first assumption is that for each update event a new memory element is created which represents the information given in the update. The second assumption is that each element is rehearsed a number of times, thus being strengthened. The remaining assumptions are part of the ACT-R theory.

According to the rational analysis basis of ACT-R, the activation of a declarative memory element reflects the probability that the element is needed in the current context and determines its retrieval. The two additive components of activation are baselevel activation and net activation. The former reflects the baserate probability, the latter the conditional probability given the current context. In this application, current context means the cues that are active and enhance retrieval of the correct memory element. Since this model makes no specific assumptions about cues, we can focus on baselevel activation.

The baselevel activation of an element is defined as the log odds that the element is needed. The odds are calculated with Equation 3, where  $n$  is the number of times the element has been needed, and  $L$  is the lifetime of the element<sup>1</sup>. Lifetime is the time that has passed since the creation of the element. The more frequently a memory element has been needed in its lifetime, the higher is its baselevel activation. If an element is not needed for some time, its baselevel activation decays. These changes of baselevels depending on use and time are referred to as *baselevel learning*.

$$odds = \frac{2n}{\sqrt{L}} \quad (3)$$

As mentioned earlier, I assume that a new memory element is created for each updating event and that this element is rehearsed a number of times after its creation. Each single rehearsal involves a retrieval of the element, which increases the respective  $n$ . The number of rehearsals is a free parameter of the model. The lifetime  $L$  is determined by the lag at which the element was created and the duration of each step (which was seven seconds in Venturino's experiment).

Odds can be transformed into probabilities using the definition  $odds = p/(1-p)$ . This gives us Equation 3a.

$$p = odds/(odds+1) \quad (3a)$$

<sup>1</sup> Equation 3 is an approximation of the original ACT-R equation. The approximation includes the default value 0.5 of the "baselevel-learning" parameter. The similarity between the time functions of Equations 1 and 3 illustrates the ACT-R notion that memory processes reflect the probabilistic structure of the environment.

With this equation, the probability of retrieval  $p$  can be predicted for each memory element that was created to represent an update event.

This probability is assumed to be degraded in the same-attribute conditions where interference is expected, depending on the number of competing memory elements. Assuming that only the elements that represent current values of the same attribute are competing, the respective numbers  $n_c$  are two, four, and six. Note that in the different-attribute conditions there is only one current value of each attribute, so no interference is expected there.

I assume further that the interference effect is "buffered" by a constant  $c$ , which is the second free parameter of the model. Equation 4 shows the degrading function. To ensure that the degraded probability value ranges between 0 and 1,  $c$  may vary between 0 and 0.5.

$$p' = \left\{ \begin{array}{l} p \mid \text{condition} = \text{different - attributes} \\ p \cdot (c + 1/n_c) \mid \text{condition} = \text{same - attribute} \end{array} \right\} \quad (4)$$

It is important to realize that the cognitive component of the model makes no assumptions about the influence of the information organization factor. This can be justified by the result that this factor accounted for only 1% of the variance in the experiment. Nevertheless, the predictions for the one-object conditions are slightly different from those for the many-object conditions, because of the different probability structure of the one-object/same-attribute condition.

Equation 5 describes how the prediction of the model is obtained by summing up for each time step the probability that its value will lead to a correct answer and dividing the sum by the number of current values (i.e. memory load).  $q_i$  is the probability that the value of step  $i$  is still current,  $p_i'$  is the probability that the memory element representing that value is retrieved, and  $n_c$  is the number of current values.

$$P = \frac{\sum_{i=1}^s q_i \cdot p_i'}{n_c} \quad (5)$$

Summing up the probabilities of all memory traces gives a generalized estimate of their potential to answer all possible probes. The prediction of the model should be the expected proportion of correct answers. Therefore, the sum must be divided by the number of current values, because, depending on memory load, all traces contain two, four, or six traces that represent current values.

The two free parameters of the model, number of rehearsals  $n$  (Equation 3) and  $c$  (Equation 4), were estimated to optimize the fit to the data. The resulting values were  $n = 12$  rehearsals and  $c = 0.5$ . With these values, the prediction of the model matched the data with an  $R^2$  of 0.89 and a root-mean-square deviation (RMS) of 0.07.

Although an  $R^2$  of 0.89 might not seem very high, one has to take into account that twelve degrees of freedom were predicted by adjusting only two parameters. For the many-objects conditions alone, the  $R^2$  is 0.97 and the RMS is 0.04.

Note that the task analysis contributes to the prediction only in combination with the memory assumptions. Since  $\sum q_i$  equals  $n_c$  (cf. Equations 1 and 2), a constant probability of retrieval  $p'$  would simplify the numerator of Equation 5 to  $n_c \cdot p'$ , and Equation 5 would yield the constant  $p'$ . The variation of probabilities of being current,  $q_i$ , would be completely neutralized by a constant probability of retrieval,  $p'$ , and no differences would be predicted.

If only the assumption about baselevel learning would be omitted, Equation 4, which models the interference effect, would still create variations in  $p'$ . I tried to fit the data without the calculation of retrieval probabilities as a function of time (i.e. without baselevel learning), using a single value for the probability of retrieval  $p$ . This value was estimated as  $p = 0.85$ . The resulting values of  $R^2 = 0.77$  and RMS = 0.08 show that the interference assumption alone accounts for a fair amount of variability, but the prediction is clearly improved by the assumption about baselevel learning.

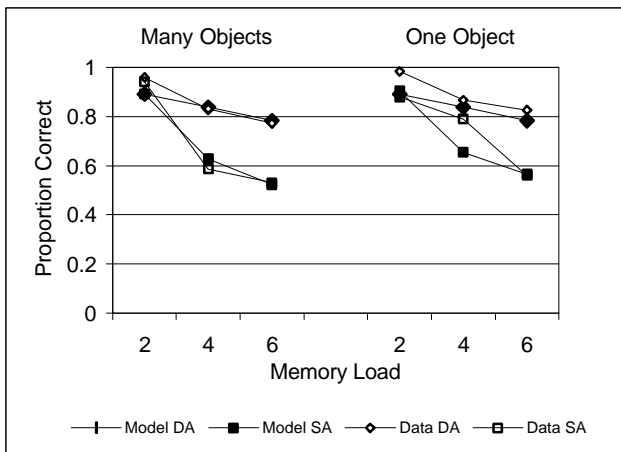


Figure 3: Mean proportions of correct answers from Venturino (1997) and the model (DA: different attributes, SA: same attribute)

## Discussion

It is remarkable that a model that combines a task analysis with a small set of basic assumptions about the dynamics of memory elements can reproduce the data so well. This demonstrates that there is no reason to distinguish between memory capacity for static information and memory capacity for dynamic information, as it was suggested by Venturino (1997). The model implies a simple rehearsal strategy in which only the most recent value is rehearsed about twelve times. This number is slightly higher than the number of rehearsals that were needed to encode the instruction in a model of serial attention by Altmann (2000). Because the present model does not include activation spread by cues, which would also in-

also increase the probabilities of retrieval, this number of rehearsals is probably overestimated.

The simplicity of the rehearsal strategy was not assumed for sake of parsimony, but is actually functional. If more than the most recent value would be rehearsed, this would strengthen older memory traces to a degree that new traces could hardly compete with the older ones, thus preventing the system from retrieving newer traces which are more likely to represent current values. This prediction of the model should be tested in future research.

Although the model is successful with standard assumptions about memory, there is one feature that points in a similar direction as Venturino's (1997) speculation about different types of memory capacity. The parameter  $c$  in Equation 4 and its estimated value of 0.5 establish a threshold of two memory elements up to which no interference occurs. This raises the question if there might be a preferential type of representation for a very small number of elements. Such an assumption, implemented in a simulation model, would remedy the model's underestimation of performance in the lowest memory load conditions. ACT-R provides opportunities to model such a preferential representation, for example if one assumes that one or two of the most recent values are always elements of the focus of attention.

Another interesting question that can be stated more precisely thanks to the model is what interferes with the correct answer. The present model assumes that only the *current* values that share the same attribute interfere with each other, resulting in no interference in the conditions with different attributes. The small memory load effect in these conditions is due to the increasing mean "age" of the memory representations with higher memory load. Also in the same-attribute conditions, the interference factor (Equation 4) depends on the number of *current* values.

This assumption, although critical for the predictions and supported by the data, can be questioned. It might be more plausible to assume that not only the current values of an attribute compete, but all of them. Interestingly, this assumption predicts more interference for lower memory loads in the different-attribute conditions. Suppose there are twelve memory elements representing the twelve most recent values, some of them current, some not. Under memory load 2, there are two different attributes, thus on average six of the elements share the same attribute. Under memory load 4, three elements, and under memory load 6, two elements share the same attribute. Thus, the lower the memory load, the more elements of the same attribute compete with each other, producing higher interference - a pattern that is contradicted by the data.

All these observations converge at the question of what happens with the memory elements that represent outdated values. The decay of baselevel activation certainly contributes to the diminishing interference potential of outdated memory elements, but the decay guarantees this effect only if no noise is assumed. If one assumes some noise, which seems to be realistic, much more interfer-



ence would be expected than predicted by the present model and found in the data. I have started to investigate this problem using a rather process oriented, symbolic type of modeling. It will be interesting to see if additional processes such as active inhibition have to be assumed to explain the rather low interference effects.

Another advantage of symbolic modeling is that it demands more details about cues. In the present model, it was implicitly assumed that only one strong cue is in effect. It is the attribute in the different attribute condition. In this condition, only one value of each attribute is a current value. This value is always the most recent - and thus the most active value of that attribute. Therefore, using the attribute as a constraint and retrieving the most active memory element delivers the correct answer.

The reason why the attribute is assumed to be the only strong cue is that the relation between an attribute and its values is the only one that stays constant throughout the experiment. In their Experiment 4, Hess et al. (1999) established a constant relation between a spatial cue and attribute values in a many-objects/same-attribute condition. This cue was strong enough to abolish the interference effect that is usually observed in that condition.

The objects on the other hand are much less potent cues, because the relation between objects and attribute values varies. This is probably the reason why the information organization factor (which is operationalized through the number of objects) exerts so little influence. The model even justifies to doubt if there is a real effect at all, because the difference between the one-object and many object conditions is partially explained by the different probability structure of the material in the one-object/same-attribute condition. One data point that contributes much to the difference is the performance in memory load 4 of that same condition where the model's predictions deviate most highly from the data. A replication would be necessary to find out if this deviation is rather due to noise in the data or to inappropriate assumptions of the model. In such a study, the probability distribution of the one-object/same-attribute condition should be approximated to the distributions of the other conditions in order to draw clearer conclusions about information organization.

## Conclusions

The model has demonstrated clearly that a task analysis combined with a small set of assumptions about the dynamics of memory traces is sufficient to reproduce the basic results related to the keeping track paradigm. No distinction between memory capacities for static and for dynamic information is needed. The model implies that interference occurs between representations of current values. Hence, an issue of future research should be to investigate what happens with the representations of outdated values. As to the factor information organization, it has been shown that the effect of this factor is partially due to the deviating probability structure of one of the conditions. To clarify the influence of information organization, the probability structures should be assim-

lated in future studies by means of the presented task analysis.

## Acknowledgments

I wish to thank Mike Venturino for providing me with detailed information about his experiment. I am grateful to my colleagues at the Human Factors and Applied Cognition Program, especially to Deborah Boehm-Davis and Erik Altmann for their comments on this paper. Thanks also to Wayne Gray who encouraged me to model this paradigm. This research has been supported by grants NAG 2-1289 from the NASA and 99-G-010 from the FAA.

## References

- Altmann, E. (2000). The anatomy of serial attention: An integrated model of set shifting and maintenance. *Proceedings of the Third International Conference on Cognitive Modeling*, Groningen, The Netherlands.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37, 32-64.
- Hess, S. M., Detweiler, M. C., & Ellis, R. D. (1999). The utility of display space in keeping track of rapidly changing information. *Human Factors*, 41, 257-281.
- Venturino, M. (1997). Interference and information organization in keeping track of continually changing information. *Human Factors*, 39, 532-539.
- Yntema, D. B., & Mueser, G. E. (1960). Remembering the present states of a number of variables. *Journal of Experimental Psychology*, 60, 18-22.

# Motivating Base-Rate Sensitivity (Sometimes): Testing Predictions of the RCCL Framework

Christian Schunn (schunn@gmu.edu) & Thuy L. Ngo

Department of Psychology; George Mason University  
Fairfax, VA 22030 USA

## Abstract

In choice situations, people are usually (but not always) sensitive to the base-rates of success of the options, and this base-rate sensitivity usually (but not always) goes up when motivation levels are increased. The RCCL framework, which emphasizes what information is represented by the individual and what strategies are used, provides an explanatory framework for these types of effects. In particular, RCCL predicts that manipulations of motivation levels should produce changes in the strategies being used, which will not produce a change in base-rate sensitivity for dimensions not represented in the strategies. This paper reports an empirical test of these predictions; changes in strategy use and a lack of change in base-rate sensitivity are found, as predicted by RCCL.

## Introduction

In making optimal choices in an uncertain world, a problem-solver must pay attention to the base-rates of success of each of the possible choices: the past success rates are usually good indicators of future success rates. For example, travel routes that were generally congested in the past are likely to be congested in the future. While one often finds base-rate insensitivity when base-rates are presented verbally in textual problems (e.g., Ginossar & Trope, 1987; Tversky & Kahneman, 1982), one usually finds extremely good base-rate sensitivity in experiential paradigms (e.g., Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Maddox, 1995). That is, when problem-solvers experience many decisions during problem solving, they are typically very sensitive to the base-rates of success that they have experienced. However, there are a few well-documented exceptions to this general trend of good base-rate sensitivity (Goodie & Fantino, 1995; Goodie & Fantino, 1996; Medin & Edelson, 1988).

A challenge for cognitive science is to come up with models that explain why and to what degree one observes base-rate sensitivity (or base-rate neglect). Recently Lovett and Schunn (1999) proposed RCCL (pronounced "ReCY-CLe") as a framework for providing such an explanation.

RCCL specifies how task representations can influence choice in experiential base-rate situations. The four main stages of processing in RCCL are: (i) Represent the task, (ii) Construct a set of action strategies consistent with that task representation, (iii) Choose among those strategies according to their success rates, and (iv) Learn new success rates for the strategies based on experience. The primary theme underlying RCCL is that a task representation constrains the set of strategies an individual will use for taking

actions in the task environment. Making choices according to the learned success rates of a certain set of strategies enables RCCL to produce base-rate sensitivity or base-rate neglect in direct-experience situations; sensitivity arises only when the constructed strategies include stimulus features that are important to success in the task. The RCCL framework also includes re-cycling through the above processes when the current representation and strategies lead to low success rates. This implies that an individual's task representation and strategy set need not be static but rather can develop with experience.

At this level of description, the components of RCCL may seem intuitive to the point of being obvious: how else could it be done? However, the central contribution of RCCL may be to forefront processes that are highly likely to be going on yet have been ignored in previous accounts of human choice processes. Moreover, there are accounts of choice processes that do not invoke (and perhaps even deny) the role of mental representations (e.g., Goodie & Fantino, 1995; Goodie & Fantino, in press).

Lovett and Schunn (Lovett & Schunn, 1999) described two experiments that provided empirical support for the RCCL framework. In one experiment they showed that people prefer representations and strategies that make use of information predictive of successful problem solutions. In the second experiment, they demonstrated that one could change the superficial characteristics of the task environment such that participants would prefer one representation or another, and that this manipulation determined what base-rates participants would learn.

The current paper seeks to further test RCCL specifically, and strategy-based accounts of choice processes more generally (e.g., ACT-R). The insight is to examine the effects of performance motivation on base-rate sensitivity in a problem-solving context.

To tease apart strategy and non-strategy-based accounts of choice processes, one needs to distinguish between simple and complex choice situations. In a simple choice situation there is a direct, one-to-one mapping between the person's strategies and external alternatives. That is, one can adequately describe the person's strategies in terms of simple external choices. For example, when presented with a left and right button to press, the person represents the choice strategies as Select-Right and Select-Left. By contrast, in a complex choice situation, there is not a simple mapping between strategies and external alternatives. That is, a given strategy might map onto different external alternatives on different trials; two different strategies may map onto the same external alternative on the given trial.

In very simple choice situations, strategy-based and non-strategy-based accounts make very similar predictions about the effects of motivation on base-rate sensitivity. The greater the value of a success, the more participants (human or otherwise) will prefer the more successful choice (see Anderson, Lebiere, & Lovett, 1998). In other words, greater motivation levels should produce higher base-rate sensitivity.

In complex problem-solving situations, however, RCCL makes two novel predictions regarding the effects of motivation. First, RCCL predicts shifts in strategy choice as a function of motivation changes when the strategies vary in terms of effort and success. That is, it is the selection among strategies (rather than externally defined alternatives) that is directly influenced by motivation. This prediction is easily formalized using various forms of expected utility theory (e.g., see Anderson et al., 1998). However, intuitively this prediction can be understood as people becoming more willing to put out the extra effort associated with a more effortful but more successful strategy when they are more motivated to succeed.

RCCL's second prediction is that this change in strategies may produce increases or decreases in base-rate sensitivity depending on whether the new or old strategies represent the external alternative feature whose base-rate is being manipulated. As an abstract example (the next section presents a concrete example), suppose there is a strategy S1 that does represent an external feature F1 (i.e., S1 makes direct use of feature F1 to make a choice) and a strategy S2 that does not represent external feature F1 (i.e., S2 makes choices without making use of feature F1). Then, when people use strategy S1, they will be sensitive to the base-rates with which F1 predicts success, whereas when they use strategy S2, they will not be sensitive to the base-rates with which F1 predicts success. Thus, if increasing motivation leads people to move from S1 to S2, then base-rate sensitivity to F1 will go down. By contrast, if increasing motivation leads people to move from S2 to S1, then base-rate sensitivity to F1 will go up. In general, for situations in which increases in motivation level cause a person to shift to a strategy that does not represent the relevant base-rate, then RCCL predicts decreases in base-rate sensitivity with increases in motivation level.

By contrast, non-strategy-based accounts would always predict an increase in base-rate sensitivity with increasing performance motivation. As the value of currently picking the best option increases, one should find better base-rate sensitivity (or even over-matching). Intuitively, the more incentive one has to do well, the more one pays attention to cues (e.g., base-rates) that will predict accurate choices.

The role of performance motivation in base-rate sensitivity and strategy adaptivity is also an important question for other reasons. Recent research (Schunn & Reder, 1998) has shown that there are individual differences in the degree to which people adapt their strategies to shifting base-rates of success, and that these base-rate sensitivity individual differences are correlated with individual differences in inductive reasoning skill. A remaining question, however, is whether these individual differences in base-rate sensitivity can also be partially explained by motivational differences (i.e., are the more base-rate sensitive participants simply the

more motivated ones). The current research will show the degree to which base-rate sensitivity is influenced by motivation levels and thus whether there is a potential confound in the individual differences research in this area.

## Methods

### Participants

Ninety-two George Mason University undergraduates participated for course credit and were randomly assigned to one of two conditions. Nine participants encountered technical difficulties with the computer setup, and their data is not included in the analyses.

### Building Sticks Task

In the building sticks task, participants are presented with 3 different-sized building sticks which they must choose among to create a given goal stick. To achieve the goal stick, participants add or subtract any combination of the buildings sticks provided.

For a given BST problem, using one of two approaches will result in the goal stick (Note, here I use the term "approach" to refer to an externally-defined alternative in contrast to a true strategy). Using the undershoot approach, participants start with a stick shorter than the goal stick and add to it to achieve the desired stick length. In the overshoot approach, participants pick a stick longer than the goal stick and subtract from it until the goal stick is created. Each problem is designed to be solved using one of the approaches, but not both.

For example, if the goal stick provided is 8 units in length and the 3 sticks A, B, and C, are 15, 6, 7, respectively, using the overshoot approach will solve this problem. To achieve the goal stick, participants start with stick A and subtract stick C ( $15 - 7 = 8$ ) to reach the solution. Using the undershoot approach in this case would never result in the desired stick length because picking stick B and adding to it will not equal 8 ( $B + C = 6 + 7 = 13$ ). Note that participants in the task are not given numerical lengths of the sticks. Instead, participants must estimate stick lengths and determine which sticks would lead to the goal stick before taking the appropriate steps. As a result, participants were forced to implicitly apply an approach (overshoot/undershoot) to solve each problem without knowing in advance whether it would work.

Participants were given 80 BST problems to solve. Participants worked through each problem until the goal stick was achieved. If a solution was not reached within 5 moves or less, participants were asked to reset the problem and start over again until the goal stick was reached. Each problem was designed to be solved by only one of the two approaches.

For the first 40 problems, the overshoot approach was biased to be more successful in solving the problems than the undershoot approach, with 70% and 30% success rates for each approach, respectively. For the second 40 problems, the success rates were reversed, with the undershoot approach biased to be more successful 70% of the time. This sequence was held constant across conditions. The degree to which participants adapted their approach choices to this

base-rate manipulation is one of two primary dependent measures.

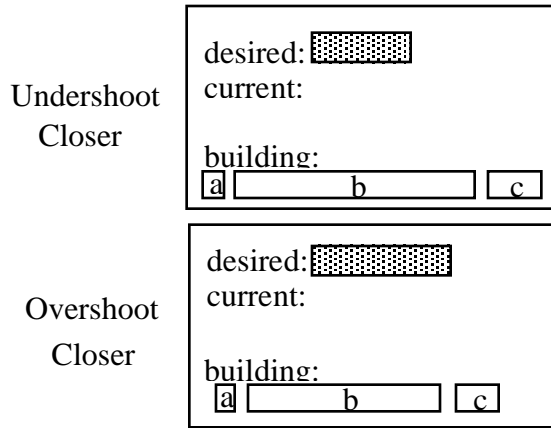


Figure 1. Examples of Undershoot looking (top) and Overshoot looking (bottom) BST problems.

In addition, each problem was designed with a feature pattern, called a relative length cue, which was predictive of the correct approach to use for a given goal stick. One of the 3 building sticks was designed to appear closer in length to the goal stick, suggesting a bias towards use of one approach over another. As shown in the top of Figure 1, stick C looks closest in length to the goal stick. Therefore, participants are more likely to start with stick C (initiating the undershoot approach) and adding segments until the desired stick length is reached. In contrast, stick B in the bottom of Figure 1 looks closer in length to the goal stick than sticks A and C. Thus, participants will pick stick B and subtract segments until the goal stick is achieved.

Of the 80 BST problems, 40 problems appeared biased towards overshoot and 40 problems were biased towards undershoot. This cue was manipulated to be successful 70% of the time—the predictiveness of the relative length cue remained constant across both conditions. Table 1 summarizes how problem types were manipulated over time for all participants (in both conditions)—overshoot success rate being changed over time, while the predictiveness of the length cue was held constant over time.

Table 1. Overshoot success rate and predictiveness of the length cue over blocks of trials (in both conditions).

Predictive cue	Trials 1-40	Trials 41-80
Overshoot success rate	70%	30%
Predictiveness of length cue	70%	70%

Consistent with the RCCL account, participants in the BST tend to report a variety of strategies (Lovett & Schunn, 1999). The two most salient strategies are the hill-climbing and exclusive strategies. In the hill-climbing strategy, participants compare the goal stick to the building sticks and select the stick that most closely matches the length of the goal stick. In the exclusive strategy, participants simply select one approach, overshoot or undershoot, without regard to which appears to be closest to the goal. As long as hill-climbing distance is predictive of solution success (as it

was in the current experiment), the hill-climbing strategy is more likely to be successful than the exclusive strategy. However, the hill-climbing strategy also involves more effort because of the visual comparison component. Which strategy participants adopt across conditions will be the second primary dependent measure.

This task is a complex problem solving situation (according to the definition given in the introduction) because there is not a simple mapping between strategies and external alternatives. Table 2 presents the choices that participants would tend to make in each of the blocks under the hill-climbing and exclusive strategies. The exclusive strategy should tend to select the most successful approach regardless of what the problem looked like. By contrast, the hill-climbing strategy should tend to select approaches according to problem appearance, independent of the base-rate of success of each approach.

Table 2. Expected modal approach (O=overshoot, U=undershoot) under each strategy in each block for undershoot and overshoot biased problems.

Strategy	Trials 1-40		Trials 41-80	
	O-biased	U-biased	O-biased	U-biased
Hill-climbing	O	U	O	U
Exclusive	O	O	U	U

## Procedure

Participants were randomly assigned to one of two conditions. In the Unpaid condition (the control group), participants received course credit only. However, in the Paid condition (the motivated group), participants received compensation in addition to course credit. Payment was based on a \$10 scale and calculated according to the percentage of problems solved correctly within 5 steps or less. That is, participants who solved 80% of the problems in so few steps received \$8.00 while participants who solved 60% of the problems in this way received \$6.00.

At the beginning of the experiment, a computer tutorial provided participants with step-by-step instructions to the task, along with an animated demonstration of the undershoot and overshoot approaches. For participants in the Paid condition, the last page of the instructions informed the participants that they were being compensated for their participation based on their performance on the task. The instructor reiterated this to ensure participant motivation.

## Predictions

The hill-climbing strategy is a more successful but more effortful strategy than the exclusive strategy. Therefore, RCCL predicts that the motivation manipulation should increase the participants' use of the hill-climbing strategy. Let us define base rate sensitivity as the difference in frequency of overshoot approach use from the first to second halves of the experiment. Then, because the exclusive strategy is more sensitive to the base-rates of overshoot and undershoot, RCCL predicts no effect of the motivation manipulation (or perhaps a decrease) on base-rate sensitivity, at least as defined in terms of external choices. By contrast, non-representational accounts (and perhaps even common

sense) would suggest that the participants given the performance incentive should show greater base-rate sensitivity.

### Strategy Coding

At the end of the Building Sticks Task, participants were asked about what strategies they used. Responses were classified into one of 5 categories: using whatever the problem looked like (hill-climbing), always using one stick size first (exclusive), using what worked previously (memory), randomly selecting sticks (trial and error), and other strategies (miscellaneous). Based on a recoding of 20% of the data by a second coder, the reliability for this coding scheme was 93%.

## Results & Discussion

### Verifying Differences in Strategy Features

The predictions of strategy shifts rest on assumptions about the differential effort and success rates associated with the various strategies. The assumptions were tested by examining the relationship between 1<sup>st</sup> mentioned strategy and participant mean success rates (across all blocks) and mean time to make the first move on each trial (across all blocks). Note that time to execute the strategy is used as an approximation of the effort required by a strategy. We expect that the participants using the hill-climbing strategy should be more successful and require less time to make choices. However, given that participants in this task have been found to typically each use several strategies during the course of the session (Lovett & Schunn, 1999), one would expect analyses averaging performance data across the whole session to show diluted trends.

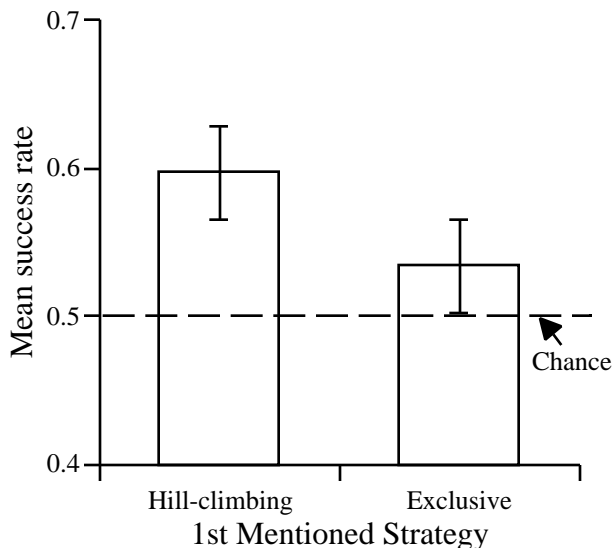


Figure 2. Mean success rate (and SE bars) for the hill-climbing and exclusive strategies.

Overall, there was a significant effect of 1<sup>st</sup> strategy mentioned on the mean success rates,  $F(2,54)=8.8$ ,  $MSE=0.004$ ,  $p<.005$  (see Figure 2). Specifically, exclusive strategies ( $n=10$ ) showed lower success rates than hill-

climbing strategies ( $n=46$ ). This trend was consistent within both conditions.

Overall, the timing data was more variable, with a non-significant overall effect of 1<sup>st</sup> strategy mentioned on the mean times to make the first move  $F(2,54)=2.1$ ,  $MSE=1.31$ ,  $p>.15$  (see Figure 3). However, exclusive strategies did show the expected lower mean times than did hill-climbing strategies. This trend was consistent within both conditions.

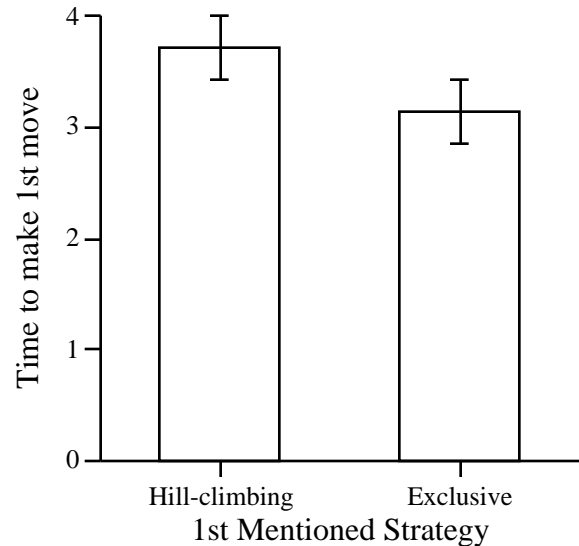


Figure 3. Mean time to make 1<sup>st</sup> move on each problem (and SE bars) as a function of the 1<sup>st</sup> mentioned strategy.

In sum, the assumptions about the differential effort and success rates between the hill-climbing and exclusive strategies were at least qualitatively supported.

### Strategy Changes

Table 3 presents the frequency of mention of each strategy type based on the first strategy mention. We see the predicted increase in the use of hill-climbing strategy and the predicted decrease in the use of the exclusive strategy.

Note also that the Memory strategy, a relatively effort intensive strategy, showed an increase in the Paid condition, and that the Trial & Error strategy, a relatively effort free strategy showed a decrease in the Paid condition. The predicted increase in reliance on effortful strategies was statistically significant,  $t(81)=1.6$ ,  $p<.05$  (one-tailed).

While these effects were not large, it is important to note that these analyses are likely to be an underestimate of the effects—participants did not indicate how often they used the mentioned strategies, and they do try multiple strategies.

Table 3. Proportion of participants mentioning each strategy within each condition.

Strategy	Unpaid (N=42)	Paid (N=41)
Hill climbing	0.50	0.61
Exclusive	0.17	0.07
Memory	0.10	0.15
Trial & Error	0.17	0.15
Misc.	0.07	0.02

Another important issue raised by RCCL is whether motivation has an impact on the degree of search for an optimal strategy. Towards this end, we examined the effect of condition on number of different strategies mentioned. There was no significant effect of condition on the number of strategies mentioned,  $F(1,81) < 1$ . Both the Paid and Unpaid participants mentioned a mean of 1.4 strategies per participant.

### Overall Base-Rate Changes Over Time

Both groups showed a rise in the amount of Overshoot use in the first half followed by a drop in the second half  $F(3,243)=16.6$ ,  $MSE=0.015$ ,  $p < .0001$  (see Figure 4). There was no main effect of condition,  $F(1,81) < 1$ , nor was there a significant interaction,  $F(3,243)=1.0$ ,  $MSE=0.015$ ,  $p > .3$ . To directly quantify the influence of condition on base-rate adaptivity, one can define base-rate adaptivity as the amount of drop in Overshoot use from the first half to the second half (difference of half means). On that measure, both Paid and Unpaid participants shifted exactly 7% in their use of Overshoot over time. As Figure 4 reveals, if anything, Paid participants were less sensitive to the base-rates. Thus, as predicted by RCCL, motivation manipulations produced changes in strategy use, not changes in base-rate sensitivity.

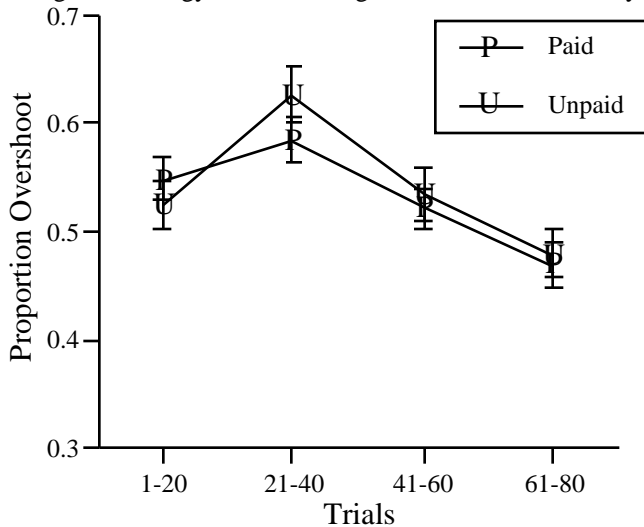


Figure 4. Proportion of overshoot choices within each set of twenty trials within each condition.

### Hill-climbing sensitivity

One can also analyze the effects of problem appearance (whether the problem appearance was biased towards overshoot or biased towards undershoot) on solution method and its interaction with condition and blocks. As one always finds in this task, there are large effect of problem appearance on the proportion of overshoot selections,  $F(1,81)=657.0$ ,  $MSE=0.038$ ,  $p < .0001$ . More interestingly, there was also a significant interaction of appearance with condition,  $F(1,81)=6.5$ ,  $MSE=0.038$ ,  $p < .02$ . In particular, Unpaid participants showed a significantly lower sensitivity to problem appearance than did the Paid participants (49% versus 60% differences between overshoot-biased and undershoot-biased problem types).

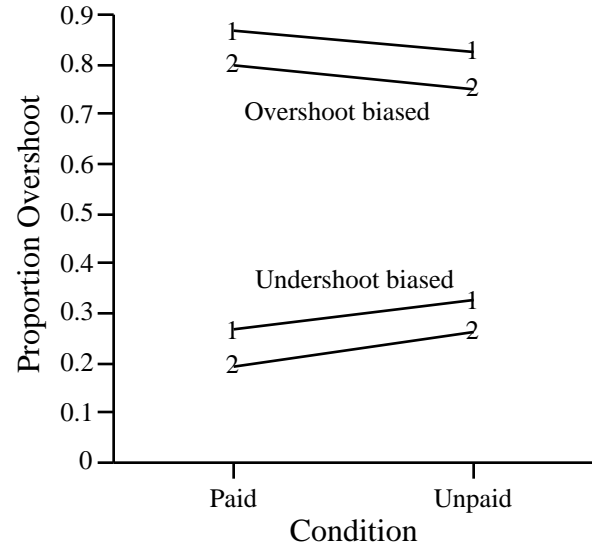


Figure 5. Proportion of overshoot choices as a function of problem appearance (Overshoot biased /Undershoot biased) and condition (Paid/Unpaid), for the first (1) and second (2) halves of the experiment.

This effect establishes that the payment manipulation did have some influence on participants, and thus clarifies the interpretation of the null effects on base-rate sensitivity. This effect is also consistent with increases in hill-climbing strategy use as a result of the manipulation.

### General Discussion

This experiment found that increasing motivation levels can produce strategy changes (as measured by self-report and patterns in choice) without producing changes in base-rate sensitivity. The changes in strategy choice were consistent with a shift in motivation levels—a shift from lower-success/lower-effort strategies to higher-success/higher-effort strategies. Thus, the key predictions of the RCCL framework with respect to the effects of motivation levels on choice patterns were met. These findings are not consistent, by contrast, with non-strategy-based theories of choice that focus entirely on external alternatives rather than internal representations and strategies.

It should be noted that RCCL is a general framework, not a detailed model. With respect to the predictions regarding the effects of motivation, there are several particular utility-based strategy models of choice processes that could be used to account for the obtained results, including ACT-R (Anderson & Lebiere, 1998), SAC (Schunn, Reder, Nhoyvanisvong, Richards, & Stroffolino, 1997), and ASCM (Siegler & Shipley, 1995).

Some of the results of the current experiment are potentially difficult to interpret because they involve null effects of a manipulation. However, the manipulation did produce some effects demonstrating that it was strong enough to influence behavior. Moreover, it is somewhat rare to find a case in which performance in a problem-solving task does not improve when undergraduates normally taking part only for course credit are suddenly paid for higher performance levels.

Our experiment is also not the first to find no effect of motivation manipulations on base-rate sensitivity. For example, Goodie and Fantino (1995) found no effect of a motivation manipulation on base-rate sensitivity. They also used conditions of course credit and pay versus course credit alone, although they paid their participants as much as \$40. While Goodie and Fantino did not explain their null result (it was also not the focus of their experiment), RCCL provides a potential explanation. The key is to examine whether the motivation manipulation produced changes in strategy use rather than changes in choices at the level of simple external alternatives. While the task used by Goodie and Fantino was not a complex problem-solving task, Lovett and Schunn (1999) established that participants do use a wide variety of strategies during that choice task as well.

Another consequence of the current experimental findings is that they resolve a question about individual differences. In particular, previous research on individual differences in sensitivity to base rates (Schunn & Reder, 1998; Stanovich & West, 1998) left open the possibility that the differences were due to motivational differences. The current research suggests that the observed individual differences in base-rate sensitivity are not so easily attributed to motivational differences.

The current experimental findings also permit some refining of the RCCL framework. RCCL posits that people will search for new representations and strategies when the success rates of the current alternatives are too low. An open question was whether motivational levels entered into determining when a search for new representations and strategies was begun. The current findings suggest that motivational levels do not have a large role of in the amount of search for alternative representations and strategies. Or, at least, all of the participants were sufficiently motivated to conduct such searches.

As a final note, the current experiment only manipulated one kind of motivation: extrinsic motivation. There are other types of motivation. For example, research (e.g., Button, Mathieu, & Zajac, 1996) has shown that people also differ in terms of their performance motivation (the degree to which they need to succeed) and learning motivation (the degree to which people prefer to learn new things). It is an open question whether those dimensions of motivation will have similar influences on choices processes generally, and base-rate sensitivity in particular.

### Acknowledgments

Work on this project was supported by funds to the first author from the Department of Psychology and the College of Arts and Sciences at George Mason University. We would like to thank Erik Altmann, Melanie Diez, Wai-Tat Fu, Eliza Littleton, Lelyn Saner, and Susan Trickett for comments made on an earlier draft.

### References

Anderson, J. R., & Lebiere, C. (1998). *Atomic components of thought*. Mahwah, NJ: Erlbaum.

- Anderson, J. R., Lebiere, C., & Lovett, M. (1998). Performance. In J. R. Anderson & C. Lebiere (Eds.), *Atomic Components of Thought*. Mahwah, NJ: Erlbaum.
- Button, S. B., Mathieu, J. E., & Zajac, D. M. (1996). Goal orientation in organizational research: A conceptual and empirical foundation. *Organizational Behavior and Human Decision Processes*, 67(1), 26-48.
- Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 556-571.
- Ginossar, Z., & Trope, Y. (1987). Problem solving in judgment under uncertainty. *Journal of Personality and Social Psychology*, 52, 464-473.
- Goodie, A. S., & Fantino, E. (1995). An experientially derived base-rate error in humans. *Psychological Science*, 6, 101-106.
- Goodie, A. S., & Fantino, E. (1996). Learning to commit or avoid the base-rate error. *Nature*, 380, 247-249.
- Goodie, A. S., & Fantino, E. (in press). Representing the task in Bayesian reasoning: Comment on Lovett and Schunn (1999). *Journal of Experimental Psychology: General*.
- Lovett, M. C., & Schunn, C. D. (1999). Task representations, strategy variability and base-rate neglect. *Journal of Experimental Psychology: General*, 128(2), 107-130.
- Maddox, W. T. (1995). Base-rate effects in multidimensional perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 288-301.
- Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, 117, 68-85.
- Schunn, C. D., & Reder, L. M. (1998). Strategy adaptivity and individual differences. In D. L. Medin (Ed.), *Psychology of Learning and Motivation* (Vol. 38, pp. 115-154). New York: NY: Academic Press.
- Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., & Stroffolino, P. J. (1997). To calculate or not to calculate: A source activation confusion model of problem familiarity's role in strategy selection. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 23(1), 3-29.
- Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In G. Halford & T. Simon (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 31-76). New York: Academic Press.
- Stanovich, K. E., & West, R. F. (1998). Who uses base rates and P(D/~H)? An analysis of individual differences. *Memory & Cognition*, 26, 161-179.
- Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

# Now They See the Point: Improving Science Reasoning Through Making Predictions

Christian D. Schunn (schunn@gmu.edu) Christine J. O'Malley (comalley@gmu.edu)

Department of Psychology; George Mason University  
Fairfax, VA 22030 USA

## Abstract

Previous research on scientific reasoning has found that many students find it difficult to think about the theoretical level when asked to design experiments. Two studies are reported that explore whether forcing students to make predictions before running their experiments improves their scientific reasoning performance. Both studies find that, if they do make predictions, students become more focused on the theories they are being asked to test. The students become more likely to make conclusions about the theories under test and they design experiments more relevant to the theories under test.

## Introduction

Science education is a core component of education throughout the industrialized world, and the ability to reason scientifically is a generally valued skill. Nevertheless, relatively little is known about the details of how students become good scientific reasoners. There is one clear fact, however, about the developmental process that has been frequently documented: students do not come naturally to many aspects of scientific reasoning, and it is not easy to teach those skills (e.g., Kuhn, 1989; Lehrer, Schauble, & Petrosino, in press; Schauble, 1990). Even towards the end of a students' college education, many basic scientific reasoning skills are weak or missing (Schunn & Anderson, 1999, In press).

Developing an understanding of what can improve scientific reasoning skills is an important problem for cognitive science. It tends to involve many disciplines of cognitive science because it is a difficult problem that requires resolving: 1) what it means to reason scientifically (philosophy, history), 2) what cognitive processes are involved (psychology), 3) what kinds of interventions are successful (education), and 4) constructing complex computer environments that model and support it (computer science). What makes it an inherently cognitive science-like problem is that these four components are intimately interconnected.

This paper explores one simple method for improving scientific reasoning: forcing students to make predictions before running an experiment. It presents two empirical studies conducted in the psychology laboratory (as opposed to a classroom situation). The studies compare the scientific reasoning performance of students forced to make predictions before each experiment with students not asked to make predictions before each experiment. Before turning to the empirical studies, we will provide additional background on this particular issue.

To make a prediction, one needs a hypothesis or theory. Science textbooks generally recommend that one should

always have a hypothesis before running an experiment. However, philosophical, historical, and, more recently, psychological accounts of science agree that one need not always have a hypothesis before running an experiment (see Okada and Shimokido, in press, for a review).

We acknowledge that there are plenty of situations in which people do not have a theory before conducting the experiment (Klahr & Dunbar, 1988). A central aspect of doing science, however, is the development and testing of formal theories—unifying or explanatory accounts that sit at a level above simple beliefs about the effects of particular variables.<sup>1</sup> Thus, having a theory to test is a common and important situation.

A separate question (and the one we examine) is whether one should always make concrete predictions before running the experiment *when one does have a theory to test*. There are plenty of situations in which one does have a theory to test. What role do explicitly made predictions serve in those situations?

Several recently developed science education computerized training environments have components that prod students into making predictions before running the experiments (e.g., Loh et al., in press; White, 1993, 1995). While these systems as a whole have been demonstrated to be effective, the value-added of the prediction-making component of these complex systems has not been tested in isolation. Thus, little is known from that research about whether forcing predictions actually improves reasoning.

There are several reasons to think that making predictions will help scientific reasoning. First, making predictions may remind students to focus on the theories that they are supposed to be testing. Schunn and Anderson (1999, In press) found that even undergraduates pay little attention to the theories they are supposed to be testing and instead simply explore the effects of different variables.

Second, making predictions may lead participants to consider alternative theories, and thus design experiments that more uniquely target the theory under test. On a related point, Koehler (1994) found that generating one's own hypothesis rather than being given the hypothesis leads to more accurate evaluations of the likelihood that the hypothesis is correct (however, see Schunn and Klahr (1993) for the exact opposite finding).

---

<sup>1</sup> We will use the term *theory* to refer to these general accounts and the term *hypothesis* to refer to beliefs about particular concrete variables. For example, ACT-R (Anderson & Lebiere, 1998) or SDDS (Klahr & Dunbar, 1988) are theories; "making predictions should improve reasoning" is a hypothesis.



There are also several reasons to think that making precise predictions will hurt scientific reasoning. First, it could be that making predictions would direct students away from the theoretical level that they are supposed to be testing and instead focus on simple empirical effects of particular concrete variables.

Second, getting students to make precise predictions could push students into an engineering rather than scientific mode (Schauble, Klopfer, & Raghavan, 1991; Tschirgi, 1980). In other words, it could lead students to focus on how to produce a particular outcome rather than on finding out why certain outcomes occur. As a variant of this theme, focusing on concrete predictions might lead students adopt a goal of trying to maximize their prediction accuracy (i.e., see how well they can predict outcomes). This new goal could be seen as a kind of engineering goal that is potentially at odds with the scientific goal of testing the theory.

In sum, there is a general belief that making predictions is important to scientific reasoning, possible reasons for it to help or hurt scientific reasoning, and little evidence one way or the other. We examine the role of making predictions on scientific reasoning in two different situations: when students are designing an experiment to choose between two alternative theories (Study 1); and when students are designing an experiment to test only one given theory (Study 2).

### **The Simulated Psychology Lab**

To examine the influence of making predictions on scientific reasoning skills we selected a real scientific question from cognitive psychology: what is the cause of the spacing effect in memory? The spacing effect itself is intuitively understood by undergraduates—that spaced practice produces better memory performance than massed practice (i.e., cramming is bad). The advantage of using this particular question is that it is relatively easy to explain to undergraduates without the use of complex domain-specific jargon and yet it is an authentic scientific problem rather than a toy problem. Recent work in the psychology and education of science suggests that it is important to use realistically complex problems (Chinn & Malholtra, in press).

As we noted earlier, not all situations require a theory to be tested in the experiment. However, since we wanted to examine the role of predictions, it was important to place students in a theory-testing situation. The spacing-effect problem may be too complex for students to quickly develop their own theories to test from the start. For this reason, we gave students theories to test. In particular, the students were presented with two theories that had been proposed to account for the spacing effect and their goal was to develop experiments to tease the theories apart (i.e., determine if either, both, or neither of these theories adequately explained the spacing-effect phenomenon).

Briefly, the first theory was the shifting context theory, which stated that memories were associated with the context under study and that context gradually shifted with time. Thus, the spacing effect occurs because spaced practice produces associations to more divergent contexts, which in turn are more likely to overlap with the test context. The second theory was the frequency regularity theory, which stated that

the mind tries to estimate how long memories will be needed based on regularities in the environment and, in particular, adjusts forgetting rates according to the spacing between items. The students were given longer descriptions of the theories (and the spacing effect itself) with concrete examples, could look at the descriptions of the theories throughout the task, and had several opportunities to ask the experimenter questions about the theories. (In Study 2, participants were only given the shifting context theory to test).

With the spacing-effect phenomenon and two theories in hand, we could have then given the students paper and pencil and asked them to describe an appropriate experiment. However, science is more than just experimental design. It also involves data analysis (among many other things). Moreover, few scientific questions are answered in the first experiment. Instead, scientists iterate and refine their methodology in response to experimental results. In order to place students in such a more realistic iterative situation that also included a data analysis process, we asked the students to design and interpret experiments using an environment called the Simulated Psychology Lab (Schunn & Anderson, 1999).

The Simulated Psychology Lab is a computer environment that allows students to design a wide variety of experiments and examine the results of those experiments. Students create experiments by selecting values for six factors, of which up to four could be simultaneously manipulated for any single experiment. They are told that the computer had been given the results of many actual experiments, and that it will display the results of any type of experiment they chose to generate.<sup>2</sup>

There were two groups of factors, source task factors and test factors, that the participants could manipulate. The source task factors included 1) repetitions—the number of times that the list of words was studied; 2) spacing—the amount of time spent between repetitions; and 3) source context—whether the participants were in the same context for each repetition or whether they changed contexts on each repetition. The test factors included 1) the test task—free recall, recognition, or stem completion; 2) delay—the amount of time from the last study repetition until the test was given; and 3) test context—whether the participants were in the same context or a different context at test relative to study. Only three of the factors are highly relevant to testing the two theories: spacing, source context, and test context. (In Study 2, since participants were asked to investigate only the shifting context theory, then only two factors are relevant: source context and test context).

For each of these factors, the participants could either vary it or hold it constant. Values had to be chosen for all of the factors before participants were allowed to continue. There were no confines on the order of value selection, and

---

<sup>2</sup> In fact, in order to produce numbers for the large number of possible combinations that the students could generate, the computer uses a mathematical model based on ACT-R (Anderson & Lebiere, 1998) that is very consistent with previous memory and spacing effect results, and includes a small level of random noise for added realism. See Schunn and Anderson (1999) for details.

the participants could change any of their selections at any time up until they chose to run the experiment.

The results were displayed using a table format and the participants could decide how to organize their tables. If participants were in an experimental condition that asked them to make predictions, then they made numerical predictions in a table. For each cell in the designed experiments, the participant must predict the percent correct of the hypothetical subjects. For example, Figure 1 presents an example table in which source context, spacing, and delay are manipulated and predictions have been already made for the first 8 cells (the bold 5 is currently being entered). Note that the table also contains information about the settings of the factors not being manipulated.

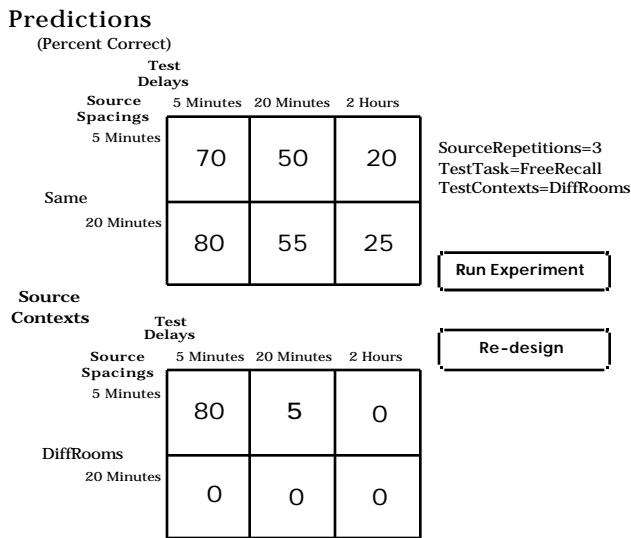


Figure 1. The interface used for making predictions.

A few words should be said about the form of the prediction task. In psychology, scientists are rarely asked to make precise numerical predictions. However, there are sciences in which one does make precise numerical predictions (indeed, in some sciences, predictions can only be made in quantitative terms because of the complexity of the theories). Moreover, it is not clear whether there is a simple method in a computer interface for asking students to make qualitative predictions for each of the factors (especially factors with 3 levels) and their interactions.

After making predictions, participants clicked on the 'Run' button and were shown the results of their experiments. Participants in an experimental condition that did not ask them to make predictions simply jumped straight to the experimental results. The results were shown in a table of the same format as was used to make predictions. If participants made predictions, the results table also showed their predictions (in smaller, italic text). Figure 2 presents an example results table (along with sample predictions). The correlation coefficient in the upper right is the Spearman correlation between the predictions and the actual outcomes, and was given to participants to provide a rough estimate of the accuracy of their predictions.

**Actual Outcome**

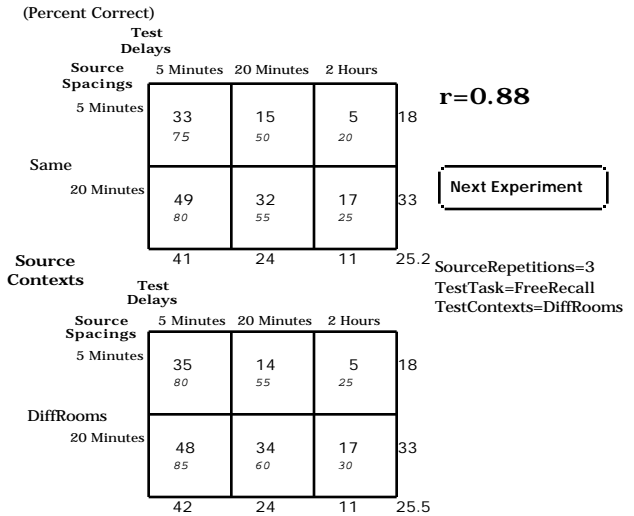


Figure 2. The interface used for displaying results (predictions, in italics, occurred only in the Prediction conditions).

For the purposes of this paper, there is one crucial performance dimension in this task: do participants focus on the theories under test? Previous research has shown that the majority of students in this task completely ignore the theories under test and simply focus on testing the effects of the 6 factors (Schunn & Anderson, 1998, 1999, In press). This focus on theories versus factors can be examined in two different ways. First, one can examine what types of conclusions the students make at the end of their experimentation: do the students make conclusions about the theories or the factors? Second, one can examine what types of experiments they design: do they focus on the factors that are actually relevant to the theories under test?

**Study 1**

**Methods**

**Participants** 56 George Mason undergraduates participated for course credit, of which 6 were removed due to computer problems. None of the participants had completed a research methods course, although a few (<10%) were currently enrolled in a research methods course.

**Procedure** Participants were randomly assigned to one of two conditions. Participants in the Prediction condition had to make numerical predictions for each cell in their experiments before viewing the outcomes of the experiment. By contrast, participants in the No Prediction condition skipped the numerical prediction phase entirely, both in the instructions and in the experiment itself.

Participants in both conditions were given a 15-minute tutorial on the computer that covered the spacing effect, the two theories, and how to use the Simulated Psychology Lab. The experimenter then reiterated the goals of the experiment (which had been presented on multiple computer screens including the very last one): to test the two theories of the spacing effect to determine whether one, both, or neither could account for the spacing effect. Participants worked on the task until they felt they understood the cause

of the spacing effect or until time had expired (40 minutes). Once finished, participants were asked what they found and their responses were recorded. They then answered a series of questions about the theories and any conclusions they came to about the effects of the six factors.

## Results & Discussion

**Overview** The results are broken into 3 sections. First, we verify that there were no background differences between the groups. Second, we examine the effects of the manipulation on what kinds of experiments the students generated. Third, we examine the effects of the manipulation on what kinds of conclusions the students made at the end of the task.

**Background Differences** To verify that the groups were roughly equivalent, we compared their reported SAT and status. There were no differences by group in either measure. For status, 18% and 21% of the undergraduates were upperclassmen in the Prediction and No Prediction groups respectively,  $\chi^2(1) < 1$ . For SAT, the combined Verbal + Quantitative scores were 1048 and 1052 for the Prediction and No Prediction groups respectively,  $F(1,53) < 1$ .

**Types of Experiments Conducted** The participants in the Prediction group ran marginally fewer experiments than did the participants in the No Prediction group, with means of 5.8 and 8.4 experiments respectively,  $F(1,55) = 3.4$ ,  $MSE = 29.4$ ,  $p < .1$ . This result is not surprising because the No Prediction subjects had more time to run experiments since they did not have to make predictions for each one.

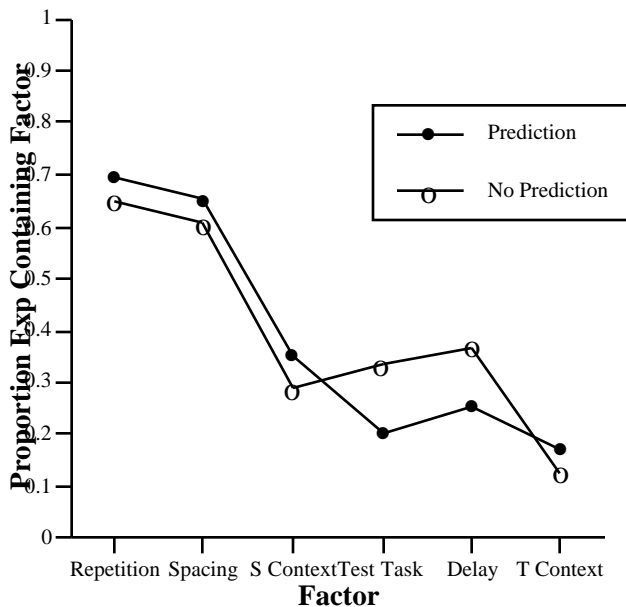


Figure 3. Mean proportion of experiments containing each of the factors within each group of Study 1.

More important than the number of experiments conducted are the types of experiments conducted. Figure 3 presents the proportion of experiments involving each factor. As one can see, the students in the Prediction group were generally more likely to focus on the factors relevant to

the theories under test (with the exception of the repetitions factor, which is the first option in the interface). Let Appropriateness be defined as the mean proportion of experiments involving Spacing, Source Context, and Test Context minus the mean proportion of experiments involving Repetitions, Test Task, and Delay. Then students in the Prediction group had a significantly higher Appropriateness score than the No Prediction students, with means of 0.01 and -0.11 respectively,  $F(1,53) = 4.1$ ,  $MSE = .05$ ,  $p < .05$ .

**Types of Conclusions Made** When the time was up or the students announced they were done, the experimenter asked the students what they had found. We coded whether the students responded to that question with a discussion of the factors that could be manipulated or a discussion of the theories under test. In the Prediction group, 31% of students mentioned the theories first, whereas in the No Prediction group, only 8% of students mentioned theories first,  $\chi^2(1) = 4.5$ ,  $p < .05$ . Thus, the manipulation did have a significant impact on whether the students focused on the theory testing nature of the task.

If they did not volunteer information about the factors at the end of the task, then the students were explicitly asked about each factor. There was no effect of the manipulation on the number of factors for which the students had correct statements about their effects, with means of 3.0 and 3.1 in the Prediction and No Prediction conditions, respectively,  $F(1,53) < 1$ . Thus, the difference in propensity to make conclusions about the theories at the end of the task was not a function of having learned less about the factors.

**Summary** Study 1 found that forcing students to making predictions did improve scientific reasoning in that problem. In particular, it led students to actually focus on the theories under test and manipulate factors relevant to those theories.

Study 2 examines whether these results generalize to a situation in which students have been given only one theory to test. Making predictions may only be helpful when it leads students to realize the key differences between theories and thus generate experiments that would tease the theories apart. Additionally, the frequency regularity theory is somewhat subtle and it may be that many of the students either did not understand it or did not know how to test it. Thus, in Study 2, students were only asked to test the shifting context theory.

## Study 2

### Methods

**Participants** 69 undergraduates participated for course credit, of which 2 were removed due to computer problems. None of the participants had completed a research methods course, although a few (<10%) were currently enrolled in a research methods course.

**Procedure** The procedure for Study 2 was identical to Study 1 with two exceptions. First, participants were never told about the frequency regularity theory and were given only the shifting context theory to test. Second, we did not

collect background information about the students (SAT, major, year, etc) since it did not prove predictive of performance in Study 1.

## Results & Discussion

**Types of Experiments Conducted** In study 2 the participants in the Prediction group ran approximately half as many experiments than did the participants in the No Prediction group, with means of 6.4 and 11.7 experiments respectively,  $F(1,65)=8.8$ ,  $MSE=52.6$ ,  $p<.01$ . Once again, this result is not surprising, since the No Prediction subjects had more time to run experiments because they did not have to make predictions for each one.

The size of the difference in number of experiments is larger than what was found in Study 1 and causes some problems for subsequent analyses. Specifically, it raises the question: are the differences in groups due to the number of experiments conducted or the cognitive consequences of the manipulation? Moreover, it appeared that in this Study, there were a significant number of participants running a very large number of experiments without much understanding (as many as 36 experiments in 40 minutes!)—they were simply clicking buttons. Therefore, we decided to remove from the remaining analyses all participants who ran more than 10 experiments (3 participants in the Prediction group and 6 in the No Prediction group). One consequence of this unequal reduction in condition Ns is that the subsequent condition comparisons should be more conservative tests of the manipulation: the ones removed from the analyses are more likely to not have understood the task and we have removed more of them from the No Prediction condition.<sup>3</sup>

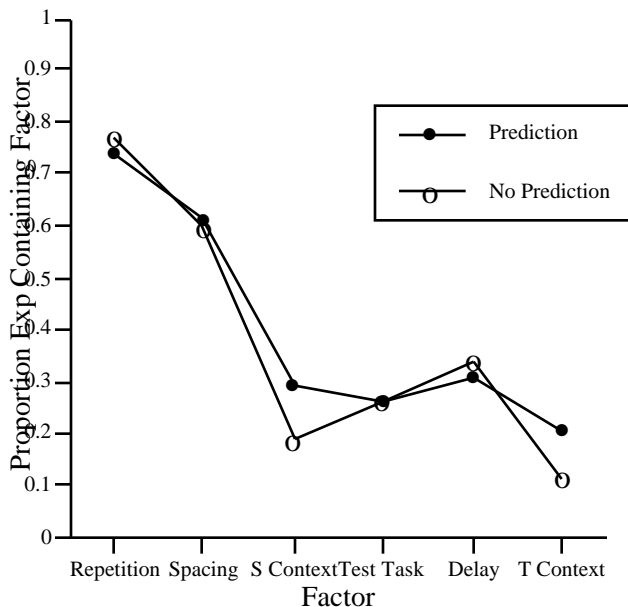


Figure 4. Mean proportion of experiments containing each of the factors within each group of Study 2.

<sup>3</sup> Moreover, the same patterns of results were found if all subjects were included in the subsequent analyses

The more central analysis was of the types of experiments conducted. Figure 4 presents the proportion of experiments involving each factor. As in Study 1, the students in the Prediction group were generally more likely to select the factors relevant to the theories under test. Since there was only the Shifting Context theory to test in Study 2, the Appropriateness measure must be redefined as the mean proportion of experiments involving Source Context and Test Context minus the mean proportion of experiments involving Repetitions, Spacing, Test Task, and Delay. Under this measure, students in the Prediction group had a significantly higher Appropriateness scores than the No Prediction students, with means of -0.12 and -0.35 respectively,  $F(1,47)=3.9$ ,  $MSE=.17$ ,  $p<.05$ .

**Types of Conclusions Made** As in Study 1, we coded whether the students responded to the final “what did you find?” question with a discussion of the factors that could be manipulated or a discussion of the theory under test. Students in the Prediction group mention the theory 11% of the time, whereas students in the No Prediction group never mentioned the theory on their own ( $\chi^2(1)=2.9$ ,  $p<.1$ ). Thus, the manipulation did have the same trend of an effect as in Study 1. This time, however, all students were quite unlikely to mention the theory on their own. It is possible that the students did not feel that the theory should be part of their final report since there was only one theory to test and they could not come up with an alternative theory.

As in Study 1, if the students did not volunteer information about the factors at the end of the task, then the students were explicitly asked about each factor. This time, however, there was a significant effect of condition on the number of factors for which the students had correct statements about their effects, with means of 3.2 and 3.8 in the Prediction and No Prediction conditions, respectively,  $F(1,46)=4.77$ ,  $MSE=0.98$ ,  $p<.05$ . That the No Prediction students had more correct responses establishes that the difference in propensity to make conclusions about the theory was not due to differences in what was learned about the factors.

Why did students in the No Prediction group produce a larger number of correct responses? It is likely that this effect occurred because the participants in the No Prediction task designed more experiments and explored more of the factors (especially the irrelevant factors). There were no differences between groups on the two most important factors. For source context, the Prediction group had a non-significantly higher proportion of correct responses (.31 versus .18,  $F(1,46)<1$ ). For the test context, the less relevant factor of the two, the Prediction group had a non-significantly lower proportion of correct responses (.54 versus .68,  $F(1,46)=1.0$ ,  $p>.3$ ).

## General Discussion

The two studies found generally quite consistent results: forcing students to making numerical predictions improves their scientific reasoning performance because it leads them to focus on the theories being tested and design more appropriate experiments.

The effects found in these studies were not large. However, the task given to the students is a very realistic scientific discovery task and was quite difficult for the students—in other words, there may have been relatively small improvements because the task was so difficult and there were possible floor effects in performance. Moreover, designing experiments which actually address the theories under test is such a central and important aspect of science. Any improvement from such a simple manipulation is important. Finally, previous research (Schunn & Anderson, In press) with this exact task has shown that even an entire course in research methods has relatively little impact on these same measures. Thus, that we found any improvement with such a simple manipulation is impressive.

While students were found to have a difficult time overall focusing on theories, we do not want to claim that most students could not focus on theories if the situation were made simple enough. However, that caveat is of little use to the educational setting in which students must learn to deal with experiments in real content domains. This consideration is what led us to use an authentic problem.

Our manipulations involve forcing students to make quantitative predictions for each cell in the design. What about other methods of generating predictions (e.g., generating more qualitative predictions)? Lehrer et al. (in press) argue that focusing on quantitative aspects of science is fundamentally important to scientific reasoning generally. However, one might imagine other methods for generating quantitative predictions. For example, what if one used graphical tools for generating predictions, or only forced predictions for each factor being manipulated and simple interactions among factors (rather than each individual cell)?

Our manipulation also focused on college students working on a problem in psychology. What about students working on problems in the physical sciences? One might imagine students in physics also losing track of the larger theories under test and focusing on the roles of particular concrete factors instead. Along those lines, Chabay and Sherwood (1999) have argued that giving physics students simulators that allow them to see the precise predictions of different theoretical assumptions improves students' understanding of the theories.

What about students in high school or elementary school? If university students lose track of the theories that are supposed to be tested, one can only imagine that this problem would be compounded in younger children. Indeed Deanna Kuhn's (1991) work suggests that children generally have a lack of differentiation between theory and evidence in scientific reasoning situations. However, whether making predictions actually improves performance for younger students (who may have other reasoning difficulties as well), is an open question.

### Acknowledgments

Thanks to Elizabeth Mazzanti for her help in running and analyzing Study 1. Work on this paper was supported by funds to the first author from the Department of Psychology and the College of Arts and Sciences at George Mason.

### References

- Chabay, R. W., & Sherwood, B. A. (1999). Bringing atoms into first-year physics. *American Journal of Physics*, 67, 1045-1050.
- Chinn, C., & Malholtra. (in press). Epistemologically authentic scientific reasoning. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implications from everyday, classroom, and professional settings*. Erlbaum.
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, 12, 1-48.
- Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 20(2), 461-469.
- Kuhn, D. (1989). Children and Adult as Intuitive Scientists. *Psychological Review*, 96(4), 674-689.
- Kuhn, D. (1991). *The skills of argument*. Cambridge, MA: Cambridge Press.
- Lehrer, R., Schauble, L., & Petrosino, A. (in press). Reconsidering the role of experiment in science education. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implications from Professional, Instructional, and Everyday Science*. Erlbaum.
- Loh, B., Reiser, B. J., Radinsky, J., Edelson, D. C., Gomez, L. M., & Marshall, S. (in press). Developing reflective inquiry practices: A case study of software, the teacher, and students. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implications from everyday, classroom, and professional settings*. Erlbaum.
- Okada, T., & Shimokido, T. (in press). The role of hypothesis formation in a community of psychology. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implications from everyday, classroom, and professional settings*. Mahwah, NJ: Erlbaum.
- Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology*, 49, 31-57.
- Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching*, 28(9), 859-882.
- Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science*, 23(3), 337-370.
- Schunn, C. D., & Anderson, J. R. (In press). Science education in universities: Explorations of what, when, and how. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implications from Professional, Instructional, and Everyday Science*. Erlbaum.
- Schunn, C. D., & Klahr, D. (1993). Self- vs. other-generated hypotheses in scientific discovery. In W. Kintsch (Ed.), *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, 51, 1-10.
- White, B. Y. (1993). ThinkerTools: Causal models, conceptual change, and science education. *Cognition & Instruction*, 10(1), 100

# Dueling Theories: Thought Experiments in Cognitive Science

Sam Scott (sscott@ccs.carleton.ca)

Department of Cognitive Science

Carleton University

Ottawa, ON K1S 5B6

## Abstract

Brook (1999) identified thought experiments as one of the key elements of philosophy's contribution to the cognitive sciences. In this paper, I tackle the question of how and why thought experiments work, and what exactly it is they do for us when they do work. I propose that thought experiments almost always involve two different theories of the world being compared to show that they do, or more often do not, fit together. Sometimes both theories are clearly articulated in the narrative of the thought experiment, but more often one of the two goes unarticulated - the thought experimenter instead relies on our shared folk theories of the world. The strength of some of the more famous and persuasive thought experiments lies in their ability to show that a given theory runs afoul of these deeply held folk intuitions. I will compare the "Dueling Theories" account of thought experiments to both Brook's "empirical" account and Brown's (1991) platonic account.

## Introduction

In Hilary Putnam's famous thought experiment, we are asked to imagine a possible world called "Twin Earth" identical in all respects to our own, but for the chemical composition of what is called "water" on both worlds. On our world, water is H<sub>2</sub>O, while on twin earth it is something else, say XYZ. Now we are asked to consider what happens when Adam on our world and his counterpart, "Twadam" on Twin Earth use the word "water". Do they mean the same thing? Apparently not. Even though their mental states are the same, the external referent is different. Hence, according to Putnam, "meanings just ain't in the head". (Putnam, 1975)

In Jackson's (1991) story of Mary the Colorblind Scientist, we are asked to imagine that Mary knows everything there is to know about color. That is to say, she knows everything that can be measured, described, and communicated about what color is and the process by which we perceive it. But she has never actually seen color before. Then one day she sees a color, say red, for the first time and learns something about color that she did not know before, namely what red looks like. The moral we are asked to draw is usually something about the uniqueness and indescribability of phenomenal experience.

In introducing these two famous thought experiments, what I am first interested in is how well they "work". Reactions tend to vary, but if you're like me, you will be immediately convinced by the story of Mary, but quite skeptical (at least at first) about the Twin Earth story. Perhaps your reactions differ, but the interesting question is

why we have the reactions we do. What makes a thought experiment "work" or "not work"? The thesis that I want to defend is that thought experimentation is a meta-activity - a duel between conflicting theories in which one appears to be a clear winner, thus challenging anyone who holds both theories. Thought experiments can never tell us something new about the world, because the world doesn't participate in the experiments. The objects of evaluation are theories of the world. On the other hand, thought experiments are a perfect device not only for revealing problems with various theories of the world, but in some cases, for making clear to us what our theories of the world actually are.

It may not be obvious that the two thought experiments described above fit this "Dueling Theories" story, but if we take the word "theory" to encompass both scientific and folk theory, then the story not only works but can be quite revealing. Consider Mary first. In one corner, we have some theory in which everything that is part of the physical world can be fully described in scientific terms. In the other corner, we have a folk intuition, based on our own experience of perception, that the phenomenal experience of color is indescribable and would therefore be unknowable to someone not directly acquainted with it. What makes this thought experiment so stunning to so many people is that they may not have realized that they held the folk theory, or that the folk theory was so difficult to give up, until it was put into direct conflict with the other theory. In the Twin Earth experiment, any theory that imparts semantic properties solely to brain states is supposed to lose to a reference theory of semantics. What makes someone embrace or reject Putnam's conclusion will be their pre-existing acceptance or skepticism about whether such reference theories of semantics can be made to work.

In what follows, I shall attempt to develop this Dueling Theories view in the context of two alternative views. The first, from Brown's *The Laboratory of the Mind* (Brown 1991), states that thought experiments can reveal a priori truths about the world through their investigation of Platonic universals. I hope to show that this idea should be rejected, mostly because it actually has very little to offer in aid of our understanding of the nature of thought experiments. The second view, from Andrew Brook's "Does philosophy offer cognitive science distinctive methods?" published last year in this forum (Brook, 1999), states that thought experiments get part of their value from their empirical content. This empirical content may make thought experiments capable of testing hypotheses against the real world. I believe that with

careful exegesis, Brook's view is actually quite close to my own. The main difference is that my conception allows in principle that there may or may not be empirical content to thought experiments, and in fact, a thought experiment need not actually have directly empirical content in order to make a useful contribution to science or philosophy.

### Platonic Thought Experiments

Brown's two paradigm cases of a priori reasoning about the world through thought experiments are Galileo's Coupled Falling Bodies and the photon-decay thought experiment by Einstein, Podolsky, and Rosen, which has become known as the "EPR Paradox". For the benefit of the uninitiated, both these experiments will be explained shortly.<sup>1</sup> But first, a short discussion of Brown's brand of Platonism is in order. According to Brown:

*A platonic thought experiment is a single thought experiment which destroys an old or existing theory and simultaneously generates a new one; it is a priori in that it is not based on new empirical evidence nor is it merely logically derived from old data; and it is an advance in that the resulting theory is better than the predecessor theory. (p.77, Brown's italics)*

Brown thinks that this kind of a priori thought experimentation is possible due to the existence of natural laws as abstract platonic universals, similar to the abstract objects that he supposes to exist in mathematics and logic.

Postulating abstract objects in a Platonic heaven is certainly a controversial move, and one that I am not sympathetic to. But rather than attempting to destroy the monolith of Platonism entirely, I will concentrate on trying to demonstrate that the Dueling Theories story does a better job of accounting for Brown's two favorite thought experiments than does his own Platonic account. But first I will digress briefly to point out two aspects of Brown's Platonism that are clearly problematic from a cognitive perspective.

### Cognitive Science and Abstract Objects

A central feature of Brown's defense of Platonism is its heavy reliance on the notion of "obviousness". According to Brown, without Platonism it is "an utter mystery why '3 > 2' seems intuitively obvious" (p. 56). Later, he continues, "if there were no abstract objects, then we wouldn't have intuitions concerning them; '2 + 2 = 4' would not seem intuitively obvious" (p. 64). Here we can take a lesson from research on mathematical cognition. What are we to make of the *non-obviousness* of mathematical facts like:

8329273847592 < 78374223847532, and  
89652 + 15265 = 104917 ?

---

<sup>1</sup> Unless otherwise referenced, more extensive descriptions of the thought experiments described here can be found in (Brown, 1991).

These mathematical facts are of the same order of complexity as '3 > 2' and '2 + 2 = 4'. That is, they use the same number of operators and the same number of arguments. Only the magnitudes differ. But the facts expressed are certainly not intuitively obvious to most people. The reason for the difference is easily accounted for in symbol-processing terms. We have memorized the order of the 10 digits, and this makes ordering judgements easy for single digit numbers, and more complicated for larger ones. Similarly, we have memorized all possible single-digit sums. Thus the obviousness of facts about small numbers and the *non-obviousness* of facts about large numbers can be accounted for in cognitive terms, whereas this difference is difficult to account for in Platonic terms. Perhaps some abstract objects are bigger than others? Or maybe some are further away than others? But what do "bigger" and "further away" mean in the abstract realm?

Brown also makes use of an analogy between sense perception of objects in the real world and our intuitions about objects in the abstract world. He defends this analogy by asserting "the perception of abstract laws of nature is certainly no more mysterious than [ordinary sense perception]." He justifies this by stating that,

at best we understand part [of visual perception] - the *physical* process starting with photons emitted by an object and ending with neural activity in the visual cortex. From there to *belief* about the object seen is still a complete mystery. (p. 87, Brown's italics)

Fair enough, but at least we do have: 1) an account of the links between an object and the visual cortex and 2) a research program capable of making progress in understanding how the visual system processes information from this link and communicates it to the rest of the brain. As far as I know, we have no account of the link between an abstract object and whatever organ of abstract sense perception we use to perceive it. Nor do we have an account of how these abstract perceptions would be processed in the brain.

The moral of this digression? Putting the logical neatness of Platonism aside and ignoring the natural fascination and attraction that many people feel towards abstract objects, there is much here to be suspicious of from a cognitive perspective.

### Coupled Falling Bodies and the EPR Paradox

Back to thought experiments. Galileo's Coupled Falling Bodies is one of Brown's favorite examples of a platonic thought experiment, and is at the very least an impressive piece of reasoning. Aristotle held the view that heavier bodies must fall faster than lighter bodies. This has now been refuted experimentally - we know that on earth, when air friction is removed, bodies of different masses fall at the same rate of acceleration. But before Aristotle's view was empirically refuted, Galileo supposedly refuted it with an act of pure thought. He did so by asking us to consider a

cannon ball and a musket ball attached by a string and asking what will happen when this entire assembly is dropped from the top of a tower.

According to Galileo, Aristotle would be forced into a contradiction. On the one hand, the combined system of musket ball, cannon ball, and string is heavier than the cannon ball alone. Therefore, the whole should fall faster than its parts. On the other hand, the lighter musket ball would try to fall more slowly than the heavier cannon ball, and would act as a drag on the entire system. Therefore the assembly should fall slower than its heaviest component, the cannon ball. So we have a contradiction. If  $C$  is the falling rate of the cannon ball, and  $S$  is the falling rate of the combined system, then both  $S > C$  and  $S < C$  are true. So Aristotle must be wrong. But the whole problem goes away if we propose that everything falls at the same rate ( $S=C$ ). Brilliant. But what is justifying this line of reasoning?

The theory that is clearly on trial here is Aristotle's. But another theory is being brought to bear here as well - our folk theory (or theories) of how objects behave in the real world. To see this, we have to poke around for the assumptions Galileo makes. The celebrated contradiction is derived from two inferences. First, Galileo assumes that if Aristotle is right, then the whole assembly should fall faster than the cannon ball alone. Why? Because if you put the whole assembly on a scale, it would weigh more than the cannon ball. So far so good. The second inference is that the musket ball must act as drag on the cannon ball. But what justifies this inference? Why does the small ball retain its autonomy as a lightweight when it becomes part of a heavier whole?

The answer to the above seems to be that we just *know* that it does! Imagine holding the two balls attached with string. If you put the musket ball in your left hand, and the cannon ball in your right, with the string stretched between the two, you just *know* that your right arm will tire more quickly. Our folk theory of the world, based on real experience, says that parts of wholes can in some cases be experienced as if they were autonomous. It is this folk theory that justifies the crucial second inference and allows Galileo to complete the contradiction. But what if our folk theory had been wrong? What if, despite our shared experience with lifting assemblies of objects, such assemblies in free fall actually do behave as integrated wholes? Well then, Galileo's conclusion would have been wrong, too. The moral of the story is that looking past the first theory on trial to the second (in this case unarticulated) theory that it is confronting leads us away from the idea that Galileo actually proved something through pure thought, and towards the more productive idea that he demonstrated that two theories were at odds and therefore one of them had to be discarded in favour of the other. The obvious choice for most people was to let go of Aristotle.

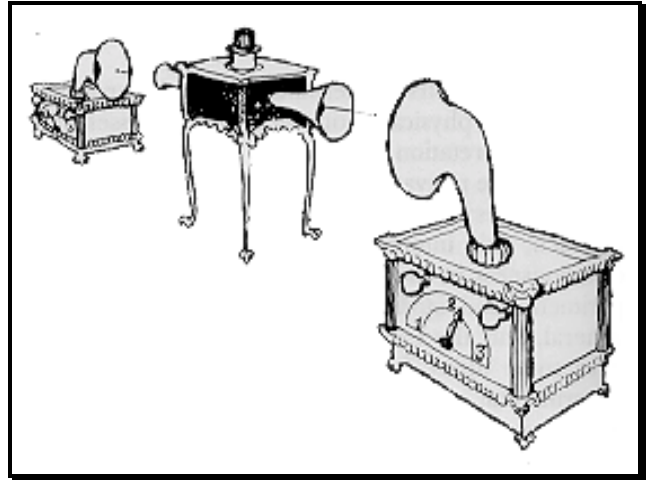


Figure 1: The apparatus for the EPR Paradox thought experiment. Picture from [www.reed.edu/~rsavage/epr.html](http://www.reed.edu/~rsavage/epr.html)

The EPR Paradox<sup>2</sup> is another of Brown's favorite platonic thought experiments, although this is a slightly less impressive story because in the end the conclusion of the thought experiment was tested and refuted in a real experiment. The target of the EPR Paradox was Heisenberg's uncertainty principle in particular, and the Copenhagen interpretation of quantum mechanics in general. Heisenberg's uncertainty principle states that we can never know the complete trajectory of a quantum particle. "Trajectory" here means position plus momentum, where momentum is a vector that encodes mass, as well as direction and velocity. The problem is that in order to observe a quantum particle, we have to bounce another quantum particle off it, and this interaction will necessarily change one of the two aspects of its trajectory. The Copenhagen interpretation of this result stated that the indeterminacy existed not in our knowledge of reality, but in reality itself. That is, the trajectory of a quantum particle is not actually fully determined until it is observed.

Einstein, Podolsky and Rosen did not accept the Copenhagen interpretation - they believed that the theory was incomplete, not that reality was indeterminate. To prove their point, they published a 1935 *Physical Review* article in which they proposed the following thought experiment. Imagine two quantum particles that interact and then fly apart. We could later measure one of the particles for either momentum or position and automatically, through conservation laws, deduce the corresponding property in the other particle, even though the particles were far apart and could no longer interact. (Einstein et al., 1935)

The crucial step in the EPR reasoning was to notice that we could make *either* measurement of the first particle, and deduce the corresponding measurement of the second

<sup>2</sup> The name of this thought experiment is rather misleading. "EPR" stands for Einstein, Podolsky and Rosen, but don't expect to find an actual paradox anywhere.



particle. Since the two particles are now far apart, and could not be influencing each other, the second particle must have predetermined real values for both position and momentum. It must not be indeterminate with respect to any property. Einstein et al. were aware that they were making a crucial assumption here, namely that distant particles cannot influence each other. They named this the *locality assumption* and even went as far as to suggest that quantum mechanics be modified to include this assumption, thus making it a complete theory again. That is to say, they put the Copenhagen interpretation into the ring with another theory (one that Brown calls *local realism*). Once again, as seems to happen so often in successful thought experiments, the second theory accords well with our folk intuitions about the world - so much so that the EPR Paradox became a powerful tool for the anti-Copenhagen crowd. Unfortunately for them, John Bell in 1964 managed to derive a mathematical result that could be used to pit local realism against the Copenhagen interpretation in a *real* experiment, and local realism made the wrong prediction. (See Brown, 1991 and Bell, 1987)

The main point in all of this is that we get further in understanding these thought experiments if we always assume that two theories must be doing battle. Brown's Platonic view of what Galileo and Einstein et al. were up to does not help us to decide whether their conclusions were right or wrong. That is to say, the assumption that they were exploring abstract universals does not help us decide whether the particular universals they discovered correspond to the universals that happen to operate as natural laws. Furthermore, the Platonic view does not give us any real insight into what the thought experimenters were actually up to in either case. Specifically, there is nothing in the Platonic story to explain what justified Galileo's reasoning or why the EPR Paradox failed in the end. Platonism just doesn't buy us anything here. On the other hand, viewing their work through the lens of Dueling Theories forces us to make clear which theories were being tested, leading in the case of Coupled Falling Bodies to the exposure of an unstated piece of folk theory, and in the case of the EPR Paradox, to a simple account of why the result, which seemed so persuasive, did not stand up in a real experiment.

### Other Types of Thought Experiments

Although Brown focuses most of his attention on Platonic thought experiments, his taxonomy actually includes a number of other categories as well. Brown's full set of thought experiment categories is reproduced in Figure 2 below. Thought experiments are broken down into two main groups, *destructive* and *constructive*. Roughly speaking, constructive thought experiments build new theories, while destructive thought experiments invalidate old ones. Constructive thought experiments are further subdivided into *direct* (those that begin with common, unproblematic phenomena and end with a well articulated theory), *mediated* (those that start with a well articulated theory and

help to reach a new conclusion), and *conjectural* (those that start with conjectured, rather than common and unproblematic phenomena and end with a well articulated theory.) Platonic thought experiments, such as Coupled Falling Bodies and the EPR Paradox, are those that are both destructive *and* directly constructive. That is, they start from common, unproblematic phenomena and end in both the destruction of an old theory and the construction of a new one in it's place.

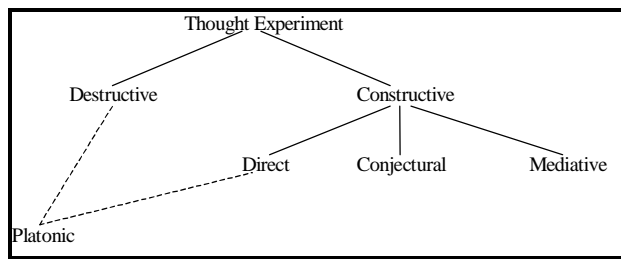


Figure 2: Brown's taxonomy of thought experiments. Adapted from (Brown, 1991)

Most of the examples Brown uses in any category are well accounted for by the Dueling Theories idea, but in some of Brown's examples, one has to dig a little deeper to get at the "other" theory involved. Perhaps the best example of this is Schrödinger's Cat, a destructive thought experiment aimed at the Copenhagen Interpretation of quantum mechanics (Schrödinger, 1935). In this experiment, a quantum event is amplified to have an effect on the macroscopic world - a cat in a box either dies or does not. The idea is that if the Copenhagen interpretation is correct, and reality is indeterminate, then the cat in the box is both alive and dead at the same time. The second theory coming into play here is just our deeply entrenched theory of the determinacy of the world around us. Calling us on this theory is a brilliant move, in which Schrödinger bets that scientists can accept quantum uncertainty only as long as they can cut it off from their folk intuitions about how the macroscopic world works - a sort of philosophical "not in my back yard" mentality.

The only problematic area for the Dueling Theories story is in Brown's "mediative constructive" category. Mediative experiments seem to play an articulatory role only. In these cases, a theory is already well established, and the thought experiment behaves like a diagram illustrating some particular, perhaps counter-intuitive, aspect of it. Most of the examples here are difficult to analyze for competing theories, but I suspect that is because they are not the same sort of thing as the thought experiments in the other categories. Consider one of Brown's examples - Maxwell's demon. One of the consequences of the theories Maxwell was developing was that heat could (with infinitely small probability) flow from a cold body to a hot body, rather than the other way around. This is, of course, wildly counter-intuitive. So Maxwell told a cute little story about a chamber

of hot gas molecules and another of cold gas molecules connected by a gate. A little demon controls the gate and only lets fast (hot) molecules enter the hot chamber and slow (cold) molecules enter the cold chamber. All this story really does for us is provide a framework for understanding a counter-intuitive theory. Maxwell was not trying to prove anything new, and as a result neither Brown's conception of thought experiments nor my own has much to say about his thought experiment. I suspect we would do much better to reclassify mediative thought experiments as "thought experimental illustrations" and leave it at that.

### Thought Experiments in Cognitive Science

The most famous thought experiments in the cognitive sciences are destructive in nature. This is true of both Putnam's Twin Earth experiment, in which the target is all non-reference-based theories of semantic content, and of Jackson's Mary the Colorblind Scientist experiment, in which the target is the scientific descriptibility of phenomenal experience. It is also true of Searle's (1980) Chinese Room, arguably the most famous thought experiment in cognitive science (practically talked to death in the last 20 years.) For reasons of space, I will apply the Dueling Theories analysis in detail only to Searle's experiment - I will not develop the analysis of Putnam or Jackson any further. I want to show: a) that the Dueling Theories analysis can be applied, and b) that it would lead Searle's opponents directly to a particular line of attack.

The theory that Searle is attempting to demolish with the Chinese Room is the "Strong AI" theory that a computer programmed to behave indistinguishably from a human who understands Chinese would *really understand* Chinese. The thought experiment works by trading on our strong folk intuition (i.e. theory) that a non-Chinese-speaking human taught to behave like [a computer programmed to behave indistinguishably from a human who understands Chinese] would *not really understand* Chinese. Laying out the two theories explicitly makes it easy to see that in order for them to be in conflict, the computer and the human must be equivalent in their ability to execute computer programs. But that would clearly not be the case in the real world. Computers are much better and faster at executing programs than humans. Therefore, Searle must have a fictional super-human in mind, which leads us to ask why we feel entitled to any a priori judgements about what such humans would or would not "understand" about the task they were performing. The Dueling Theories analysis thus leads directly to a productive line of attack on Searle - one that has, in fact, been pursued in the literature (Hofstadter and Dennett, 1982).

### The Empirical Basis of Thought Experiments

Brook's (1999) paper emphasizes a slightly different way of looking at thought experiments - one which seems at least partially at odds with Brown's conception, and which differs in emphasis from mine. He asks:

...are thought experiments also empirical, at least in part? Yes; they are merely a particular way of manipulating material stored in memory, material originally gained from experience. (p. 104)

Since the materials and at least most of the relationships of the imagined situation are derived from experience, thought experiments are thus a *kind* of empirical investigation. (p. 106, Brook's italics)

Two related questions lurk in the above. First, what is what the nature of the "material stored in memory", and second, what is the sense of "empirical" that Brook is employing?

Concerning the first question, it seems clear that the material referred to must not simply be memories of events that the thought experimenter has witnessed. If that were the case, then most observers would have no opinion about what might happen even in the somewhat ordinary situation imagined in the Coupled Falling Bodies experiment, to say nothing of the strange situations of the EPR Paradox, Twin Earth, Schrödinger's Cat, the Chinese Room, and Mary the Colorblind Scientist. The material stored in memory must have some kind of predictive power in order for it to be applied to the novel situation of a thought experiment. Therefore, it must consist of generalizations over past experience - that is, theories, folk or scientific, about the world.

Concerning the second question, the word "empirical" is a loaded one, and calling an activity empirical can have a number of different meanings. In the strongest sense, an empirical activity might consist of observing events in the world and carefully recording them for future use. But that is clearly not the sense being used here. In a weaker sense, an empirical activity might be any activity that takes into account memories of actually experienced events. But for reasons stated above, this cannot be the correct sense either - we could not perform thought experiments based solely on remembered events. But there is another sense in which an activity is empirical if it involves generalizations (theories) that were derived, at least in part, from actual empirical experiences. This must be closest to the sense that Brook has in mind, and it also fits reasonably well with the Dueling Theories idea.

Brook goes on to situate thought experiments within an abductive, Popperian view of scientific progress, pointing out that thought experiments have historically played a role on both sides of the generate and test paradigm. It is easy to see how hypothesis generation can be aided by thought experiments. As Brook notes, "hypothesis generation is pretty much a pure act of the imagination." As such, it's not hard to see that thought experimentation can help here. Lots of counterfactual "what ifs" are bound to be involved, each of which is likely to be a thought experiment. On the Dueling Theories account, a thought experiment can also help to crystallize a folk theory that previously went unnoticed. This is not exactly the same thing as generating a hypothesis, but for the purposes of scientific investigation, it

is the same thing - in order to test a theory, it must first be articulated and acknowledged.

Where I take some issue with Brook is concerning the role of thought experiments on the testing side of the generate and test paradigm. On the one hand, Brook makes the uncontroversial claim that thought experiments *have* historically played a role in testing hypotheses - the paradigm case being Galileo's rejection of Aristotelian physics based on the Coupled Falling Bodies experiment. Where Brook and I may part company is on the question of whether thought experiments *ought to* play such a strong role in testing hypotheses. Recall that on the Dueling Theories account, Galileo's thought experiment made use of an unarticulated folk theory of ordinary objects to make it's point. Hence, his conclusion should not have been that Aristotle's theory was ruled out as a possibility but that the theory ran afoul of some very ordinary intuitions about the world. This is a very important result in and of itself. It's just that the appropriate reaction was not for Galileo to smile triumphantly from his armchair, but for him to get up and figure out how to test the two theories with a *real* experiment.

The goal of Brook's treatment (and my own) is to see how much thought experiments can do for us in the cognitive sciences. Brook's emphasis on the empirical nature of thought experiments leads to the idea that they can in some cases stand in for real hypothesis testing.<sup>3</sup> What I hope to have shown with the Dueling Theories account is that the usefulness of thought experiments actually lies in their meta-level ability to test theories of the world against each other, and not in their ability to test the world directly. There are two further reasons why I would avoid using the term "empirical" to describe thought experiments - one technical and one psychological. The technical reason is that folk theories and intuitions may not need to be based on anything empirical at all. In fact, there are some thought experiments, such as Putnam's Twin Earth, Searle's Chinese Room, and even the EPR Paradox for which it is difficult to nail down exactly what the relevant empirical evidence supporting our intuitions actually is. For that reason, it's safer to view thought experiments as a theory evaluation activity and leave the question of the empirical nature of the theories under test open to be evaluated on a case by case basis. The psychological reason is that we might mistakenly encourage the idea that thought experiments are empirical in a stronger sense than what we really mean. And if we do that, we've done ourselves a disservice.

## Conclusion

If I am right, then thought experiments are best viewed as a picturesque arena in which two competing theories of the world do battle. Sometimes the Dueling Theories are explicitly stated ahead of time, but often one of the two

---

<sup>3</sup> From personal communication. Brook also discusses the use of thought experiments for the elimination of possibilities. This idea also relies on the characterization of thought experimentation as an empirical activity.

lurks in the background assumptions. In fact, what gives thought experiments such as Coupled Falling Bodies and the EPR Paradox their rhetorical power is precisely their accord with some of our most deeply entrenched folk theories of the world of middle-sized dry goods. Thought experiments are more useful if they clearly contain some kind of empirical material. This empirical material is not direct observation, but instead takes the form of predictions encoded by theories formed on the bases of empirical data. But this is not the same as saying that thought experiments can be empirical in the way that real experiments are. Despite their potential for polemical power, thought experiments can never actually tell us anything about the world. No thought experiment can ever be as good as the corresponding real experiment.

## Acknowledgements

I would like to acknowledge the help and support of Andy Brook. Thanks also to Ron Boring, Leo Ferres, and Finn Makela for many useful discussions. This work was supported in part by Canada's National Sciences and Engineering Research Council (NSERC).

## References

- Bell, J. S. (1987). *Speakable and Unsayable in Quantum Mechanics*. Cambridge: Cambridge University Press.
- Brook, A. (1999). Does philosophy offer cognitive science distinctive methods? *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 102-108). New York: LEA.
- Brown, J. R. (1991). *Thought Experiments in the Natural Sciences*. London: Routledge.
- Einstein, A., Podolsky B. and Rosen N. (1935). Can quantum mechanical description of reality be considered complete? *Physical Review*.
- Hofstadter, Douglas R. and Dennett, Daniel C. (1982) Reflections. In Douglas R. Hofstadter and Daniel C. Dennett (Eds.) (1982). *The Mind's I: Fantasies and Reflections on Self and Soul*. Toronto: Bantam Books. 373-382.
- Jackson, F. (1991). What Mary didn't know. *Journal of Philosophy*. 83-5: 291-295. Reprinted in D. Rosenthal (Ed.) *The Nature of Mind*. New York: Oxford University Press.
- Putnam, H. (1975). The meaning of meaning. *Mind, Language and Reality: Philosophical Essays*. Cambridge: Cambridge University Press.
- Schrödinger, E. (1935). The present situation in quantum mechanics. Translated and reprinted in J. Wheeler and W. Zurek, *Quantum Theory and Measurement*. Princeton, NJ: Princeton University Press.
- Searle, John R. (1980). Minds, brains, and programs. *BBS*. Volume 3. Reprinted in Hofstadter and Dennett. *Op. Cit.* 353-373.

# Temporal Progression of the Cortical Potential Distribution for the AEP P300 Component in Mild Traumatic Brain Injury

Robert D. Sidman (rds7637@louisiana.edu)

Department of Mathematics & Institute of Cognitive Science

University of Louisiana at Lafayette

Lafayette, LA 70504-1010

Lan Ke (ke@louisiana.edu)

Department of Mathematics, University of Louisiana at Lafayette

Lafayette, LA 70504-1010

Martin R. Ford

Consultant, Gainesville VA

## Abstract

An objective scoring system has been developed to quantify degrees of auditory evoked potential (AEP) abnormalities in patients suffering from mild traumatic brain injury (mTBI). In this study the AEP P300 responses for 20 patients with scores in the abnormal range were compared to the responses in 20 age and gender equivalent controls. The cortical imaging technique (CIT) was used to calculate the evolution of potential changes on the surface of the brain during a 50-millisecond epoch containing the AEP P300 response for both groups of subjects. AEP P300 condition recordings were obtained from 20 EEG and 2 EOG artifact channels. Previously published CIT results showed anterior and posterior peaks, implying multiple sources, for the P300 component. This study suggests that the anterior sources are significantly attenuated in the patient group and this anterior P300 attenuation appeared relatively more focal than that seen with scalp topographical maps alone and that the effects of the injury appear to be selective at the anterior sites.

## Introduction

This study compares the auditory P300 evoked responses of a group of subjects who have suffered mild traumatic brain injuries (mTBI), with those of a group of age and gender matched normal controls. Analyses were conducted using the cortical imaging technique (CIT), a mathematical procedure for constructing activity as it theoretically would appear on the cortical surface. The goal is to detect subtle differences in the evoked response that are not apparent from the scalp recordings, and may suggest the intracerebral site of injury.

A growing body of radiological, neuropsychological, electrophysiological, neuropathological and experimental evidence indicates that mild brain injury may occur in the absence of direct impact to the head, and in cases where there is no loss of consciousness [Binder, 1986; Evans, 1992a,b]. MRI, CT, routine EEG and conventional evoked potential (EP) recordings are often unremarkable in these cases, yet the patients may have sequelae that affect professional functioning and activities of daily living for months or years.

Sequelae of mild traumatic brain injury (TBI) have been

thoroughly reviewed [Evans, 1992c]. The most common of these are headaches, dizziness, blurred vision (or other visual disturbances), memory impairment, attention/concentration difficulties, increased reaction and information processing times, and personality changes including increased irritability, anxiety and depression.

Ford and Khalil [1996a,b] have developed an objective scoring system for AEP, VEP and EEG findings from patients with suspected mTBI. In these studies, significant group differences were found between patients and controls, a relatively objective scoring system was developed, and the patterns were used to identify about 60% of the patients, with no false positives. The following analysis is limited to the auditory P300 response.

The auditory P300 component is one of the most investigated of all the cognition-driven evoked responses. It is a prominent, posterior vertex positive component, peaking at approximately 300-350 ms post-stimulus in response to randomly presented "oddball" stimuli that are **counted** or otherwise **identified by the subject**.

The generator sources of the auditory P300 component have been suggested in the hippocampus [Halgren et al., 1980; Wood et al, 1984; Neshige & Luders, 1992] and, possibly, with neocortical localization with deep frontal and thalamic contributions [Yingling & Hosobuchi, 1984; Wood et al., 1984; Neshige & Luders, 1992]. Recent work [Pilgreen, 1995; Gevins & Cutillo, 1995] expand on these earlier studies.

In this study we compared the auditory P300 responses of two groups - a patient group identified by the scoring system referred to above with a normal group of age and gender matched controls. The comparison was performed as follows: the potential field for each subject was approximated at the cortical surface using the cortical imaging technique [Sidman, 1991] for a 50 msec epoch extending from 25 msec prior to the latency of the maximum voltage at Pz (the P300 latency) to 25 msec after the maximum. The average normal response and the average abnormal response on the cortical

surface were calculated and compared using t-scores. The average normal response during the N2a/P300/N3 endogenous complex of responses to the rare "oddball" auditory stimulus has been discussed and analyzed in [Ford, Sidman & Ramsey, 1993].

### **Mathematical Analysis-The Cortical Imaging Technique (CIT)**

The mathematical method that was used to perform the analyses reported here is the cortical imaging technique [Sidman, 1991]. Briefly, this method is a way of simulating the potential field on the cortical surface, presumably closer to the actual generators of the scalp-recorded field.

### **Subjects and Measuring Procedure**

Twenty normal subjects and twenty mTBI patients (as identified in [Ford and Khalil, 1996a]) were included in this study. Twenty channels of EEG and two channels of EOG activity were recorded with a NeuroScience Brain Imager, Series III (for amplification and on-line filtering) and a NeuroScan, Inc. software based system using two Dell 433/L PCs (for stimulus presentation, A-to-D conversion, data recording and off-line baseline correction, filtering, artifact minimization, averaging and group statistics).

Electrode sites were the twenty standard International 10-20 system placements (FP1/2, F3/4, F7/8, T3/4, C3/4, T5/6, P3/4, O1/2, Fz, Cz, Pz and Oz), with a linked ear reference. Vertical and horizontal eyeball and eyelid movements (VEOG and HEOG) were also recorded for possible artifact rejection. The recording protocol for all subjects included both resting and P300 auditory.

In the AEP P300 recording, 165 total responses were recorded to the frequent tone (1kHz, 95dB, 50ms duration, binaural) and to the "oddball" or rare tone (2kHz, 95dB, 50ms duration, binaural). There was a variable inter stimulus interval of 1.9-2.1 secs. The probability of an "oddball" or rare tone occurrence was 20%, and the same sequence of frequent and rare stimuli was presented to all subjects. The subjects were instructed to count the rare tone silently to themselves, and their answers were recorded at the end of the session.

The low pass filter was set at 100Hz, high pass filter set at 1.05Hz, with the 60Hz notch filter activated. Sweeps were automatically rejected for artifact if the voltage in either VEOG or HEOG channels exceeded  $\pm 100 \mu\text{v}$ . Analog to digital processing was performed at 320 Hz yielding 256 points and an EP epoch from -100ms to 700ms (3.125ms resolution) for all EP recordings.

### **Results**

All of the mTBI patients (20) and all but one of the normal controls (19) exhibited a well defined AEP P300 component,

as identified by a distinct voltage maximum primarily at Pz. The standard deviation of the latencies was 25 ms. In each case scalp (Figure 1) and cortical surface (Figure 2) potential maps were constructed for the epoch extending from 25 ms (one standard deviation) prior to the latency of the peak to 25 ms after the peak. The time point  $t_0$  is the latency of the AEP P300, different for each individual, and pairs of consecutive time points are 3.125 ms apart. The evoked responses were aligned at time point  $t_0$  for each group of subjects to obtain the average normal response and average patient response for 50 ms (see [Ford, Sidman & Ramsey, 1993]), on the scalp (Figure 1) and cortical surface (Figure 2). The SURFER® software was used throughout for making the graphical displays. The right-most column of pictures for each figure compares the two sequences of pictures by plotting the t-scores (degrees of freedom = 37) at each of the 20 scalp sites, for Figure 1, and for 160 cortical surface sites, for Figure 2.

Ford et al., 1993 [Ford, Sidman & Ramsey, 1993], contains a similar analysis for the entire ~250 ms epoch containing the N2a/P300/N3 complex of responses for normal subjects.

T-score differences between groups show significant attenuation in the patient group at time points leading into and including the P300 component peak, but not afterward. The differences were significant at the anterior source only.

### **Conclusions**

Scalp topography of the auditory P300 component typically shows the peak amplitude at posterior vertex area sites, with no strong indication of more than one generator site. In previous studies with CIT, using 28 channel data [Ford, Sidman & Ramsey, 1993], we have demonstrated that there are several apparent sources in normals, including an anterior one and symmetric, bilaterally homologous, centro-parietal sources. In another previous study, we found that individuals with histories and symptoms consistent with mild TBI showed attenuation of the scalp-recorded auditory (and visual) P300 component, with significant differences spread widely across all anterior sites. The primary purpose of the present study was to analyze the P300 recordings from the mild TBI cases using CIT to determine whether CIT was more sensitive in identifying group differences.

The results of the P300 scalp recordings from both groups showed the expected P300 peak in the area of Pz, with no anterior source indicated. Statistical differences between groups were negligible. CIT analyses of the recordings in both groups showed two clear sources - including one from Fz and the surrounding area. When compared statistically, the differences between groups were significant at the anterior - but not posterior - source. These results indicate that whatever injury is present in these cases differentially affects the anterior source contributing to the generation of the P300 component, while the posterior source remains unaffected.

Although it is not possible to determine the exact mechanism nor locus of injury, based upon the analysis of 20-channel scalp-recorded EP data, the results are not inconsistent with other reports of abnormal findings involving anterior or subcortical regions. (Parenthetically, the availability of 128- and 256-channel recordings may make the elucidation of the mechanism involved possible.) The purpose of the study was to determine whether CIT analyses provided additional information beyond that obtained from scalp recordings, and the results suggest a focal effect at the anterior midline area, thus suggesting subcortical involvement, since the differences from adjacent areas over frontal cortex were not significant.

In summary, the P300 has been shown to have significant amplitude attenuation and/or latency increases in association with a host of clinical disorders and conditions, including schizophrenia, dementia, alcoholism, brain injury, and some instances of learning disability, to name a few. The typical scalp recorded P300 waveform shows amplitude attenuation and/or latency increases across diagnostic groupings, and is thus concluded to be a nonspecific indicator of dysfunction. However, further analyses using analytical techniques such as CIT may ultimately show differential effects on the generator sources for the P300 component among groups, thereby yielding information that is valuable in understanding the underlying processes involved.

## References

- Binder, L.M. (1986), Persisting symptoms after mild head injury: A review of the postconcussive syndrome. *Journal of Clinical and Experimental Neuropsychology*, 8, 323-346.
- Evans, R.W. (1992), Mild traumatic brain injury. In: G.H. Kraft and S. Berrol (Eds.), *Physical Medicine and Rehabilitation Clinics of North America, Traumatic Brain Injury, Vol.3*, W.B. Saunders Company, Philadelphia.
- Evans, R.W. (1992), The postconcussion syndrome and the sequelae of mild head injury. In: R.W. Evans (Ed.), *Neurologic Clinics, The Neurology of Trauma, Vol.10*, W.B. Saunders Company, Philadelphia.
- Evans, R.W. (1992), Some observations on whiplash injuries. In: R.W. Evans (Ed.), *Neurologic Clinics, The Neurology of Trauma, Vol.10*, W.B. Saunders Company, Philadelphia.
- Ford, M.R. and Khalil, M. (1996), Evoked potential findings in mild traumatic brain injury 1: Middle latency component augmentation and cognitive component attenuation. *J. Head Trauma Rehabil.*, 11(3) 1-15.
- Ford, M.R. and Khalil, M. (1996), Evoked potential findings in mild traumatic brain injury 2: Scoring system and individual discrimination. *J. Head Trauma Rehabil.*, 11(3) 16-21.
- Halgren, E., Squires, N.K., Wilson, C.L., Rohrbaugh, J.W., Babb, T.L. and Crandall, P.H. (1980), Endogenous potentials generated in the human hippocampal formation by infrequent events. *Science*, 210, 803-805.
- Wood, C.C., McCarthy, G., Squires, N.K., Vaughan, H.G., Woods, D.L. and McCallum, W.C. (1984). Anatomical and physiological substrates of event-related potentials: two case studies. *Annals of the New York Academy of Sciences*, 425, pp. 681-721.
- Neshige, R., and Luders, H. (1992), Recording of event-related potentials (P300) from human cortex. *J. Clinical Neurophysiology*, 9(2), 294-298.
- Yingling, C.D. and Hosobuchi, Y.A. (1984), A subcortical correlate of P300 in man. *Electroenceph. Clin. Neurophysiol.*, 59, 72-76.
- Pilgreen, K.L. (1995), Physiologic, medical, and cognitive correlates of electroencephalography. In: Paul L. Nunez (Ed.) *Neocortical Dynamics and Human EEG Rhythms*, Oxford University Press.
- Gevins, A.S. and Cutillo, B.A. (1995), Neuroelectric measures of the mind. In: Paul L. Nunez (Ed.) *Neocortical Dynamics and Human EEG Rhythms*, Oxford University Press.
- Sidman, R.D. (1991), A method for simulating intracerebral potential fields: the cortical imaging technique. *J. of Clinical Neurophysiology*, 8(4), 432-441.
- Ford, M.R., Sidman, R.D. and Ramsey, G. (1993), Spatio-temporal progression of the AEP P300 component using the cortical imaging technique. *Brain Topography*, 6(1) 43-50.

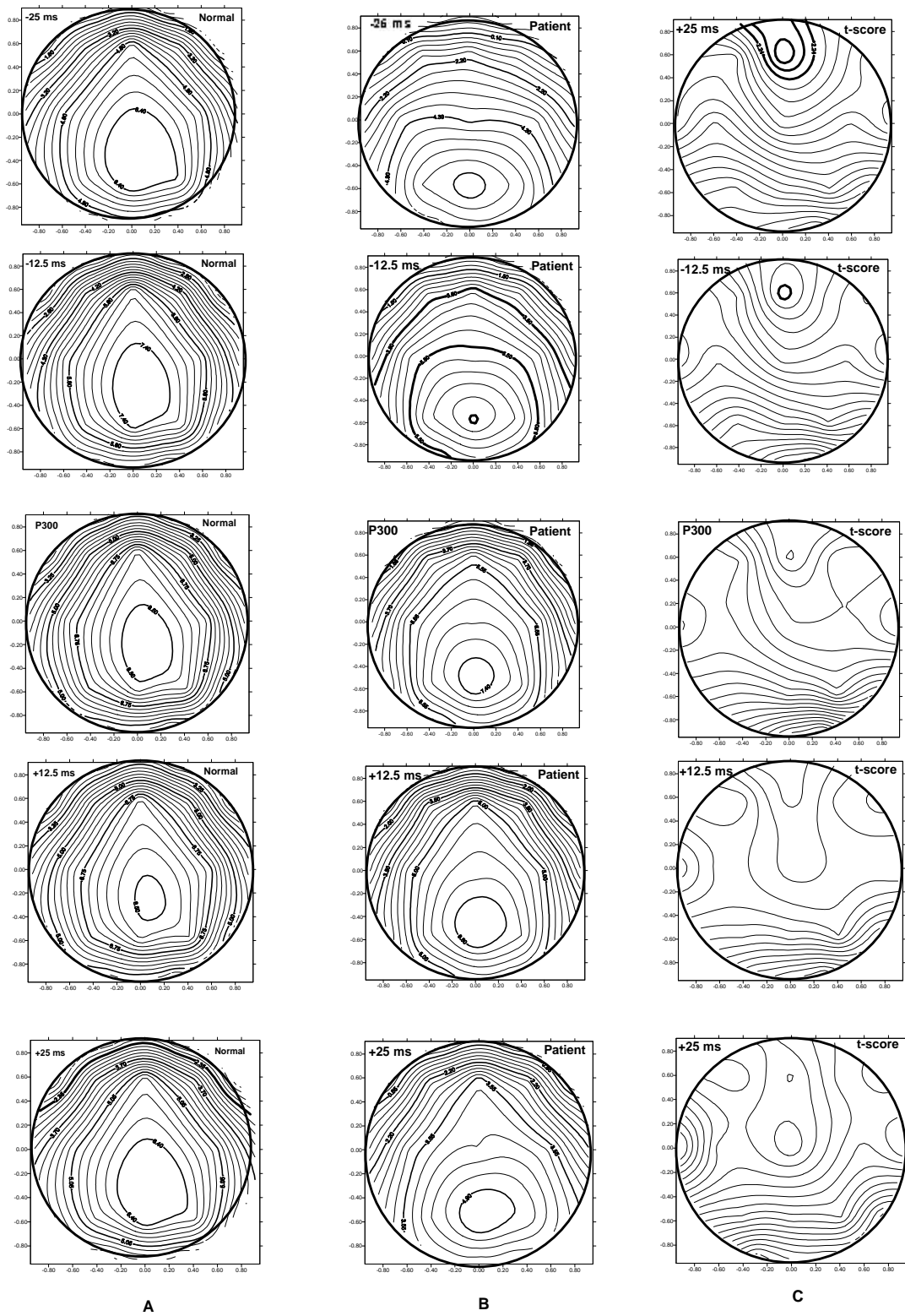
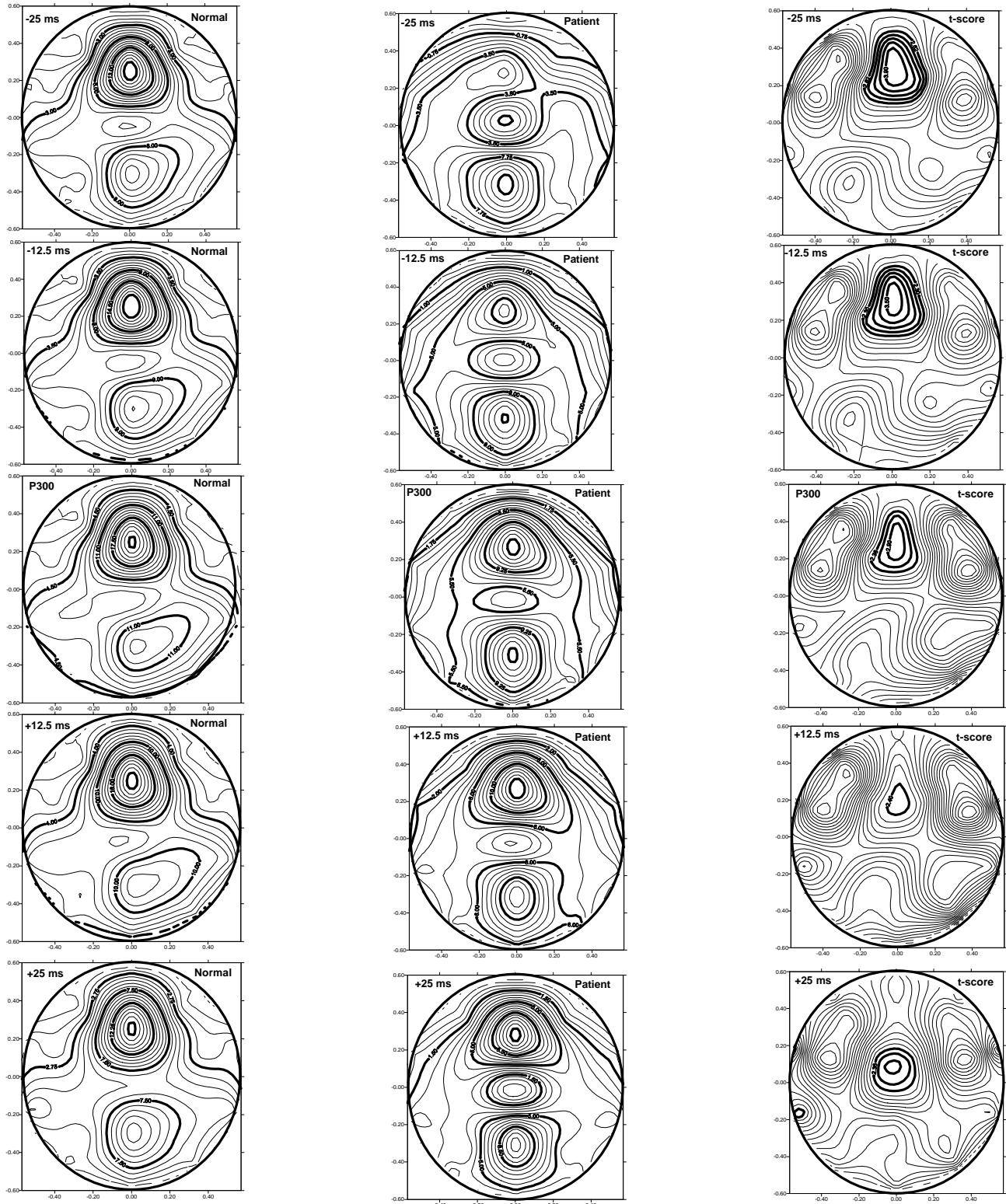


FIGURE 1



A

B

C

FIGURE 2



### Figure Legends

**Figure 1 - 1a)** Scalp-recorded voltages for the average normal subject at latencies ranging from 25 ms. prior to P300 to 25 ms. after P300. Values are in microvolts. **1b)** Scalp-recorded voltages for the average mTBI subject at the same relative latencies as in 1a). **1c)** t-score comparison of the scalp contour plots in 1a) and 1b). Regions where the scalp-recorded voltages for the average patient are significantly attenuated ( $p < .02$ ) in comparison with the average normal subject are highlighted with **bold contours**. These pictures each represent an overhead view of the scalp, modeled as a sphere of radius 1.0, in which the top-center is the nasion.

**Figure 2 - 2a)** Voltages on the cortical surface (as computed by **CIT**) for the average normal subject at latencies ranging from 25 ms prior to P300 to 25 ms after P300. Values are in microvolts. **2b)** Voltages on the cortical surface for the average mTBI subject at the same relative latencies as in 2a). **2c)** t-score comparisons of the cortical surface contour plots in 2a) and 2b). Regions where the cortical surface voltages for the average patient are significantly attenuated ( $p < .02$ ) in comparison with the average normal subject are highlighted with **bold contours**. These pictures each represent an overhead view of the cortical surface, modeled as a sphere of radius 0.6, in which the top-center lies under the nasion.

# What are fallacies good for? Representational speed-up in propositional reasoning

Vladimir M. Sloutsky (sloutsky.1@osu.edu)

School of Teaching & Learning & Center for Cognitive Science, The Ohio State University  
21 Page Hall, 1810 College Road, Columbus, OH 43210, USA

## Abstract

Two experiments examine speed-up in argument pairs of various propositional forms. In the first experiment, participants were presented with pairs of conditional arguments. Some of these pairs had a form of a valid Modus Ponens (MP) inference, whereas other pairs had a form of a fallacy of Affirming the Consequent (AC). In both argument pairs, presentation of the prime led to a significant speed-up in the probe argument. In the second experiment, in addition to AC-AC and MP-MP pairs, AC-MP and MP-AC pairs were also included. Results indicated that AC primes led to a speed-up of MP probes, and MP primes led to a speed-up in AC probes. The results are discussed in relation to theories of propositional reasoning.

## Introduction

The ability to reason deriving conclusions from available information is an integral aspect of human cognition. A large component of this ability is propositional reasoning, or reasoning with logical connectives AND, OR, IF...THEN, and NOT. There are two major theoretical approaches to propositional reasoning, the syntactic approach and the semantics approach. According to the former, reasoners extract the syntactic form of the argument and apply certain formal rules of inference, or inferential schemata, to the extracted form (Braine & O'Brien, 1991; Rips, 1994). For example, reasoners easily conclude that *B* is the case, using the modus ponens (MP) schema, when presented with the following premises:

$A \rightarrow B$  (If *A* then *B*)

*A*.

The syntactic approach thus hinges on assumptions that reasoners (a) veridically represent information in the premises and (b) automatically apply inferential schemata to these representations.

According to the semantic approach, the untrained mind is not equipped with formal rules of inference. Furthermore, reasoning, to a large extent, is a function of representations of information in the premises. In turn, these representations are not veridical but are often incomplete or defective (Johnson-Laird & Byrne, 1991; Evans & Over, 1996; Sloutsky & Goladvarg, 1999; Sloutsky, Rader, & Morris, 1998).

One of the semantic theories of propositional reasoning, the Mental Model Theory (Johnson-Laird & Byrne, 1991) suggests that inferences, such as considered above, occur in the following manner. First, the reasoner constructs the initial representation of the premises:

First premise	Second premise
$A \quad B$	$A$
...	

The first line in the leftmost column makes explicit the possibility in which both *A* and *B* co-occur, and the second model (ellipses) corresponds to those possibilities in which the antecedent of the conditional is false. The theory accordingly assumes that individuals do not normally make these possibilities explicit (Johnson-Laird & Byrne, 1991). The line in the rightmost column represents the second premise. Combining the two models together leads to the inference that *B*.

There is a plethora of empirical studies contrasting predictions stemming from the two approaches. One major result of these comparisons is that the Mental Model Theory is capable of accounting for a variety of systematic errors observed in reasoning (Johnson-Laird & Savary, 1999; Newsome & Johnson-Laird, 1996; Sloutsky & Johnson-Laird, 1999; Yang & Johnson-Laird, in press; see also Johnson-Laird, 1999; Johnson-Laird & Byrne, 1991, for reviews). One of these errors accounted for by the Mental Model Theory is the fallacy of Affirming the Consequent (AC). AC has the following form:

$A \rightarrow B$

*B*.

Therefore *A*.

The inference is a fallacy because there is nothing in the argument suggesting that *B* could not occur without *A*. The mental model explanation of this fallacy is that initial representations of MP arguments and AC arguments are identical. As a result, people tend to draw conclusions, both when presented with valid MP arguments and invalid AC arguments.

This paper offers a further examination this issue. If inferences in Modus Ponens arguments occur due to the MP schema, as specified by the syntactic approach, then the use of the schema should lead to a speed-up in subsequent applications of the schema (see Smith, Langston, & Nisbett, 1992). At the same time, inferring conclusions from AC arguments should not lead to a speed-up in MP arguments, because there is no schema for AC. However, if people reason from mental representations, as according to the semantic approach, then arguments that have identical mental representations should speed up each other. We therefore, predicted that (1) AC arguments should speed-up AC arguments and (2) MP arguments should speed-up MP

arguments. We further predicted that (3) AC arguments should speed-up MP arguments and (4) MP arguments should speed-up AC arguments. The first two hypotheses were tested in Experiment 1, whereas the last two were tested in Experiment 2.

There was also a critical point added to Experiment 1. According to the syntactic theory of mental logic (Braine & O'Brien, 1998), conjunctive arguments (CONJ) of the form *A* & *B* could be simplified using conjunction elimination schema of the form:

*A* & *B*

Therefore *A*.

On the other hand, the semantic theory of mental models

suggests that conjunctions have similar (although not identical) representations as conditionals. Therefore, an important question is whether or not conjunctive arguments can also be speeded up by subsequent use. There is evidence that during text comprehension, conjunctions do not result in automatic, on-line inferences, whereas conditionals do (Gernsbacher, 1997; Millis, Golding, & Barker, 1995). The importance of this question is that, if hypotheses are confirmed, the examination of conjunctive arguments will allow us to assess the generality of findings: whether all forms that have similar representations prime each other, or if priming is limited to If...Then forms only.

Table 1: Sample stimuli by argument type and prime type.

Prime Type	Argument Type		
	MP	CONJ	AC
Related Prime (select a conclusion)	If there is an Ace then there is a Jack. There is an Ace. <ul style="list-style-type: none"> <li>• No conclusion follows</li> <li>• There is a Jack</li> <li>• There is no Jack</li> <li>• There is a Two</li> </ul>	There is an Ace and there is a Jack. There is an Ace. <ul style="list-style-type: none"> <li>• No conclusion follows</li> <li>• There is a Jack</li> <li>• There is no Jack</li> <li>• There is a Two</li> </ul>	If there is an Ace then there is a Jack. There is a Jack. <ul style="list-style-type: none"> <li>• No conclusion follows</li> <li>• There is an Ace</li> <li>• There is no Ace</li> <li>• There is a Two</li> </ul>
Unrelated Prime (select a conclusion)	There is a Three or there is a Seven, but not both. There is a Three <ul style="list-style-type: none"> <li>• No conclusion follows</li> <li>• There is a Seven</li> <li>• There is no Seven</li> <li>• There is a Jack</li> </ul>	There is a Three or there is a Seven, but not both. There is a Three <ul style="list-style-type: none"> <li>• No conclusion follows</li> <li>• There is a Seven</li> <li>• There is no Seven</li> <li>• There is a Jack</li> </ul>	There is a Three or there is a Seven, but not both. There is a Three <ul style="list-style-type: none"> <li>• No conclusion follows</li> <li>• There is a Seven</li> <li>• There is no Seven</li> <li>• There is a Jack</li> </ul>
Probe (answer Yes or No)	If there is a Queen then there is a Six. There is a Queen. <ul style="list-style-type: none"> <li>• There is a Six</li> </ul>	If there is a Queen then there is a Six. There is a Six. <ul style="list-style-type: none"> <li>• There is a Queen</li> </ul>	There is a Queen and there is a Six. There is a Queen. <ul style="list-style-type: none"> <li>• There is a Six</li> </ul>

## Experiment 1

The first goal of this experiment was to test hypotheses 1 and 2, suggesting that there is AC-AC and MP-MP priming. The second goal was to examine whether or not there is priming of conjunctive arguments.

### Method

**Participants** A total of 86 participants from Ohio State University took part in the experiment for an introductory psychology course credit. These participants represented three groups, with each group receiving a particular argument type. There were 31 participants in the Modus Ponens (MP) group, 29 participants in the Conjunction group (CONJ) and 26 participants in the Affirmation of Consequent (AC) group. All participants were fluent English speakers.

**Materials** In each of the three groups, stimuli considered of 60 critical items and 120 filler items. Critical items consisted of 30 prime-probe pairs. Half of primes had the same argument form as probes, whereas another half of the primes had a different argument form. Examples of stimuli for each of

the group are presented in Table 1. Filler items consisted of primes and probes that were similar to those in the Table, except that they had a different logical form. Primes had a form of inclusive OR (e.g., There is a Joker or an Ace, or both), whereas probes had a form of exclusive or (There is a Joker or an Ace, but not both). Participants were randomly assigned to one of the three groups.

**Design and Procedure** The experiment had a 3 Argument Type (AC, MP, CONJ) by 2 Prime Type (Related, Unrelated) mixed design with Prime Type as a repeated measure. Stimuli were presented on a PC screen and controlled by Superlab Pro for Windows (Cedrus Corporation, 1997). Participants were tested individually. Participants were told that they would read arguments on the computer. They were further told that sometimes they would need to select a conclusion from a set of conclusions, and sometimes to determine whether or not a conclusion follows logically from the premises by answering either Yes (follows) or No (does not follow). Participants were asked to respond as quickly and accurately as possible. Then they were given examples of conclusions that do and do not follow logically from premises accompanied with explanations. Finally, they were

presented with four practice trials, two of which included selecting a conclusion from a list, and another two included a Yes/No response. These practice trials were accompanied by feedback, such that participants were told whether or not their inference was warranted and why. After finishing the practice trials, participants were presented with experimental trials. Each participant received 30 experimental items (60 arguments) and 60 filler items (120 arguments) with a total of 180 arguments. Participants read each argument in self-paced fashion. Once the argument that served as a prime was answered, a probe argument appeared on the screen. Argument pairs were separated by 300 ms interstimuli intervals. The experiment took approximately 40 minutes.

## Results and Discussion

In all reported analyses, degrees of freedom are based on subjects \* item variability. Accuracy by argument and prime type are presented in Table 2 and response times are presented in Figure 1. For AC arguments accuracy was below chance  $t(389) < -7, p < .0001$ , two-tailed. For MP and CONJ arguments, accuracy was above chance  $ts(389) > 7, p < .0001$ , two-tailed. Because comparisons of response times across between-subject conditions could be misleading, we perform only comparisons across within-subject conditions. For CONJ condition, there were no significant differences in responses to the probe questions between related and unrelated primes. In fact, unrelated primes resulted in slightly (but not significantly) faster responses than related primes. At the same time, in the AC condition,  $t(353) = 3.5, p < .0001$  and MP condition,  $t(425) = 3.8, p < .0001$ , related primes resulted in a significant speed-up of responses to the probe questions. These data indicate that while there was no speed-up in the CONJ condition, both AC and MP arguments were speeded-up by more 500 ms when preceded by a related prime. Having established that responses to both AC and MP arguments could be speeded up by a related prime, we deemed it necessary to answer another question: what constitutes a related prime? According to syntactic theories of reasoning, related prime would be the one that is based on the same inference rule. However, syntactic theories do not posit a rule for AC inferences. In accordance with the semantic approach, we hypothesized that the prime is related whenever it has a identical mental representation with the probe. For example, according to the Mental Model Theory, AC and MP have identical mental representations. In this case, both an AC prime should speed up both AC and MP probes, and MP prime should speed up AC and MP probes. This prediction was tested in Experiment 2. Note, that there was not priming in CONJ-CONJ pairs; this issue will be addressed in the General Discussion section.

Table 2: Percent of accurate responses by argument type and prime type.

Argument Type	Prime Type	
	Related	Unrelated
AC	32.80	31.28
MP	95.90	92.20
CONJ	68.50	68.00

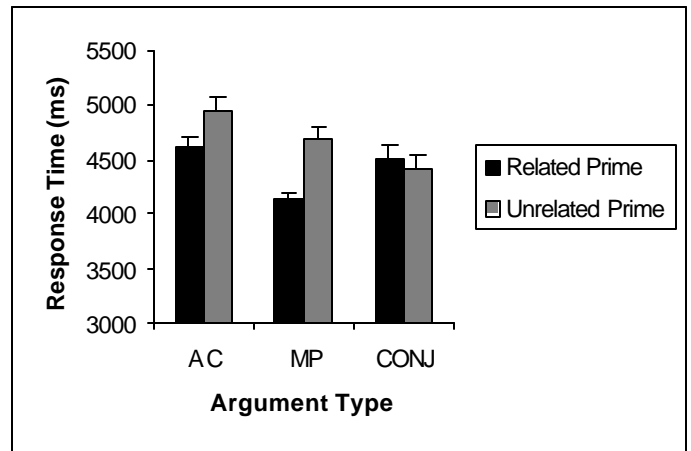


Figure 1: Response times by argument type and prime type. Error bars represent standard errors of the mean

## Experiment 2

Experiment 2 differs from Experiment 1 in two respects. First, in Experiment 2, both types of argument and types of prime varied within subjects. Each participant received two types of arguments (AC and MP) and three types of prime (AC, MP, and XOR, which was considered unrelated). Types of argument were fully crossed with types of prime. Second, because there was no CONJ-CONJ speed-up, conjunctive arguments were eliminated.

## Method

**Participants** A total of 26 participants from Ohio State University took part in the experiment for an introductory psychology course credit. All participants were fluent English speakers.

**Materials & Procedure** The experiment had a 3 Argument Type (AC, MP) by 3 Prime Type (AC, MP, XOR) within-subject design. The experimental procedure was identical to that of the first experiment, except that the total number of items in the current experiment was 240. The experiment took approximately 55 minutes.

## Results and Discussion

As in the previous experiment, degrees of freedom are based on subjects \* item variability. Accuracy rates by argument and prime type are presented in Table 3. These rates were subjected to one-sample t-tests. The analyses indicate that for AC arguments accuracy was below chance,

$t < -7$ ,  $p < .0001$ , whereas for MP argument accuracy was above chance  $ts > 7$ ,  $ps < .0001$ . Figure 2 presents relative speed-up for prime-probe pairs; estimates for relative speed-up were derived as  $RT_{\text{unrelated prime}} - RT_{\text{related prime}}$ . Speed-up rates presented in Figure 2 were subjected to one-sample t-tests. Recall that it was predicted that for both types of arguments, MP and AC primes should lead to a speed-up above XOR primes that were considered unrelated. As depicted in Figure 2, all "related" primes resulted in a speed-up. Speed-up, however, did not reach significance above 0 for AC-MP pairs, while it was significantly above 0 for the other prime-probe pairs  $ts > 2$ ,  $ps < .05$ . Speed-up effects presented in Figure 2 were also subjected to a repeated measures ANOVA. The analysis reveal no overall differences between different prime-probe pairs,  $F(3, 320) = 2.13, p = .1$ .

These findings are consistent with predictions that speed-up occurs due to a common mental representation. Note that even though speed-up in the AC-MP pair did not reach significance, the difference between XOR-MP pairs and AC-MP pairs was in the predicted direction.

Table 3: Percent of accurate responses by argument type and prime type

Prime Type	Argument Type	
	AC	MP
AC	39.74	95.64
MP	39.75	98.21
XOR	39.49	97.69

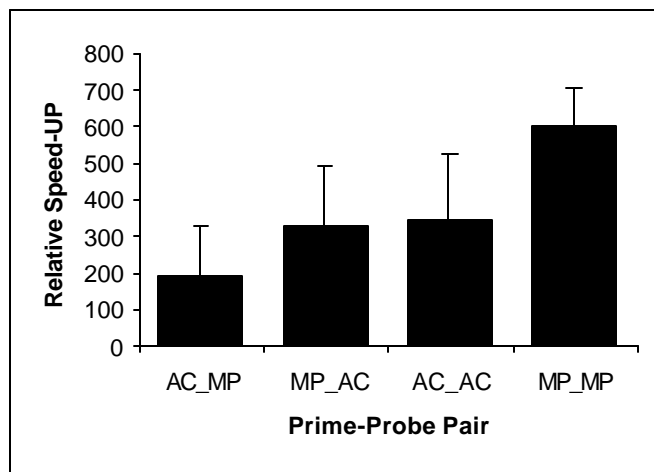


Figure 2: Relative speed-up by prime-probe pairs. Error bars represent standard errors of the mean.

## General Discussion

The results of the two reported experiments indicate that both MP and AC arguments speed-up each other. These findings support predictions that priming could be due to a common mental representation rather than due to a common syntactic rule. Indeed, what do AC and MP arguments have

in common? First, they have the common linguistic form "If...then," and second, they have a similar mental representation. It seems more likely that the observed speed-up is due to the similarity of mental representation rather than due to the similarity of linguistic form. This suggestion is based on indirect evidence (e.g., Lea, 1995; Rader & Sloutsky, 2000) that when inference in the priming argument is blocked (e.g., *If there is an Ace then there is a King. I really need an Ace.*), priming does not occur.

It also seems important that there was no speed-up in CONJ-CONJ pairs, even though these arguments have identical linguistic form and identical representation. One important difference of conjunctive arguments is that, unlike conditionals, conjunctions do not lead to an automatic, on-line inference (Gernsbacher, 1997; Millis, Golding, & Barker, 1995). Taken together, these findings suggest that the identical mental representation is not sufficient for priming; only those forms exhibit speed-up that (a) have the identical mental representation and (b) lead to an automatic inference. Furthermore, priming effects occur in both valid (MP) and invalid (AC) conditional arguments.

One finding that deviates from predictions is that in the Experiment 2, where argument forms varied within subjects, AC-MP pairs resulted in a smaller speed-up than AC-AC, MP-AC, and MP-MP pairs. Recall that in the Experiment 1, where argument forms varied across subjects, both AC-AC and MP-MP pairs resulted in a comparable (approximately 500 ms) speed-up. Taken together, results of the two experiments suggest that the presence of MP arguments may lead participants to consider AC arguments as invalid arguments (after all the participants are college undergraduates who may be familiar with basic principles of logic). This consideration did not lead to an increase in accuracy, but could have slowed down their responses.

There are several issues that are to be tested in future research. In particular, it could be predicted that strengthening of the associative link between the antecedent and the consequent in the AC argument (*If it flies then it is a bird. It is a bird.*) should result in an increase in the speed-up in AC-AC pairs. This is because when the antecedent and the consequent are highly associated, people are less likely to notice that the inference is invalid (Markovits, 1993; Markovits, Fleury, Quinn, & Venet, 1998). Alternatively, weakening of the associative link between the antecedent and the consequent in the AC argument (*If you throw a watermelon to the window, the window breaks. The window is broken.*) should result in a decrease in the speed-up in AC-AC pairs.

While these possibilities will be tested in further experiments, results of the current experiments seem to indicate that independently of the validity, MP and AC conditional arguments tend to speed-up each other. This finding seems to support the idea that those arguments that share mental representation and lead to an automatic, on-line inference are likely to get primed by each other independently of their validity.

## Acknowledgements

This research has been supported by a grant from the James S. McDonnell Foundation.

## References

- Braine, M. D. S., & O'Brien, D. P., Eds. (1998). *Mental logic*. Mahwah, NJ: Erlbaum.
- Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human Reasoning: The Psychology of deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.
- Johnson-Laird, P. N., & Savary, F. (1999). Illusory inferences: A novel class of erroneous deductions. *Cognition*, 71(3), 191-229.
- Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology*, 50, 109-135.
- Lea, R. B. (1995). On-line evidence for elaborative logical inferences in text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1469-1482.
- Markovits, H. (1993). The development of conditional reasoning: A Piagetian reformulation of mental models theory. *Merrill-Palmer Quarterly*, 39, 131-158.
- Markovits, H., Fleury, M., Quinn, S., & Venet, M. (1998). The development of conditional reasoning and the structure of semantic memory. *Child Development*, 69, 742-755.
- Millis, K. K., Golding, J. M., & Barker, G. (1995). Causal connectives increase inference generation. *Discourse Processes*, 20, 29-49.
- Newsome, M. R., & Johnson-Laird, P. N. (1996). An antidote to illusory inferences? In Cottrell, G.W. (Ed.) *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates, p. 820.
- Rader, A. W., & Sloutsky, V. M. (2000). Conditional inferences during comprehension: Is *If* really logical? Under review.
- Sloutsky, V. M., & Goldvarg, Y. (1999). Effects of externalization on representation of indeterminate problems. In M. Hahn & S. Stones (Eds.), *Proceedings of the XXI Annual Conference of the Cognitive Science Society* (pp. 695-700). Mahwah, NJ: Erlbaum.
- Sloutsky, V. M., & Johnson-Laird, P. N. (1999). Problem representations and illusions in reasoning. In M. Hahn & S. Stones (Eds.), *Proceedings of the XXI Annual Conference of the Cognitive Science Society* (pp. 701-705). Mahwah, NJ: Erlbaum.
- Sloutsky, V. M., Rader, A., & Morris, B. (1998). Increasing informativeness and reducing ambiguities: Adaptive strategies in human information processing. In Gernsbacher, M.A., & Derry, S.J. (Eds.) *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. (pp. 997-999). Mahwah, NJ: Erlbaum.
- Yang, Y., & Johnson-Laird, P. N. (in press). Illusions in quantified reasoning: How to make the impossible seem possible, and *vice versa*. *Memory & Cognition*.

# The Primacy of One-to-One Generalization in Young Children's Induction

Vladimir M. Sloutsky (sloutsky.1@osu.edu)

School of Teaching & Learning & Center for Cognitive Science, The Ohio State University  
21 Page Hall, 1810 College Road, Columbus, OH 43210, USA

Ya-Fen Lo (lo.37@osu.edu)

School of Teaching & Learning & Center for Cognitive Science, The Ohio State University  
21 Page Hall, 1810 College Road, Columbus, OH 43210, USA

## Abstract

The paper compares predictions derived from the similarity-based and the theory-based accounts of young children's induction. The former predicts the primacy of induction from one single entity to another single entity (one-to-one induction), whereas the latter does not predict such primacy. Predictions were tested in three experiments where 4-5 year-olds and 11-12 year-olds were asked to perform inductive generalization of biological properties. Participants could generalize properties either from a single animal to another single animal (one-to-one induction) or from a group of animals to a single animal (many-to-one induction). Experiments 1 and 2 revealed that under various stimuli presentation conditions, young children exhibited a strong preference of one-to-one induction, performing generalizations in a similarity-based manner. At the same time, preadolescents exhibited a strong preference of many-to-one induction, performing generalizations in a theory-based manner. In Experiment 3, an alternative explanation that one-to-one induction stems from a tendency to match quantifiers or label endings was tested and eliminated. Results are discussed in relation to cognitive and developmental aspects of inductive inference.

## Introduction

Inductive generalization is prominently present both in low-level processes, such as sensation and perception, and in high-level processes, such as learning and transfer, categorization, analogy, rule discovery, and inductive inference (see Shepard, 1987, for a discussion). Inductive generalization involves at least two stimuli (or stimuli sets): the source and the target of generalization.

One issue that has been hotly debated is what aspects of the source and the target support inductive inference. One possibility that has been extensively discussed in the literature is that inductive generalizations are driven by similarity construed as featural overlap between the source and the target (see Estes, 1994; Medin, 1975; Nosofsky, 1986; Shepard, 1987; Tversky, 1977, for specific models of computing similarity). In this case, the more similar the source and the target, the more likely there will be generalization from the source to the target (see Medin & Smith, 1984; E. Smith, 1995; L. Smith, 1989 for discussions).

However, it has been counter argued that similarity construed this way does not sufficiently constrain

generalization processes (see Carey, 1985; Keil, 1989; 1994; Medin et al., 1993, for discussions). For example, there are many more overlapping features between a live monkey and a mechanical monkey than between the live monkey and a worm. However, people deem it more appropriate to generalize biological properties from a live monkey to a worm than to generalize biological properties from the live monkey to a mechanical monkey (Carey, 1985). Therefore, not all featural overlaps are equally important. Somehow people intuitively realize that it is appropriate to generalize certain biological properties from Elephant to Hippopotamus (as they both are mammals) and it is inappropriate to generalize these properties from Elephant to Paris (as they both are smaller than China). Hence, it has been argued that generalization must be constrained by some deep "theoretical" beliefs that could not be reduced to simple featural similarity. Proponents of this view have suggested that generalization processes are constrained by a set of core beliefs about the "essence" of a category. Those entities that have common "essential" properties (e.g., the same biological origins) should be also considered as members of a common group. Those biological properties that stem from the essence (and therefore from the common membership) could be legitimately generalized from one entity onto the whole group and subsequently to each member of this group (Murphy & Medin, 1985; Gelman & Coley, 1991; Gelman & Wellman, 1991; Keil, 1994).

In this paper we attempt to derive predictions from these positions and to empirically test these predictions. Inductive generalizations can be performed over individual entities (e.g., This dog has property X, therefore that dog has property X) or over classes (e.g., Dogs have property X, therefore cats have property X). Quantification of the source and the target define several types of induction. The current research focuses on two of these types of induction over individual entities, one-to-one induction and many-to-one induction. In the case of one-to-one induction, attributes or relations could be generalized from one single entity to another single entity (e.g., This sparrow has biological property X, therefore that sparrow has biological property X). In the case of many-to-one induction, attributes or relations could be generalized from a group of entities to a single entity (e.g., Sparrows have biological property X, therefore this sparrow has biological property X). Note that induction could be also performed strictly over classes (see

Osherson, Smith, Wilkie, Lopez, & Shafir, 1990, for a discussion of induction over classes).

The distinction between the two types of induction affords deriving specific and testable predictions from each of the above mentioned positions. If induction in young children is similarity-based, there should be primacy of one-to-one induction over many-to-one induction, whereas if induction is category-based, there should not be such primacy. As shown below, each prediction follows directly from the respective position.

Proponents of the similarity-based position have argued that induction in young children is not category-based, and that both induction and categorization are products of featural similarity between compared stimuli (Sloutsky & Lo, 1999). They have also suggested that (a) different attributes and attribute dimensions have different weights in the computation of similarity, (b) young children consider linguistic labels as attributes with greater weights than other attributes (Sloutsky & Lo, 1999). Finally, according to the similarity-based approach, when entities are novel, computation of similarity between two single novel entities should be simpler than computation of similarity between many novel entities and one novel entity. This is because it is possible to directly compute similarity between single entities, whereas computation of similarity between a group and a single entity is difficult. The latter requires one first to construe a composite representation of the group and then to compute similarity between the single entity and the group. Note that the argument may not apply to familiar entities, for which a composite representation had been established (see Estes, 1994, for a discussion). Therefore, if induction is a function of overall similarity, one-to-one induction should be easier for young children than many-to-one induction.

Recall that according to the theory-based approach, young children have abstract representations of categories, such as biological kinds. When an entity is familiar, it is represented as a member of a familiar category, whereas when an entity is novel, it is represented as a member of a novel category. These novel categories are devoid of representational specifics; they rather exist as category "templates" or "placeholders" (see Gelman, Coley, & Gottfried, 1994; Gelman & Coley, 1991 for discussions), and linguistic labels point to this category placeholder. When a perceptual input indicates that compared novel entities are animals, a set of beliefs about "natural kinds" is activated. These include beliefs about growth, inheritance, reproduction, self-generated movement, and so forth (Gelman, Coley, & Gottfried, 1994). These beliefs in conjunction with the common category membership suggest that both entities belong to the same natural kind, and, therefore, they should share unobservable biological properties (Murphy & Medin, 1985; Gelman & Coley, 1991; Keil, 1994). Thus, according to the theory-based explanation, induction is category-based (i.e., it is a function of categorization) (Gelman & Coley, 1991), and the process underlying induction should be as follows. (1) The description of a single entity or multiple entities (e.g., This Gubla has biological property X or These Gublas have

biological property X) activates the essence placeholder "GUBLA." (2) Other members of the category GUBLA (this membership is denoted by the linguistic label) should have biological property X. (3) As indicated by the common label, this GUBLA is a member of the category GUBLA (or these Gublas are members of the category GUBLA), and therefore, it (or they) should have biological property X. Therefore, because both one-to-one and many-to-one induction follows from the category membership, there should be no primacy of one-to-one induction over many-to-one induction.

To test predictions derived from both approaches, we developed the following task. Suppose that the child is presented with a set of realistically looking novel animals having novel labels (e.g., "Look, these are Gublas"). Then, one Gubla is presented as a Target, another Gubla is presented as Test 1 and the rest of Gublas are presented as Test 2. The child is also told that this Gubla (Test 1) has biological property X, whereas these Gublas (Test 2) have biological property Y. Does the Target Gubla have biological property X or Y?

The putative processes that, according to each model, underlie the child's inference are as follows. According to the theory-based approach, the encounter with a group of novel biological objects that have the same linguistic label (i.e. Gubla) should activate the category placeholder GUBLA. Once the category is activated, the child should be equally likely to generalize from Test 1 (one Gubla) or from Test 2 (many Gublas) to the Target. On the basis of the theory-based approach, it should be inferred that in the task like this, young children should be at chance, or have a slight preference for many-to-one over one-to-one induction. The slight preference might stem from the fact that many identically looking Gublas should be more representative of the category than a single Gubla. Furthermore, normatively it is more appropriate to generalize from many Gublas than from a single Gubla, because a single entity is more likely to be an exception than many entities. Of course, we should not expect many young children to take into account this consideration, therefore, if any, only a small many-to-one preference should be predicted.

The similarity-based approach yields different predictions. As described above, all other things being equal, the computation of similarity between two entities should be simpler than computation of similarity between many entities and one entity. In addition, because similarity between two identical entities is the unity (Estes, 1994; Medin, 1975; Sloutsky & Lo, 1999), this similarity could not be less than similarity between several entities and one entity. Therefore, similarity between the Test Gubla and the Target Gubla should be no less than similarity between Test Gublas and the Target Gubla, and the former should be more easily computed. Based on these considerations, the similarity-based approach predicts a large preference of one-to-one induction over many-to-one induction.

These considerations led us to formulation of the following specific predictions. If young children base their induction on similarity between compared entities, they should generalize from a single Gubla to another single Gubla more often than chance. At the same time, according



to the theory-based account, young children should perform at chance (or with a slight preference of many-to-one induction).

## Experiment 1

### Method

**Participants** Participants were 31 children aged 4 to 12 years. The first group consisted of 16 four-to-five-year-old children enrolled in two daycare centers in an upper middle class suburb of Columbus, Ohio ( $M = 4.5$  years,  $SD = 0.6$  years, 11 boys and 5 girls). The second group consisted of 15 eleven-to-twelve year-olds selected from a public middle school located in an upper middle class suburb of Columbus, Ohio ( $M = 11.7$  years,  $SD = .31$  years, 8 girls and 7 boys).

**Materials** Eight sets of line-drawing pictures were used in the present experiment. Each set consisted of two single pictures and a stack of pictures (see Figure 1), with each picture measuring approximately 3" by 5". Both single pictures depicted realistically looking animals, whereas the stack was turned faced down such that pictures in the stack were not visible. The Target (a single picture) looked identical to Test 1. Materials also included artificial labels and a set of biological properties. The animals presented in each set of pictures were given the same artificial label (e.g., a Gubla). Children were taught that each of the Test stimuli had a particular biological property (e.g., has salt inside the body or has sugar inside the body). The task consisted of generalization of biological properties from one of the Test stimuli to the Target. The experiment had a between-subject design with age as a factor. The dependent variable of interest was the proportion of inductive generalizations from each of the Test stimuli, either one-to-one induction (choosing Test 1) or many-to-one induction (choosing Test 2). Each participant received eight trials.

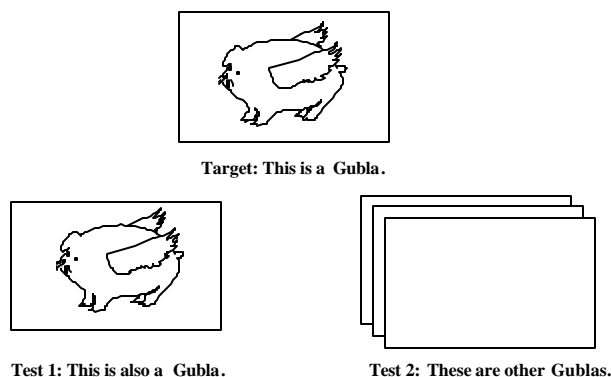


Figure 1: Layout of stimuli in Experiment 1.

**Design and Procedure** The experiment was conducted in a single 15-20 minute session that included three phases: stimuli presentation, comprehension/memory check, and inductive inference. Each participant was tested individually in a separate room at their daycare center or school.

**Stimuli presentation.** Each participant was presented with eight stimuli triads, one triad at a time. Each triad was

referred to using a two-syllable artificial linguistic label and was introduced as a group of animals (e.g., I will show you several Famos). The experimenter then presented participants with three stimulus items: (a) a single card depicting a single animal (the Target), (b) another single card depicting another single animal (Test stimulus 1), and (c) a stack of cards that were face down (Test stimulus 2). At this point, participants were asked to repeat the label. After presenting the stimuli items, the experimenter introduced two biological properties, one characterizing Test 1, and another characterizing Test 2 (e.g., This Famo has a lot of sugar inside the body. These Famos have a lot of salt inside their bodies). The order of presentation of the Test stimuli, their positions relative to the Target, and the order of introduction of biological properties were counterbalanced across trials. Stimuli items were randomly paired with biological properties.

**Comprehension/memory check and inductive inference phases.** After the stimuli items were presented, participants were asked to repeat the labels and biological properties. The labels and biological properties were reintroduced when participants failed to answer correctly. All participants successfully completed this comprehension/memory check phase. After repeating the labels and biological properties, children moved to the inductive inference phase, in which participants were asked which of the two biological properties was likely to be shared by the Target.

### Results and Discussion

In this section, we present proportions of generalizing from each of the Test stimuli across the two age groups. Results of this experiment are presented in Figure 2.

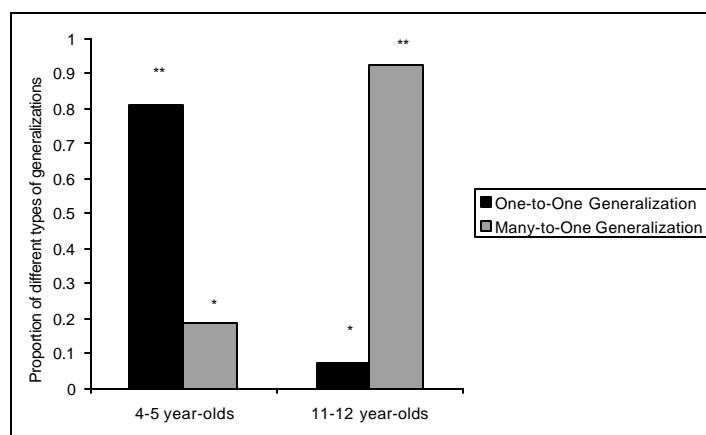


Figure 2: Proportions of one-to-one and many-to-one generalizations by age. Note: \*\* above chance,  $p < .0001$ ; \* below chance,  $p < .0001$ .

To determine the difference from chance, these results were subjected to one-sample t-tests. The analysis indicates that while 4-5 year-olds generalized from Test 1 (one animal) to the Target significantly above chance (81% of all responses),  $t(15) = 11.2$ ,  $p < .0001$ , 11-12 year olds generalized from Test 2 (many animals) to the Target,

significantly above chance (93% of all responses),  $t(14) = 9.4$ ,  $p < .0001$ . Percentages of one-one choices and percentages of many-one choices (both aggregated across the 8 trials) were subjected to a one-way ANOVA with age as a factor. The analyses indicate that 4-5 year-olds were significantly more likely to generalize from Test 1 (one animal) to the Target than 11-12 year-olds, whereas 11-12 year-olds were significantly more likely to generalize from Test 2 (many animals) to the Target than 4-5 year-olds,  $F_s(1, 29) > 97.1$ ,  $p_s < .0001$ . In addition, 10 out of 14 preadolescents explicitly pointed that the Target is more likely to share properties with a larger group of animals than with a single animal.

These results support our predictions describing inductive generalizations of young children and point to important differences in inductive generalizations of 4-5 year-olds and 11-12 year olds. While preadolescents' inductive generalizations conform to what should be expected when induction is category-based (they did in fact generalize in a category-based manner, thus both supporting predictions and validating the task), young children's inductive generalizations conform to what should be expected when induction is similarity-based. This experiment, however, constitutes a rigorous test of whether participants performed category-based induction, and a not so rigorous test of whether participants performed similarity-based induction. This is because one-to-one induction was supported by a picture, whereas many-to-one was not (see Figure 1).

Therefore, reported findings are indicative of the category-based induction of 11-12 year-olds, whereas they are ambiguous with respect to induction of 4-5 year-olds. Indeed, generalization in the latter group could point either to the primacy of one-to-one induction or to the preference of young children of depicted stimuli over non-depicted stimuli. Although such preference in itself might be indicative of similarity-based induction (indeed, category placeholders are not accompanied by pictures), we deemed it necessary to conduct a more rigorous testing of predictions generated by the similarity-based model. To this end, we conducted Experiment 2, where both Test 1 and Test 2 were both accompanied by pictures (Condition 1) or both were presented without pictures (Condition 2).

## Experiment 2

### Method

**Participants** A group of 30 children aged 4 to 5 years participated in the two conditions. These children were selected from daycare centers in an upper middle class suburb of Columbus, Ohio on the basis of permission slips returned by parents. The No-Picture condition group of 15 children consisted of 7 boys and 8 girls ( $M = 4.4$  years,  $SD = 0.48$  years). The Picture condition group of 15 children consisted of 9 boys and 6 girls ( $M = 4.4$  years,  $SD = 0.39$  years).

**Materials, design, and procedure** Materials, design, and procedure were identical to those in Experiment 1. The only

differences were that the design included an additional between-subject factor, the picture presentation condition that had two levels, Picture and No-Picture conditions, and that pictures were presented differently from those in Experiment 1. In the Picture condition, both Test stimuli were accompanied by pictures, whereas in the No-Picture condition neither of the Test stimuli was accompanied by a picture.

## Results and Discussion

Results of this experiment indicate that in both Picture and No-Picture conditions young children reliably generalized in a one-to-one manner. In the No-Picture condition in 78% of responses children generalized from Test 1, whereas in the Picture condition 77% of responses children generalized from Test 1, both above chance,  $t_s(14) > 6.2$ ,  $p_s < .0001$ . The response patterns in the Picture and No-Picture conditions were practically identical,  $t < 0.5$ .

These findings replicate those of Experiment 1 for young children, ruling out the possibility that young children's responses in Experiment stemmed from the fact that Test 1 (single animal) was accompanied by a picture, whereas Test 2 (many animals) was presented without a picture. Results of Experiments 1 and 2 also point to a difference in inductive generalization of young children and preadolescents: while the later perform inductive generalizations in a manner compatible with the category-based model, the former perform in a manner compatible with the similarity-based model.

The fact that young children equally frequently generalized from a single animal in both Picture and No-Picture conditions deserves special consideration. This finding could be indicative of several factors. First, it is possible that young children generalize from Test 1 (single animal) rather than from Test 2 (many animals) because they merely match quantifiers (e.g., one and one vs. one and many) or linguistic labels (e.g., Gubla and Gubla vs. Gubla and Gublas). Another possibility is that because computation of similarity is easier between single objects, young children are biased to compute similarity between single objects prior to computing similarity between a single object and multiple objects. Experiment 3 was conducted to distinguish between these possibilities.

## Experiment 3

### Method

**Participants** A group of 16 children aged 4 to 5 years ( $M = 4.1$  years,  $SD = 0.4$  years, 10 girls and 6 boys) participated in this experiment. These children were selected from daycare centers in an upper middle class suburb of Columbus, Ohio on the basis of permission slips returned by parents.

**Materials, design, and procedure** Materials, design, and procedure were identical to those in Experiment 1. The only difference was that the Target was presented as many entities, Test 1 as a single entity, and Test 2 as many entities. All stimuli were presented face up.

## Results and Discussion

Results of this experiment indicate that in 52% of responses young children generalized from Test 2, whereas in 48% of responses they generalized from Test 1; both types of generalization were indistinguishable from chance,  $t(15) < 0.3$ . Furthermore, the analysis of patterns of individual responses indicates that the chance-level performance does not stem from a bi-modal distribution where a part of the sample consistently generalized from Test 1, whereas the other part consistently generalized from Test 2. This performance rather stemmed from inconsistency within-participants. In particular, 2 out of 16 participants consistently (on 6 or more out of 8 trials) generalized from Test 2 (many entities), and another 2 out of 16 participants consistently (on 6 or more out of 8 trials) generalized from Test 1 (single entity), while 12 out of 16 participants were inconsistent in their choices of the two test items.

These findings allow us to rule out the matching hypothesis. Indeed, if participants were exhibiting matching, they should have generalized from Test 2 (many entities) to the Target (many entities) most often, which was not the case. Therefore, it seems plausible that patterns of responses observed in Experiments 1 and 2 (i.e., the tendency to generalize from a single entity to another single entity) stem from the fact that it is easier to compute similarity between several single entities than it is to compute similarity between a group of entities and a single entity.

## General Discussion

Results of the three reported experiments are as follows. Young children more readily generalize biological properties from one single entity to another single entity, whereas older children more readily generalize biological properties from many entities to a single entity. At the same time young children performed at chance when asked to generalize from many entities either to a single entity or to many entities. The latter finding undermines the possibility that young children's preference for generalization from one single entity to another single entity stems from their tendency to match quantifiers or label endings.

Taken together, these findings support predictions of the similarity-based approach. In particular, they indicate that when the computation of similarity is relatively simple (such as computation of similarity between single entities) young children more readily generalize biological properties than when computation of similarity is relatively complex (such as computation of similarity of many entities to a single entity). At the same time, when computation of similarity is comparably difficult (such as computation of similarity of many entities to many entities or a single entity to many entities), young children perform at chance.

These results point to the primacy of the one-to-one induction over the many-to-one induction in young children, an effect that has been predicted by the similarity-based position, but not by the theory-based position. Recall that in the case of category-based induction advocated by the theory-based position, there should be no primacy of the

one-to-one induction, and preadolescents, who supposedly perform induction in a category-based manner, did not exhibit the primacy of one-to-one induction.

Results also point to important developmental differences between young children and preadolescents. While preadolescents' inductive generalizations conform to what should be expected when induction is category-based, young children's inductive generalizations conform to what should be expected when induction is similarity-based. Therefore, it seems reasonable to infer that there should be a developmental transition from similarity-based to category-based induction. Such transition could be due to developmental and educational factors that lead to understanding that common category membership is a better predictor of unobservable properties than similarity (e.g., a whale looks more similar to a fish, but has internal structure similar to other mammals). However, additional research is needed to discern and tease apart these contributing factors.

The reported findings afford the differentiation between the theory-based and the similarity-based approaches to young children's induction, undermining the former and supporting the latter. Recall that according to the theory-based position, linguistic labels activate "essence placeholders" that should be equally applicable to all members of the category, independent of the quantity of these members. Therefore, if induction had been performed in a category-based manner, young children should have equally often generalized from a single animal and from a group of animals, or have a slight preference for many-to-one over one-to-one induction. In addition, the theory-based position does not predict dramatic differences between young children and preadolescents: both groups should perform induction in the category-based manner. At the same time, the similarity-based position (e.g., Sloutsky & Lo, 1999) predicts that while young children should generalize in a similarity-based manner (generalizing from a single entity to another single entity), preadolescents should generalize in a category-based manner. The primacy of one-to-one induction in young children and major differences between young children's and preadolescents' induction fit predictions of the similarity-based position, while not fitting predictions of the theory-based position.

Of course, the current results could not conclusively rule out the possibility of young children having representations of category templates, and it is hard to imagine any empirical findings capable of conclusively ruling out this possibility. The results of current experiments, however, support a parsimonious account of young children's induction that is based on a set of a priori predictions. We believe that a priori predictions are favored over post hoc accounts by both inferential statistics and philosophy of science, and, therefore, they should weigh more than post hoc accounts (cf. Barsalou, 1999).

In short, while the similarity-based approach is not capable of conclusively ruling out the proposal that young children rely on categories when performing inductive inference, it is capable of undermining such a possibility. In particular, the similarity-based approach is capable of predicting phenomena (such as those reported above) that

could not be predicted by the category-based position.

### Acknowledgements

This research has been supported by a grant from the James S. McDonnell Foundation to the first author.

### References

- Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4), 577-660.
- Carey, S. (1985). Are children fundamentally different kinds of thinkers and learners than adults? In S. Chipman, J. Segal & R. Glaser (Eds.), *Thinking and learning skills (Volume 2)*. (pp. 485-518). Hillsdale, NJ: Lawrence Erlbaum.
- Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.
- Gelman, S. A., & Coley, J. (1991). Language and categorization: The acquisition of natural kind terms. In S. A. Gelman, S. & J. P. Byrnes (Eds.), *Perspectives on language and thought: Interrelations in development* (146-196). New York, NY: Cambridge University Press.
- Gelman, S. A., & Markman, E. (1986). Categories and induction in young children. *Cognition*, 23, 183-209.
- Gelman, S. A., & Wellman, H. M. (1991). Insides and essence: Early understandings of the non-obvious. *Cognition*, 38(3), 213-244.
- Gelman, S. A., Coley, J. D. & Gottfried, G. M. (1994). Essentialist beliefs in children: The acquisition of concepts and theories. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 234-254). New York, NY: Cambridge University Press.
- Keil, F. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 234-254). New York, NY: Cambridge University Press.
- Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.
- Medin, D. (1975). A theory of context in discrimination learning. In G. Bower (Ed.), *The psychology of learning and motivation* (pp. 263-314), Vol. 9. New York: Academic Press.
- Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology*, 35, 113-138.
- Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100(2), 254-278.
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289-316.
- Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39-57.
- Osherson, D. N., Smith, E. E., Wilkie, O, Lopez, A, & Shafir, E. (1990). Category-based induction. *Psychological Review*, 97, 185-200.
- Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, 237, 1317-1323.
- Sloutsky, V. M., & Lo, Y.-F. (1999). How much does a shared name make things similar? Part 1: Linguistic labels and the development of similarity judgement. *Developmental Psychology*, 6, 1478-1492.
- Smith, E. E. (1995). Concepts and categorization. In E. E. Smith & D. Osherson (Eds), *An invitation to cognitive science*, Vol. 3: Thinking (2nd ed.), (pp. 3-33). Cambridge, MA: MIT Press.
- Smith, L. (1989). A model of perceptual classification in children and adults. *Psychological Review*, 96(1), 125-144.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327-352.

# Simulating Conditional Reasoning Containing Negations: A Computer Model and Human Data

Jacques Sougné (J.Sougne@ulg.ac.be)  
University of Liège  
Department of Psychology  
Bât B32, Sart Tilman  
4000 Liège Belgium

## Abstract

Modeling human conditional reasoning of the type “if p then q” containing negations poses a challenge for connectionism. A network of spiking neurons (INFERNET) was used to model this type of conditional reasoning. This model also provides insights on certain human limitations. The model is compared to empirical data, and classical explanations. Statistical analysis shows that the model’s performance not only surpasses classical explanations but also provides a very good overall fit to empirical data. INFERNET simulator results are also compared to human performance. The simulations compare well with both human performance and limitations.

## Introduction

INFERNET (Sougné, 1996, 1998a, 1998b, Sougné & French, 1997) achieves variable binding through temporal synchrony of node firing. In short, when one node fires in synchrony with another, they are temporarily bound together. It has a limited Working Memory (WM) span and the content of WM is maintained by oscillations. Once a node is activated, it tends to fire rhythmically at a particular frequency. This technique is used to represent n-ary predicates (Sougné, 1996), relational reasoning with multiple instantiation (Sougné, 1998a; Sougné, 1998b), working memory (Sougné & French, 1997) and conditional reasoning (Sougné, 1996). This paper shows how the model handles negated conditionals.

Many psychological studies in the area of deductive reasoning have focused on conditional reasoning of the type “if p then q.” Of course, some logicians would deny that material implication is really what humans mean by “if...then”. Nonetheless, here are transcribed rules related to material implication: modus ponens (MP) *If p then q; p; infer q* and modus tollens (MT) *If p then q; ~q; infer ~p* (~ stands for not). While most humans follow modus ponens, it is different for modus tollens. People also use two inappropriate rules related to material equivalence: Denial of the antecedent (DA) *If p then q; ~p; infer ~q*, and Affirmation of the consequent (AC) *If p then q; q; infer p*. Throughout this paper the “if p then q” form will be called the “major premise”, p the antecedent, q the consequent.

What happens when negations are introduced into the major premise? Negation can affect the antecedent or the consequent. It produces four forms of major premises.

Table 1 shows these four forms and the inferences resulting from the application of the four rules (MT, DA, AC, MT).

Table 1: Combination between form of major premises and the result of applying the four inference rules

	MP		DA		AC		MT	
	given	infer	given	infer	given	infer	given	infer
<i>If p then q</i>	p	q	not p	not q	q	p	not q	not p
<i>If p then not q</i>	p	not q	not p	q	not q	p	q	not p
<i>If not p then q</i>	not p	q	p	not q	q	not p	not q	p
<i>If not p then not q</i>	not p	not q	p	q	not q	not p	q	p

Empirical studies reveal that negations do modify the frequencies of rule application (Evans, 1977, Wildman & Fletcher, 1977, Pollard & Evans, 1980). Pollard & Evans (1980) explain these data with what they call “negative conclusion bias” which is a tendency to prefer accepting a conclusion in the negative form. This is effectively the case for DA and MT. This is not the case for MP, but one could invoke a ceiling effect. Finally the effect is not clear for AC. As stated by Evans, Newstead & Byrne (1993), this bias could be explained by people’s caution. Concluding that “the letter is not an X” would have a higher probability (25/26) than concluding that “the letter is an X” (1/26). Oaksford & Chater (1994) provide a similar explanation. There is also an interpretation of negation effect in terms of a “Matching bias”: a tendency to verify cases that are stated in the major premise. However, this bias concerns only certain procedures like the “Wason Selection Task”, the “Truth Table Task” or the “Evans construction task” in which participants have to test or verify a major premise instead of applying it. Moreover, matching bias is closely related to implicit negation (Evans, 1998). The present study focuses on explicit negation. While negation in conditionals is known to create difficulties (Oaksford & Stenning, 1992), little is said about double negation (for an exception, see Sperber, Cara, & Girotto, 1995 or Evans, Clibbens & Rood, 1995).

In this paper, the INFERNET simulator’s performance will be compared with human data. INFERNET suggests hypothesis related to the difficulty of removing double negations. An experiment was also done in order to collect reaction time data in a production task which were not available in previous studies.

## INFERNET

INFERNET is a network of spiking neurons (Maass & Bishop, 1999). In INFERNET, nodes can be in two

different states: they can fire (be on), or they can be at rest (be off). A node fires at a precise moment and transmits activation to other connected nodes with some time course. When a node activation or potential  $v_i^{(t)}$  reaches a threshold, it emits a spike. After firing, the potential is reset to some resting value  $V_r$ . Inputs increase the node potential, but some part of the node potential is lost at each time step. Spiking neuron models use a post synaptic potential function. Integration of input in INFERNET is a variation of standard input integration. In INFERNET there are two main types of connections: either they act on nodes (synaptic link) or on synapses (presynaptic link). Unlike most links, these latter links act on *connections* rather than nodes (French, 1995). Moreover each of these connections can be excitatory or inhibitory. There are six types of connections: synaptic excitation, synaptic inhibition, presynaptic amplification of an excitation, presynaptic inhibition of an excitation, presynaptic inhibition of an inhibition and presynaptic amplification of an inhibition (figure 1). In addition to the weight of a connection, there is a delay parameter associated with each connection. A delay of 10 means that the effect of the presynaptic node firing on the postsynaptic node will take 10 units of time. A unit of time has been taken to simulate 1 ms. In addition, connection weights are modified by a random factor that injects white noise into the signal propagation.

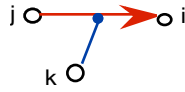


Figure 1: Example of synaptic and presynaptic connection in INFERNET. The node  $k$  inhibits the excitatory connection from  $j$  to  $i$

The potential of node  $i$  at time  $t$ ,  $v_i^{(t)}$  is:

$$V_i^{(t)} = \sum_{j \in \Gamma_i} \sum_{t_j^{(f)} \in F_j} \left[ w_{ij} + \sum_{k \in \mathcal{K}_{ij}} \sum_{t_k^{(f)} \in F_k} w_{k \rightarrow ij} \varepsilon_{k \rightarrow ij}(x) \right] \varepsilon_{ij}(x) - \eta_i(u) \quad (1)$$

The potential of node  $i$ :  $v_i^{(t)}$  is affected by connection weights coming from presynaptic node  $j$ :  $w_{ij}$  but also by the connection weights that modify this connection  $w_{k \rightarrow ij}$ . The set of presynaptic to node  $i$  is  $\Gamma_i = \{j | j \text{ is presynaptic to } i\}$ .  $F_j$  is the set of all firing times of presynaptic nodes  $j$ :  $t_j^{(f)}$ . The set of presynaptic to synapse  $ij$  is  $\mathcal{K}_{ij} = \{k | k \text{ is presynaptic to } ij \text{ synapse}\}$ .  $F_k$  is the set of all firing times of  $k$  nodes:  $t_k^{(f)}$ . These are the nodes from which start a connection acting on the connection  $ij$ . The connection weight linking node  $k$  to synapse  $ij$  is designed by  $w_{k \rightarrow ij}$ . The equations  $\varepsilon_{ij}(x)$  and  $\varepsilon_{k \rightarrow ij}(x)$  express the postsynaptic potential function. A value  $\eta_i(u)$  associated with the refractory state of nodes is subtracted. When  $v_i^{(t)}$  reaches the threshold  $\theta$ , node  $i$  fires and  $V_i$  is reset to a resting value  $V_r$ ,

## Representation in INFERNET

How does the brain represent the world? Two contrasting hypotheses are often presented in neuroscience: the code used by neurons is either a rate code or a pulse code. INFERNET relies on a pulse code, specifically, phase and

synchrony. In INFERNET, a symbol is represented by a cluster of nodes and is activated if its nodes fire in synchrony (the firing distribution is tightly concentrated around the mean: figure 2). Different symbols share nodes, so representations are distributed (see Sougné, 1998b), or more accurately, semi-distributed.

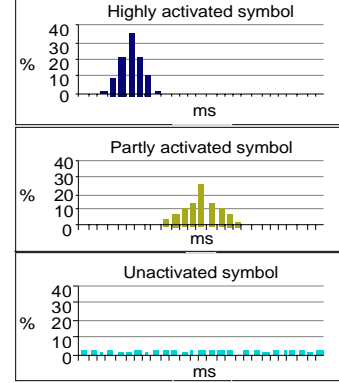


Figure 2: Symbols are represented as a set of nodes firing in synchrony.

There is considerable neurobiological evidence for considering synchrony as a possible binding mechanism in the brain (Roelfsema, Engel, König & Singer, 1996, Singer, 1993, Singer & Gray, 1995). In INFERNET, attributes are bound to an object and objects are bound to their roles by synchronous firing. For example, to represent “the red rose on the green lawn”, the attribute “red” must fire in synchrony with the object “rose” and they must fire synchronously with nodes belonging to the role “supported object” (Figure 3).

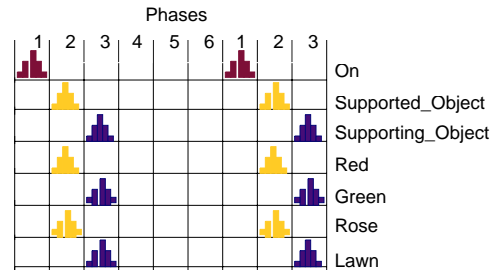


Figure 3: The “red rose on the green lawn” requires binding of symbols with their roles.

Discrimination is achieved by successive synchronies, for example, to discriminate a red rose on a green lawn. The nodes belonging to “red”, “rose” and “supported object” must fire in synchrony and those corresponding to “green”, “lawn” and “supporting object” must also fire in synchrony. Further, these two sets of nodes must fire asynchronously in different phases for “the red rose on the green lawn” to be perceived. Engel, Kreiter, König, & Singer (1991) provide evidence to show that if several objects are present in a scene, several groups of cells are clustered in distinct windows of synchrony.

A number of neurobiological parameters are involved in representations that rely on clusters of nodes firing simultaneously. The first is the frequency of oscillation. Certain specific oscillatory activities seem to facilitate synchronization (Roelfsema et al., 1996, Singer, 1993). In INFERNET once a node is activated, it tends to begin oscillating at a  $\gamma$  frequency range, whose lower limit is 30Hz and upper limit varies, according to various authors, from 70Hz (Abeles, Prut, Bergman, Vaadia & Aertsen, 1993), 80 Hz (MacKay, 1997) to 100 Hz (Wilson & Shepherd, 1995). The temporal gap between 2 spikes of a node is therefore from 10 to 33 ms. These  $\gamma$  waves have been observed to be associated with attention (Wang & Rinzel, 1995) and with associative memory (Wilson & Shepherd, 1995) and therefore seem to be a primary candidate for enabling synchronization and binding (Singer, 1993). The second key parameter is the precision of the synchrony at this frequency range. According to Singer and Gray (1995) this precision is between 4 to 6 ms., while for Abeles and al. (1993), it is about 5 ms, sometimes less, and depends on the oscillation frequency. This allows us to approximate the number of windows of synchrony that can be differentiated, i.e., approximately  $25/5 = 5$ , based on a typical frequency of 40Hz. If we assume that a window of synchrony corresponds to an item, a word, an idea, an object in a scene, or a chunk in working memory (WM), this puts WM span at approximately 5, with a small amount of variance since precision is proportional to oscillation frequency. This corresponds to current estimates of human WM span (see Cowan, 1998). The more the system needs to discriminate objects in WM, the more precise the synchrony should be. Since this parameter is bounded, it can lead to WM overload where windows of synchrony can no longer be distinguished. Therefore, the number of distinct items and the number of predicate arguments (Sougné, 1996) in WM is limited. Finally, following Lisman and Idiart (1995), the representation is maintained in WM by bursts of  $\gamma$  waves. Similar explanations for the brain's ability to store short-term memory items can be found in the literature (Hummel & Holyoak, 1997; Jensen & Lisman, 1998; Lisman and Idiart 1995; Shastri & Ajjanagadde, 1993).

### Inference in INFERNET

INFERNET implements logical gates sensitive to input timing. AND-gates require all inputs to reach the target at the same time. This is achieved by a set of excitatory and inhibitory links combined with presynaptic inhibition and facilitation (see Hawkins, Kandel, and Siegelbaum, 1993, for neurobiological evidence of this mechanism). Similarly, XOR-gates are only on when one of the inputs is active. These gates are related to the phenomenon of *coincidence detection* (Konnerth, Tsien, Mikoshiba, & Altman, 1996, Singer, Engel, Kreiter, Munk, Neuenschwander, & Roelfsema, 1997).

INFERNET has a Long Term Knowledge Base that is used for encoding premises and answering queries. Figure 4 shows the knowledge necessary to make conditional inferences with negations. Arrows represent connections; they are tagged with numbers that indicate the time required

to propagate activation. Specifically, in this example, a delay of 30ms corresponds to the lag between two spikes of a node oscillating at 33Hz. This delay ensures that these symbol-node spikes will synchronize after 30ms. The knowledge encoded, as shown in Figure 4, can correctly answer queries related to material implication.

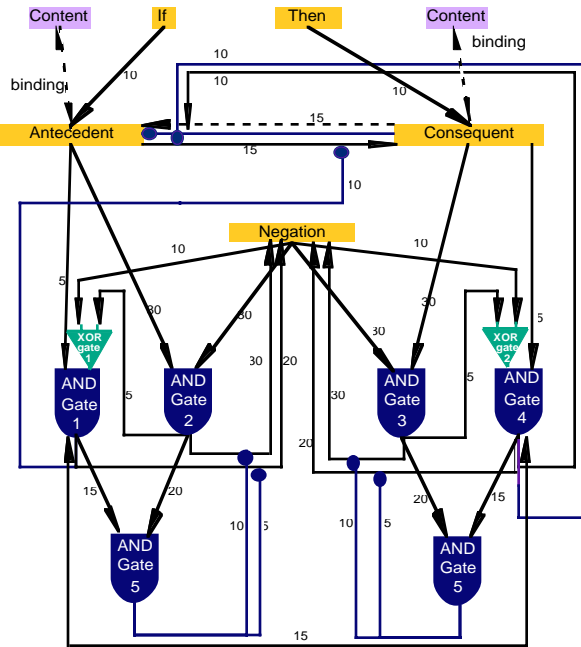


Figure 4: The encoded knowledge necessary to deal with negated conditionals

The first capacity that INFERNET must have is the ability to distinguish negations in the major premise. *AND-gate 2* detects when the antecedent is negated in the major premise and *AND-gate 3* detects a negated consequent. During the premise-encoding phase, if an antecedent is negated, for example: *If ~p then q*, the connection between the *AND-gate 2* and *p* will be strengthened as well as connections between *p* and *Antecedent*. After this phase, the firing of *p* nodes will be sufficient to induce the synchronous firing of nodes of *AND-gate 2*. The second ability of INFERNET is to detect whether in the question (minor premise) the antecedent or the consequent (as it occurred in the major premise) is negated and that is done by *AND-gate 1* and *AND-gate 4*. By following the diagram carefully, one can see that *AND-gate 1* detects the denial of the antecedent, and *AND-gate 4* detects the denial of the consequent. If the antecedent and the consequent has a negative form in the major premise (e.g. *If ~p then ~q*), and if the minor premise is in the affirmative form (e.g. *p*), *AND-gate 1* will be activated by *AND-gate 2* by the means of an *XOR-gate*. The same principle activates *AND-gate 4*. The role of *AND-gate 5* is to detect double negations. This gate will be active whenever *AND-gate 1* and *2* or *AND-gate 3* and *4* are active. This gate prevents nodes representing *negation* from firing. In order to do correct inferences, *Antecedent* and *Consequent* must be linked. The detection of the antecedent in the question

must enable firing of *consequent* nodes, unless *AND-gate 1* is active (thereby avoiding Denying the Antecedent). The detection of the consequent in the question must enable firing of *antecedent* nodes if *AND-gate 4* is not active (it avoids Affirming the Consequent). Finally, if *AND-gate 1* is active, *AND-gate 4* will be activated, and vice-versa.

## Hypotheses

Classical explanations of negation effects in conditional reasoning rely on the notion of “negative conclusion bias”: a tendency to prefer inferences in the negative form with the exception of MP (Pollard & Evans, 1980).

The first hypothesis that follows from INFERNET is that it should be easier to apply Modus Ponens than any other rule. This effect is attributed to the stronger links from antecedent nodes to consequent nodes. The second hypothesis states that whenever *AND-gate 5* (see figure 4) is needed, a decrease in performance should occur. This effect is due to an increase of the number of steps required to propagate the activation and to this gate’s role of blocking the oscillation of *negation* nodes. *AND-gate 5* is required to treat double negations. Therefore this hypothesis predicts a decrease in DA errors for major premises *If p then ~q* and *If ~p then ~q* and a poorer MT performance for major premises *If ~p then q* and *If ~p then ~q*.

In order to contrast classical and INFERNET hypotheses frequencies of inference and reaction times will be used.

## INFERNET Simulation Results

Normalized correlation between obtained data and different possible answers was computed for the 40 trials. This is a correlation between data observed and data for perfect answers. The proportion of correct responses was obtained by combining the correlations obtained on different trials, taking care to ensure that correlations are not additive (see Sougné, 1999 for computation details). INFERNET simulator results are reported in figure 5.

As expected, MP is more often applied than any other rule. There is also an effect of double negation which is responsible for the low frequencies of DA when the consequent is negated and of MT when the antecedent is negated.

Response times for the simulator are measured by monitoring the encoding phase. After each  $\gamma$  wave burst, the questions are presented and responses are collected. Since the INFERNET simulator has a resolution of 1ms, the response time is determined by the time (in ms) for the normalized correlation to reach a threshold. INFERNET simulator mean reaction times are reported in figure 6. The reaction times show that MP responses are faster than others and that a double negation results in slower reaction times.

An experiment was conducted to provide data that could be compared with INFERNET. Normally, data about negation effects on conditional reasoning do not provide reaction times and are collected with forced choice responses. The comparison between machine and human data will allow us to test INFERNET.

## Experiment and comparison with INFERNET simulator

### Participants and Design

The experiment has a within-subjects design. Forty participants received four major premises in a random order and had to answer four questions for each major premise in a random order. The 40 participants were undergraduate psychology majors, 31 females 9 males, mean age was 21.3 and SD was 2.1.

### Material

Four major premises were constructed, alternating positive and negative antecedents and consequents. Positive antecedent, positive consequent: *If the number is 3 then the letter is X*, Positive antecedent, negative consequent: *If the number is 3 then the letter is ~X*, Negative antecedent, positive consequent: *If the number is ~3 then the letter is X*, and Negative antecedent, negative consequent: *If the number is ~3 then the letter is ~X*. Each major premise presentation was followed by four questions: *The number is 3, what do you conclude?*, *The number is ~3, what do you conclude?*, *The letter is X, what do you conclude?*, *The letter is ~X, what do you conclude?*.

### Procedure

Each participant was seated approximately 50 cm in front of the monitor. One of the randomly chosen major premises appeared on the screen. Participants were asked to read it and to indicate when they understood it. The major premise stayed on the screen when the subsequent questions were displayed. Questions then appeared on the screen, one at the time and in random order. Participants had to answer each question. The computer recorded the time required to respond. The experimenter recorded the response. When the participant answered the four questions, the next major premise appeared on the screen with the same procedure until the four major premises had been presented. Before presenting the experimental material, participants received training exercises with the same procedure, but with an arithmetic content.

### Results

Frequencies of stating each inference are shown on figure 5. According to the “Negative Conclusion Bias” hypothesis, there should be more DA type inferences for major premises *If 3 then X* and *If ~3 then X*, more AC type inferences for major premises *If ~3 then X* and *If ~3 then ~X*, more MT type inference for major premises *If 3 then X* and *If 3 then ~X*, and finally more MP. According to the INFERNET prediction, there should be more MP type inferences than any other, fewer DA type inferences for major premises *If 3 then ~X* and *If ~3 then ~X*, and fewer MT type inferences for major premises *If ~3 then X* and *If ~3 then ~X*.

Data were analyzed by a Loglinear analysis which provides a means to analyze multi-way frequency table. Loglinear analysis evaluates the effect of each variable and



of their interaction<sup>1</sup>. Moreover, loglinear analysis evaluates each model that could explain the data, this gives us a way to compare INFERNET and the classical “Negative Conclusion Bias” model.

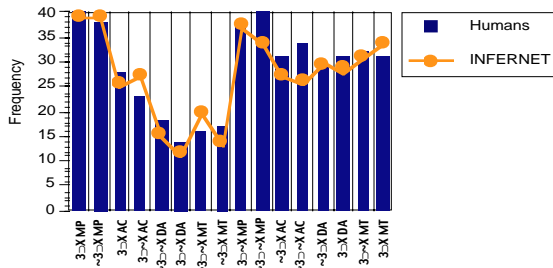


Figure 5: Graph of comparison between human and INFERNET simulator frequencies of inference.

In addition to the effect of the conclusion sign (i.e. with or without “not” in the conclusion) being significant (194 positive and 265 negative conclusions,  $G^2_{(1)}=44.135$ ,  $p<.0001$ , other effects are also. The effect of expected sign is significant, which means that DA + MT (188 inferences) are less often applied than MP + AC (271 inferences),  $G^2_{(1)}=59.358$ ,  $p<.0001$ . Forward inferences (MP+DA) are more often done (247 inferences) than backward inferences (AC+MT) (212 inferences),  $G^2_{(1)}=11.092$ ,  $p<.001$ . The interaction between the expected sign and the conclusion sign is also significant: among the positive conclusions those which involve a double negation are less often inferred (65 inferences) than others (129 inferences) while for negative conclusions cases, expected positive cases (142 inferences) are more comparable with expected negative cases (123 inferences),  $G^2_{(1)}=4.893$ ,  $p<.03$ . There is also an interaction between the expected sign and Forward and backward inferences. MP are more often applied (155 inferences) than AC (116 inferences) while DA (92 inferences) and MT (96 inferences) are sensibly equal,  $G^2_{(1)}=30.226$ ,  $p<.0001$ . The INFERNET model is the best fitting model  $G^2_{(20)}=12.88$ ,  $p=.88$ , while Negative Conclusion bias with the exception of MP cases provides a poor fit:  $G^2_{(22)}=21.92$ ,  $p=.46$ . The difference between these two models is significant:  $G^2_{(2)}=9.04$ ,  $p<.01$ .

The INFERNET data are not significantly different from these results. The comparison with human data can be done by adding one group factor to the analysis (Human vs INFERNET simulator). The effect of group is not significant, none of the interactions are significant.

Figure 6 shows mean reaction times for the 4 major premises and the four questions. The two hypotheses to compare are the same as above. The use of ANOVA<sup>2</sup> with

4 within-subject factors reveals a significant effect of the variables “Positive or Negative Conclusion”:  $F(1,9)=11.02$   $p<.01$  (negative conclusion bias). However, a Post Hoc Tukey test reveals that cases in which the conclusion is negative are only faster than those involving double negation. The double negation effect is significant:  $F(1,9)=12.79$   $p<.01$ . A post hoc Tukey reveals that reaction times for cases of “Double Negation” are significantly longer than others. INFERNET reaction times are faster than those of humans, but INFERNET does not account for the time of reading the question and producing an utterance.

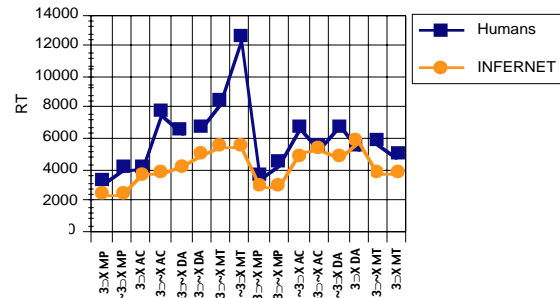


Figure 6: Graph of comparisons between human and INFERNET simulator mean reaction times.

## Conclusions

Connectionist modeling of human reasoning is a difficult challenge. Even though Holyoak & Spellman (1993) have described human reasoning in terms of constraint satisfaction, few connectionist systems has been designed for modeling reasoning. INFERNET shows how reasoning might be possible based on certain low-level neurobiological mechanisms. These properties constrain the reasoning process and explain human limitations. People are sensitive to negated conditionals. INFERNET’s account of the phenomenon involves the type of inference and double negation effects and challenges classical explanations that rely on the notion of “Negative Conclusion Bias”. It was predicted that the number of steps required to perform an inference constrained the reasoning process. Removing double negations requires a long chain of gates opening. The longer the chain of successive gates, the higher the number of errors, and the less opportunity for binding fixation. This paper presented INFERNET’s predictions and results. These results confirmed that INFERNET is sensitive to double negations. A similar experiment has been conducted on human participants. Results confirmed INFERNET’s prediction and showed that the INFERNET explanation is better than classical explanation in terms of “negative conclusion bias”. Finally, INFERNET and humans data were compared and there is a high degree of qualitative similarity between the two.

## Acknowledgments

This research was supported by the Belgian PAI Grant p4/19 Special thanks to Robert French for his assistance in the work presented here.

<sup>1</sup> All the following  $G^2$  are underestimated because data were analysed with a between subjects design. A method for analysing within designs exists but in this case, it would require a  $2^{16}$  table to analyse. However, this would not be feasible. Note, however, that a within-subjects ANOVA gave the same results.

<sup>2</sup> In this analysis, degrees of freedoms have been corrected because of violation of the sphericity assumption: Box correction  $\hat{\epsilon} = .23$

## References

- Abeles, M., Prut, Y., Bergman, H., Vaadia, E. & Aertsen, A. (1993). Integration, Synchronicity and Periodicity. In A. Aertsen (Ed.) *Brain Theory: Spatio-Temporal Aspects of Brain Function*. Amsterdam: Elsevier.
- Cowan, N. (1998). Visual and auditory working memory capacity. *Trends in Cognitive Sciences*, 2, 77-78.
- Engel, A. K., Kreiter, A. K., König, P., & Singer, W. (1991). Synchronisation of oscillatory neuronal responses between striate and extrastriate visual cortical areas of the cat. *Proc. Natl. Acad. Sci. U. S. A.*, 88, 6048-6052.
- Evans, J. St. B. T. (1977). Linguistic factors in reasoning. *Quarterly Journal of Experimental Psychology*, 29, 297-306.
- Evans, J. St. B. T. (1998). Matching Bias in Conditional Reasoning: Do we understand it after 25 years? *Thinking and Reasoning*, 4, 45-82.
- Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human Reasoning: The psychology of deduction*. Hove: Lawrence Erlbaum Associates.
- Evans, J. St. B. T., Clibbens, J., & Rood, B. (1995). Bias in conditional inference: Implications for mental models and mental logic. *Quarterly Journal of Experimental Psychology*, 48A, 644-670.
- French, R. M. (1995). *The Subtlety of Sameness: A theory and computer model of analogy-making*. Cambridge, MA: MIT Press.
- Hawkins, R.D., Kandel, E. R. and Siegelbaum, S. A. (1993). Learning to modulate transmitter release: Themes and variations in synaptic plasticity. *Annu. Rev. Neurosci.*, 16, 625-665.
- Holyoak, K. J., & Spellman, B. A. (1993). Thinking. *Annual Review of Psychology*, 44, 265-315.
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representation of structure: A theory of analogical access and mapping. *Psychological Review*, 104, 427-466.
- Jensen, O. & Lisman, J. E. (1998). An Oscillatory Short-Term Memory Buffer Model Can Account for Data on the Sternberg Task. *The Journal of Neuroscience*, 18, 10688-10699.
- Konnerth, A., Tsien, R.Y., Mikoshiba, K. and Altman, J. (1996). *Coincidence detection in the nervous system*. Strasbourg: Human Frontier Science Program.
- Lisman, J. E., & Idiart, M. A. P. (1995). Storage of  $7 \pm 2$  Short-Term Memories in Oscillatory Subcycles. *Science*, 267, 1512-1515.
- Maass, W. & Bishop, C. M. (1999). *Pulsed Neural Networks*. Cambridge, MA: MIT Press.
- MacKay, W. A. (1997). Synchronized Neuronal Oscillations and their Role in Motor Process. *Trends in Cognitive Sciences*, 1, 176-183.
- Oaksford, M., & Stenning, K. (1992). Reasoning with Conditionals Containing Negated Constituents. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 835-854.
- Oaksford, M., & Chater, N. (1994). A Rational Explanation of the Selection Task. *Psychological Review*, 101, 608-631.
- Pollard, P. & Evans, J. St. B. T. (1980). The influence of logic on conditional reasoning performance. *Quarterly Journal of Experimental Psychology*, 32, 605-624.
- Roelfsema, P. R., Engel, A. K., König, P. & Singer, W. (1996). The role of neuronal synchronization in response selection: A biologically plausible theory of structured representations in the visual cortex. *Journal of Cognitive Neuroscience*, 8, 603-625.
- Shastri, L. and Ajjanagadde, V. (1993). From Simple Associations to Systematic Reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Science*, 16, 417-494.
- Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annu. Rev. Physiol.*, 55, 349-74.
- Singer, W. & Gray, C. M. 1995. Visual Feature Integration and the Temporal Correlation Hypothesis. *Annual Review of Neuroscience*, 18, 555-586.
- Singer, W., Engel, A. K., Kreiter, A.K., Munk, M. H. J., Neuenschwander, S. & Roelfsema, P. R. (1997). Neuronal assemblies: necessity, signature and detectability, *Trends in Cognitive Sciences*, 1, 252-261.
- Sougné, J. (1996). A Connectionist Model of Reflective Reasoning Using Temporal Properties of Node Firing. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Ass.
- Sougné, J. (1998a). Connectionism and the problem of multiple instantiation. *Trends in Cognitive Sciences*, 2, 183-189.
- Sougné, J. (1998b). Period Doubling as a Means of Representing Multiply Instantiated Entities. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Ass.
- Sougné, J. (1999). *INFERNET: A neurocomputational model of binding and inference*. Unpublished doctoral dissertation, Université de Liège.
- Sougné, J. and French, R. M. (1997). A Neurobiologically Inspired Model of Working Memory Based on Neuronal Synchrony and Rythmicity. In J. A. Bullinaria, D. W Glasspool, and G. Houghton (Eds.) *Proceedings of the Fourth Neural Computation and Psychology Workshop: Connectionist Representations*. London: Springer-Verlag.
- Sperber, D., Cara, F., & Girotto, V. (1995). Relevance Theory Explains the Selection Task. *Cognition*, 57, 31-95.
- Wang, X. & Rinzel, J. (1995). Oscillatory and Bursting Properties of Neurons. In A. Arbib (Ed.) *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press.
- Wildman, T. M. & Fletcher, H. J. (1977). Developmental increases and decreases in solutions of conditional syllogism problems. *Developmental Psychology*, 13, 630-636.
- Wilson, M. & Shepherd, G. M. (1995). Olfactory Cortex. In A. Arbib (Ed.) *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press.

# A Dynamic Field Model of Location Memory

John P. Spencer

john-spencer@uiowa.edu

Department of Psychology, University of Iowa

E11 Seashore Hall, Iowa City, IA 52242

Gregor Schöner

gregor@lnf.cnrs-mrs.fr

Center for Research in Cognitive Neuroscience, C.N.R.S.

31, ch. Joseph Aiguier

13402 Marseille Cedex 20, France

## Abstract

One of the fundamental questions in cognitive science is how people remember the locations of important objects in the world with enough accuracy to find these objects when they are no longer in view. Evidence from a variety of studies suggests that people rely on visible reference axes—streets, walls, the edges of a table—to help them remember the locations of target objects. Use of such perceptual information can help stabilize memory, but there is a cost: when people are asked to reproduce the location of a hidden object, they exaggerate the distance between the reference axis and the actual location of the object. These memory errors increase in magnitude as memory delays increase. Thus, errors away from reference axes may provide a window into the second-to-second processes that serve to maintain location information in memory. In the present report, we describe a dynamic field model that captures in quantitative detail how information is maintained in memory near reference axes. This model explains the time-dependent integration of memory and perceptual processes, thereby moving beyond current models of location memory.

## Introduction

To interact successfully with the world, people must remember the locations of important objects with enough fidelity to find these objects when they are no longer visible. This is relatively easy when the target object shares a clear relationship with a visible landmark (e.g., under the landmark). In more challenging situations, however, there may be long delays between perception of the location and action toward it, and the target object may be hidden within a continuously varying scene or a field of view with relatively few perceptual landmarks. In these situations, how do people accurately maintain location information in memory?

Research on the short-term characteristics of location memory has generally focused on what people represent in memory when asked to remember the location of a hidden object. For instance, Smyth and colleagues have demonstrated that people represent locations relative to both egocentric body position and an allocentric reference frame (Smyth, Pearson, & Pendleton, 1988). Other data suggest that humans and non-human primates encode locations in retinotopic coordinates, head-centered coordinates, and shoulder or body-centered coordinates (e.g., Feigenbaum &

Rolls, 1991; Graziano, Hu, & Gross, 1997; Woodin & Allport, 1998).

Although the question “what is represented” is central to the study of location memory, it is equally important to understand how represented information is maintained in memory over short-term delays. In the past decade, this issue has been the focus of several neurophysiological studies (e.g., Constantinidis & Steinmetz, 1996; Rao, Rainer, & Miller, 1997). These studies have demonstrated that the sustained activation of neurons in prefrontal cortex, premotor cortex, and posterior parietal cortex underlie the maintenance of location information over short-term delays. Nevertheless, there have been relatively few behavioral studies investigating the short-term characteristics of location memory. The small number of behavioral studies is surprising given that virtually all of the studies examining what people represent in memory ask participants to reproduce remembered locations following a delay. A prerequisite for understanding the effects in many of these studies, then, is to understand how information about one location is maintained for several seconds.

The relative lack of behavioral data on maintenance processes has also led to a de-emphasis on time-dependent models of location memory. The goal of the present report is to introduce a dynamic model of location memory. This model represents the first attempt to explicitly capture how location information is maintained in memory over short-term delays.

## How Is Location Information Maintained?

One way to help maintain an accurate memory of location over delays is to encode locations relative to visible reference cues in the environment. People might, for instance, encode locations relative to salient perceptual landmarks. This can help stabilize memory, particularly if the landmark is visible from a variety of vantage points (e.g., the Eiffel Tower).

Although the use of landmarks has been well documented (e.g., Sadalla, Burroughs, & Staplin, 1980), the present report emphasizes people’s use of a different, but equally prevalent type of reference cue—visible reference axes. The environments in which people typically act are naturally sub-divided by visible reference axes. Reference axes such

as streets, rivers, and walls sub-divide far spaces, while axes such as the edges of tables, the edges of a computer screen, and the edges of a counter top sub-divide near spaces. Data from a variety of spatial memory studies suggest that, as with visible landmarks, people use reference axes to help them remember locations (e.g., McNamara, Hardy, & Hirtle, 1989). However, the use of reference axes may have a cost. Specifically, when people are asked to reproduce the location of a target object near a reference axis after different delays, responses become systematically distorted away from the reference axis on a second-by-second basis (Spencer, 2000). These delay-dependent effects are central to the present report because they provide insights into the processes that serve to maintain location information in memory.

### Location Memory Biases Near Reference Axes

Spatial priming and free recall studies have shown that adults use reference axes to help them remember locations. Specifically, adults group locations in memory relative to reference axes. McNamara and colleagues, for example, asked adults to learn the locations of multiple objects in a room subdivided by tape on the floor (McNamara et al., 1989). After the layout of objects was learned, participants read pairs of object names presented sequentially on a computer screen and judged if the second object was present in the original layout of objects. Adults responded more quickly when the two objects were in the same spatial region than when they were in different regions. This occurred even if the objects in different regions were physically closer than the objects in the same region. Similarly, free recall of objects and places is ordered relative to reference axes (McNamara et al., 1989). For example, adults use reference axes such as streets and rivers, recalling buildings and businesses from one region before recalling items in adjoining regions.

Although these data demonstrate that people use reference axes to organize location memory, it is difficult to isolate how such axes are used in these tasks because people are asked to remember multiple locations in the presence of many reference cues. Other studies have used much simpler tasks in which people remember a single location on each trial in the context of simple reference cues. In these studies, participants are typically shown a dot inside a geometrical figure. The dot is then covered up, there is a short delay, and participants are asked to reproduce the dot's location in a second, blank figure (e.g., Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher, Newcombe, & Sandberg, 1994).

These studies allow a more complete view of the processes that maintain location information in memory, because factors central to these processes can be directly manipulated. For instance, the length of memory delays and the separation between the target location and the axes of the geometrical figure can be manipulated across trials. In addition, two types of error can be measured—the mean or constant error across responses to the same location, and the variability of these responses. These two measures provide complementary views of how location information is maintained. Constant error indicates both the direction and

magnitude of memory biases. Variable error indicates how stably location information is maintained.

Data from several location reproduction studies demonstrate that location memory is systematically distorted near reference axes. Specifically, memory is biased away from visible reference axes when the to-be-remembered locations are close to these axes. Huttenlocher and colleagues (1994) asked children and adults to reproduce the locations of dots within a rectangular frame. Responses in this study were biased away from the left and right edges of the frame. Similarly, when older children and adults were asked to reproduce the locations of dots within a circle, they made errors away from the edges of the circle (Huttenlocher et al., 1991; Sandberg, Huttenlocher, & Newcombe, 1996). Finally, Engbretson and Huttenlocher (1996) asked adults to reproduce the direction of a line within a “V” frame. Responses were once again biased away from the edges of the frame (see also, Schiano & Tversky, 1992).

In addition to biases away from visible reference axes, responses in these studies were biased away from “mentally imposed” axes. For example, participants made errors away from the vertical midline axis of the rectangular and V frames (Engbretson & Huttenlocher, 1996; Huttenlocher et al., 1994). Similarly, in the circle task, adults made errors away from both vertical and horizontal axes, suggesting they mentally sub-divided the circle into quadrants (Huttenlocher et al., 1991; Sandberg et al., 1996).

Finally, data from a recent study demonstrate that biases away from reference axes increase systematically over short-term delays. Spencer and Hund (2000) asked adults to reproduce the location of targets at different angular distances from the midline axis of a large, homogeneous task space. Participants moved to these remembered locations after delays that ranged from 0 to 20 s. As the delay increased, participants' responses were biased away from midline and became more variable. These delay-dependent effects suggest that errors away from reference axes may be a product of the processes that maintain information in memory. As such, these errors may offer unique insights into how location information is maintained over short-term delays.

### Perceptual Processes and Reference Repulsion

Although delay-dependent results from Spencer and Hund (2000) indicate that memory decay plays a key role in response biases near reference axes, data from several studies suggest that memory processes are not the sole cause of these biases. Instead, perceptual processes contribute to biases near reference axes. Specifically, perceptual judgements of dot location and line orientation are biased away from reference axes. Importantly, these biases occur even though, in many studies, reference and target displays are presented simultaneously. Thus, errors away from reference axes in these studies cannot be caused by memory processes.

For instance, when a test line abuts a visible reference line forming an acute angle, people report the angle is larger than it actually is. This *acute-angle expansion* or *tilt contrast* effect is maximized at small angles and if the

reference line is horizontal or vertical (e.g., Blakemore, Carpenter, & Georgeson, 1970). Judgements of line orientation are also repelled from “virtual” reference axes (e.g., Beh, Wenderoth, & Purcell, 1971). Virtual reference axes result from the symmetry properties of geometrical figures. A square, for example, has four virtual reference axes—two diagonals and horizontal and vertical axes. Beh et al. (1971) showed that when adults are asked to judge the orientation of a rod in the context of a square frame, adults’ judgements are repelled from the closest axis of symmetry defined by the square frame. Such repulsion is particularly strong near horizontal and vertical axes.

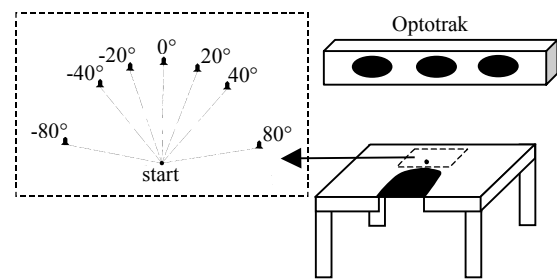
Finally, adults’ judgements of dot position are repelled from visible reference axes. Rauber and Treue (1998) asked adults to judge if two sequentially presented locations were identical. When these locations were close to a vertical reference line, adults’ judgements were repelled from the reference line. This effect decreased as the separation between the target location and the reference line increased.

To summarize, data suggest that both memory processes and perceptual processes contribute to response biases near visible reference axes. Response biases increase systematically over delays, suggesting that these errors are caused, in part, by how location information is maintained in memory. However, responses are also biased away from reference axes when reference and target displays are presented simultaneously, suggesting that perceptual processes play an important role.

Here we present a formal model that brings together perceptual and memory processes to explain the origin of response biases near reference axes. Central to this account is the proposal that initial biases in perceptual processes are amplified in memory over short-term delays. Specifically, our model demonstrates how enhanced perceptual processing of visual information near reference axes can produce both biases in perceptual judgements and biases in how information is maintained in memory.

### Empirical Results to be Modeled

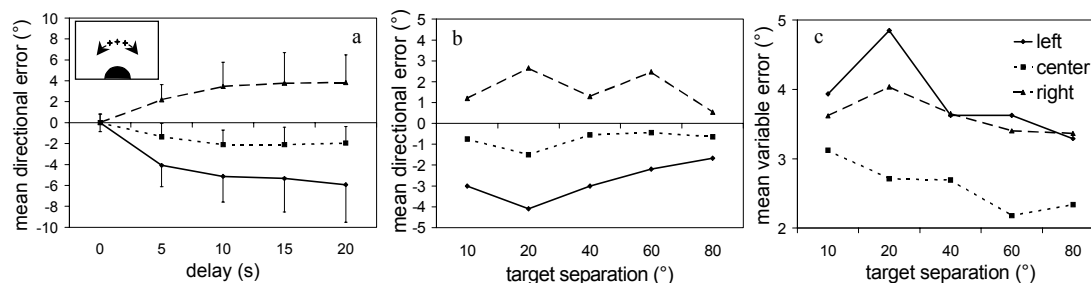
The model we propose here was designed to capture data from several experiments that have explicitly investigated the time-dependent processes that underlie reference repulsion (Spencer & Hund, 2000). In these experiments, participants pointed to target locations projected on a large, opaque tabletop. Pointing movements were tracked using an optical-electronic motion-analysis system (Optotrak, Northern Digital, Inc.). Importantly, the layout of all visible



**Figure 1.** Schematic of apparatus. Targets projected on tabletop from below. Movements recorded using an Optotrak motion analysis system. Inset shows sample target positions relative to starting point.

reference cues were explicitly controlled (Figure 1). The experimental table was quite large (4' x 5') and the surface was homogeneous. Experimental sessions were conducted in dim lighting in a room with black curtains covering the walls and ceiling. This prevented participants from using external landmarks. Nevertheless, the hand, body, and table were clearly visible. Participants sat in a chair positioned within an arc cut out from one edge of the table (Figure 1). This limited their ability to use the front edge of the table as a reference location. Thus, in this task setting, visible reference cues included the edges of the table and its axes of symmetry, the location of the participant’s body and hand, and any reference locations projected onto the surface of the table.

In the first experiment (Spencer & Hund, 2000), participants moved to three target locations—a left, center, and right location—15 cm from a starting position located on the midline axis of the table. The delay (0-20 s) between the offset of a target light and a “go” signal was varied as was the angular distance of the targets from midline (10°, 20°, 40°, 60°, 80°; see Figure 1). Based on the results of studies by Huttenlocher and colleagues, we expected participants to make errors away from the midline reference axis, particularly when targets were close to 0°. The key question was how these errors would change as a function of delay. As the delay increased in the 10°, 20°, and 40° conditions, constant directional errors to the left target became significantly more negative (larger counterclockwise errors), and errors to the right target became significantly more positive (larger clockwise errors) (Figure 2a). Errors to the center target remained small across all delays. At larger target separations, the magnitude



**Figure 2.** (a) Mean directional errors over delays for movements to the left, center, and right targets in the 20° condition. Inset shows a schematic of target locations and mean directional error. Error bars = ½ standard deviation. (b) Constant (mean) and variable (standard deviation) directional errors for movements to each target location across target separation conditions.

of the errors away from midline decreased (Figure 2b) and no longer depended on delay. Variable (standard deviation) directional errors also increased over delays (see errors bars in Figure 2a). As with the constant errors, this effect was larger for movements to the left and right targets than to the center. Variability was largest in the 20° condition and decreased significantly at the other target separations (Figure 2c).

Results from this study indicate that location memory decays over short-term delays. More importantly, however, these results reveal a specific pattern of decay near a reference axis. Both constant and variable error increased over delay when participants moved to the left and right locations, and both types of error remained small when participants moved to midline. In addition, errors were largest at 20°, and decreased as the left and right targets were moved farther from midline. Thus, reference repulsion decreased for targets far from midline. Finally, there was a reduction in both constant and variable error very close to midline (at 10°).

In a second experiment, we found similar delay-dependent effects near reference axes, demonstrating that decay effects generalize to conditions in which the three targets are not symmetrically positioned around the midline axis of the table (Spencer & Hund, 2000). Specifically, we rotated the three targets clockwise and counterclockwise around the midline axis. For example, in one condition, targets were located at -60°, -40°, and -20°, while in another condition, targets were located at -40°, -20°, and 0°. Across all modified layout conditions, participants' responses to non-0° targets were repelled from midline as delays increased. In addition, the magnitude of these errors decreased as the targets were rotated away from midline (e.g., from -20° to -40° to -60°). Finally, participants' responses to targets along the midline axis were accurate with low variability.

## The Model

To explain the pattern of memory decay near reference axes, we propose the following dynamic field model. This model specifies how perceptual and memory processes are integrated over delays to produce reference repulsion. Although this model represents a new approach to location memory, several of the concepts we discuss here have been used to capture how adults plan reaching movements to visually specified target locations (Schöner, Dose, & Engels, 1995).

The starting point for our dynamic field model is the concept of an activation field, where “activation” indicates the likelihood that a participant will move to a specific location at a particular moment in time. Plans to move to a target can be thought of as distributions of activation values across all possible target locations, with higher values indicating that a person is more likely to move to these locations than to others.

Two different types of information are integrated within the activation or action planning field. The first input—target input—captures the appearance and disappearance of the target light. The second input—P-ACT input—represents a participant's memory of previously activated

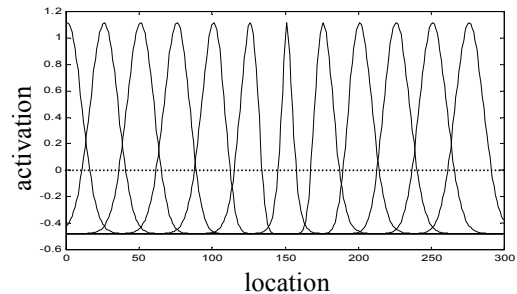
locations. The integration of these inputs in the planning field is governed by an interaction function that determines how activation at one site in the field influences activation at other sites. We use a local excitation and lateral inhibition interaction function. Thus, activation at one site increases the activation of its neighbors and decreases the activation of sites far away. One consequence of this function is that strong input can produce “self-sustaining” activation. Such patterns of activation maintain themselves, even after input is removed. In this way, *the planning field can maintain a memory of the input*.

The main concepts of the dynamic field model are captured in Equation 1. This equation specifies how activation in the planning field changes from time step to time step. Specifically, the change in activation at the next time step is a function of the current activation, the current inputs (target, P-ACT), and the way current above-threshold activation at each site enhances or suppresses activation at all other sites (local excitation/lateral inhibition function). Once computed, the change in activation is added to the current activation to produce the activation in the planning field at the next time step.

$$\tau \dot{u}(x, t) = -u(x, t) + h + \int dx' w(x, x') f(u(x', t)) + S_{tar}(x, t) + S_{pact}(x, t) + S_{noise}(x, t)$$

**Equation 1.** Rate of change in field activation =  
- current activation + base activation + sum(current thresholded activation weighted by the interaction function) + target input + P-ACT input + spatially-correlated noise

Central to our dynamic field model is the way the model integrates perceptual and memory processes. We propose that interaction is not identical across all sites in the field, i.e., interaction is inhomogeneous. Specifically, interaction is more precise at sites associated with visible reference axes due to enhanced spatial tuning of visual processes at these locations. Thus, local excitation will be narrowly distributed at sites associated with a reference axis—the focus of enhancement—and broadly distributed across sites associated with regions of “empty” space. This is depicted in Figure 3. This figure shows the local excitation/lateral inhibition functions (positive/negative values) at twelve different locations in a sample field. The focal point in this example is at location 150, reflecting the presence of a reference axis at this central location. Local excitation is



**Figure 3.** Sample modulation of local excitation/lateral inhibition function around the focal point  $x_0$  (location 150).

most narrow at location 150, and expands to the left and right of this location.

The type of inhomogeneity we propose is conceptually similar to a proposal made by Suzuki and Cavanagh to account for repulsion effects in visual processing (1997); however, according to our proposal, inhomogeneities in visual processes couple directly into the processes that maintain location information in memory (via the interaction function). This has two effects on activation in the planning field over memory delays. First, activation peaks situated on or near the focus of enhancement will be attracted toward this focal point and remain stably positioned over delays. Second, activation peaks further from the focus of enhancement will be repelled from this focal point, because peaks gravitate toward sites with broader local excitation.

The form of inhomogeneous interaction we use is shown in Equations 2 and 3. We use an exponential form of inhomogeneity centered at the focal point  $x_0$ —the site associated with the reference axis. The width of local excitation is modulated across the spatial range specified by  $\sigma_\sigma$ , and the magnitude of the modulation is specified by the amplitude parameter ( $A_\sigma$ ).

$$w(x, x') = \frac{G_{\text{int}}}{\sqrt{2\pi}\sigma_{\text{int}}} \left\{ -w_i + \exp\left[-\frac{(x-x')^2}{2\sigma(x)^2}\right] \right\}$$

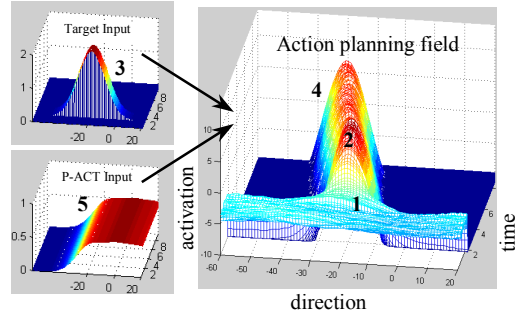
Equation 2. The interaction function is specified by  $w(x, x')$ , a local excitation/lateral inhibition function.

$$\sigma(x) = \sigma_{\text{int}}(1 - A_\sigma \exp[-(x - x_0)/\sigma_\sigma])$$

Equation 3. The interaction function (Equation 2) is inhomogeneous because the effective width of local excitation,  $\sigma(x)$ , depends on the field location  $x$ .

## Model Results

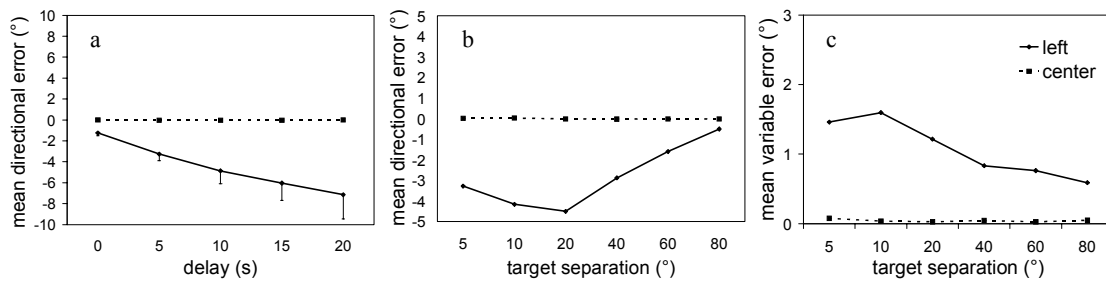
Figure 4 shows a simulation of the dynamic field model that captures delay-dependent results from the 20° condition in Spencer and Hund (2000). Recall that in this condition, participants moved to three targets positioned symmetrically about midline. Figure 4 shows how activation in the planning field evolves from second to second during one trial in which the target is presented at -20°. The lower left



**Figure 4.** Simulation of the dynamic field model. “Input” axes are identical to axes in field graph. Numbers mark events during a single trial. See text for details.

panel of Figure 4 shows the P-ACT input. Activation in this panel is high near -20°, 0°, and 20°, reflecting a participant’s memory of activation at these sites on previous trials. For simplicity, we assume that this input is relatively constant during a 20 s trial. The upper panel shows the target input. Activation in this panel is zero at the start of the trial when the target is not visible, high at -20° when the target is turned on, and zero again when the target is turned off. The P-ACT and target inputs are integrated within the action planning field shown in the right panel. At the start of the trial, the planning field is slightly “pre-activated” at previously moved-to locations (1). This reflects the P-ACT input. Next, a target is turned on and a peak of activation builds up in the planning field at the target direction driven by the strong target input (2). This input generates a peak that is maintained even after the target is turned off (3). Finally, during the delay, the peak is repelled from the focus of enhanced interaction (0°) and drifts away from midline (4). This effect is partially counteracted by the P-ACT input which attracts the peak toward previously activated locations, in this case, toward 0° (5).

The model depicted in Figure 4 not only captures how information is maintained in memory on a single trial, but also the delay-dependent pattern of constant and variable errors reported in Spencer and Hund (2000). Figure 5 shows simulation results from 100 iterations of the model in which the location of maximal activation in the field was read-out at different delays. As can be seen in this figure, the constant and variable errors computed from simulations of the model capture the pattern of error shown in Figure 2 across both delays and target separations.



**Figure 5.** (a) Directional errors over delays for simulated trials to the left and center targets in the 20° condition. (b) Constant and (c) variable directional errors for simulated trials to these targets across target separation conditions.

## Discussion

The dynamic field model effectively integrates the perceptual and memory processes that the literature suggests underlie biases away from reference axes. This model moves beyond the capabilities of previous location memory models in three fundamental ways. First, the field model proposes a specific integration mechanism that captures how location information is maintained from second-to-second over short-term delays. No current models of location memory are explicitly time-based. Second, the field model effectively reproduces time-dependent changes in both constant and variable errors near reference axes. Most models of location memory account for biases near reference axes; however, we know of no models that capture patterns of both bias and variability. Third, due to the “intrinsic” properties of the planning field, this field can generate its own activation in the *absence of input*. Thus, the field model can generate behavior that does not directly mirror the characteristics of input. Consequently, our model moves beyond models of location memory that posit that response biases are due to the relative weighting of inputs (e.g., Huttenlocher et al., 1991).

Finally, it is important to note that response biases in the dynamic field model are not solely a function of inhomogeneous interaction. As noted in Figure 4, the localization of activation peaks in the field is caused by the relative strength of repulsion effects (inhomogeneous interaction) and attraction effects (attraction toward P-ACT input). This has two important consequences. First, the field model may account for a second class of response biases prevalent in the spatial memory literature—attraction toward “prototypical” locations. We are currently exploring this possibility. Second, by changing the relative strength of repulsive and attractive effects, we may be able to capture striking differences in the performance of individuals. Consequently, the dynamic field model may offer insights not only into group effects, but also into the origins of individual differences in memory performance.

## Acknowledgements

This work was supported by a CIFRE grant to John P. Spencer from the University of Iowa and an Interdisciplinary Research Grant to John P. Spencer and Gregor Schöner from the Obermann Center for Advanced Studies.

## References

- Beh, H., Wenderoth, P., & Purcell, A. (1971). The angular function of a rod-and-frame illusion. *Perception and Psychophysics*, *9*, 353-355.
- Blakemore, C., Carpenter, R., & Georgeson, M. (1970). Lateral inhibition between orientation detectors in the human visual system. *Nature*, *228*, 37-39.
- Constantinidis, C., & Steinmetz, M. A. (1996). Neuronal activity in posterior parietal area 7a during the delay periods of a spatial memory task. *Journal of Neurophysiology*, *76*, 1352-1355.
- Engelbreton, P. H., & Huttenlocher, J. (1996). Bias in spatial location due to categorization: Comment on Tversky and Schiano. *Journal of Experimental Psychology: General*, *125*(1), 96-108.
- Feigenbaum, J. D., & Rolls, E. T. (1991). Allocentric and egocentric spatial information processing in the hippocampal formation of the behaving primate. *Psychobiology*, *19*(1), 21-40.
- Graziano, M. S. A., Hu, X. T., & Gross, C. G. (1997). Coding the locations of objects in the dark. *Science*, *277*, 239-241.
- Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, *98*(3), 352-376.
- Huttenlocher, J., Newcombe, N., & Sandberg, E. H. (1994). The coding of spatial location in young children. *Cognitive Psychology*, *27*, 115-147.
- McNamara, T. P., Hardy, J. K., & Hirtle, S. C. (1989). Subjective hierarchies in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 211-227.
- Rao, S. C., Rainer, G., & Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, *276*, 821-824.
- Rauber, H.-J., & Treue, S. (1998). Reference repulsion when judging the direction of visual motion. *Perception*, *27*, 393-402.
- Sadalla, E. K., Burroughs, W. J., & Staplin, L. J. (1980). Reference points in spatial cognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 516-528.
- Sandberg, E. H., Huttenlocher, J., & Newcombe, N. (1996). The development of hierarchical representation of two-dimensional space. *Child Development*, *67*, 721-739.
- Schiano, D. J., & Tversky, B. (1992). Structure and strategy in encoding simplified graphs. *Memory and Cognition*, *20*(1), 12-20.
- Schöner, G., Dose, M., & Engels, C. (1995). Dynamics of behavior: Theory and applications for autonomous robot architectures. *Robotics and Autonomous Systems*, *16*, 213-245.
- Smyth, M. M., Pearson, N. A., & Pendleton, L. R. (1988). Movement and working memory: Patterns and positions in space. *Quarterly Journal of Experimental Psychology*, *40A*, 499-514.
- Spencer, J. P., & Hund, A. M. (2000). Location memory biases induced by experience-dependent and visually based reference frames. *Manuscript in preparation*.
- Suzuki, S., & Cavanagh, P. (1997). Focused attention distorts visual space: An attentional repulsion effect. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(2), 443-463.
- Woodin, M. E., & Allport, A. (1998). Independent reference frames in human spatial memory: Body-centered and environment-centered coding in near and far space. *Memory and Cognition*, *26*, 1109-1116.



# A Simple Categorisation Model of Anaphor Resolution

Andrew J. Stewart (Andrew.Stewart@Unilever.com)  
Frederic Gosselin (gosselif@psy.gla.ac.uk)  
Department of Psychology, 52 Hillhead Street,  
University of Glasgow, Glasgow G12 2QQ,  
United Kingdom.

## Abstract

In this paper we examine the way in which approaching the task of anaphor resolution as a categorisation problem can shed light on the possible mechanisms underlying pronoun resolution. We formulate a model of anaphor resolution data within SLIP (Strategy Length & Internal Practicability) (Gosselin & Schyns, 1997, 1999), a general categorisation framework. We chiefly focus on pronominal anaphors in this paper but we also report the results of modelling repeat name anaphor reading time data collected by Stewart, Pickering and Sanford (in press). The success of adopting the redefinition of anaphor resolution as a categorisation problem suggests that problems faced by the cognitive system that have been considered unique to particular processing domains might be understood at a more cognitively general level.

## Introduction

In this article we bring together work on categorisation and work on psycholinguistics. We adopt a particular psycholinguistic phenomenon as a case study and examine it within a categorisation framework. We illustrate what a categorisation perspective can offer psycholinguistics in terms of theoretical apparatus. We examine the performance of a model formulated within the SLIP (Strategy Length & Internal Practicability) categorisation framework (Gosselin & Schyns, 1997, 1999), and show that it can account for human behaviour in pronoun resolution, a problem common in language processing.

We begin by reviewing existing work on pronoun resolution. Then we move on to our proposal which redefines the task of pronoun resolution as a categorisation problem. Following this we turn to outlining the SLIP framework. Finally, we discuss the consequences of redefining pronoun resolution as a categorisation problem and examine the correspondence between our model's predictions and experimental data.

## Existing Psycholinguistic Work on Pronoun Resolution

Anaphors are expressions that refer back to characters mentioned in a text. One example of an anaphor is a pronoun. Consider the fragment of sentence (A) up to but including the pronoun 'he'.

(A) John blamed Bill because he had damaged John's car.

This pronoun could refer to either character. Based on the information conveyed by the pronoun itself, the only restriction is that it refers to a singular male character. As both potential antecedents match on these features the sentence could plausibly continue like sentence (A) or (B):

(B) John blamed Bill because he didn't really like Bill.

In (A) the pronoun is coreferential with the character 'Bill', while in (B) it is coreferential with the character 'John'. There are a number of cues available in the text to facilitate the process of identifying the appropriate pronominal referent.

## Grammatical role cues

One cue is the grammatical positions occupied by the potential antecedents. The word 'John' occupies the grammatical subject position, while 'Bill' occupies the grammatical object position. A number of psychological theories, e.g. Subject Assignment Strategy (Stevenson, Nelson, & Stenning, 1995) and Parallel Function Strategy (Sheldon, 1974), predict a preference to interpret the referentially ambiguous pronoun in the above examples as coreferential with the grammatical subject (although for different reasons).

Note that in the examples discussed in this paper the character occupying the grammatical subject position is also the first mentioned character. Gernsbacher (Gernsbacher & Hargreaves, 1988; Gernsbacher, 1989) proposed that the first mentioned character occupies a privileged position in the reader's discourse model. A similar first mention privilege has been observed in other tasks (e.g. Neath, 1993; Neath & Knoedler, 1994). One of the consequences of the first mention preference found in language comprehension is that later in a sentence it is relatively easy to refer to the first mentioned character.

## Gender cues

In addition to grammatical position information, other cues may also be present. Consider sentences (C) and (D) below.

(C) John blamed Mary because she broke the window.

(D) John blamed Mary because he was in a bad mood.

The gender differentiation between the two characters serves as an additional (strong) cue as to which character the pronoun can refer. However, even under conditions where gender information can unambiguously identify the appropriate pronominal referent, there is much evidence to suggest that the system does not immediately take advantage of this (Stevenson & Vitkovitch, 1986; MacDonald & MacWinney, 1990; Tyler & Marslen-Wilson, 1982). It appears that gender information is treated simply as another cue, not in any way qualitatively distinct from other factors.

## Semantic cues

A particularly strong semantic cue known as implicit causality (Garvey & Caramazza, 1974) can also facilitate interpreting the pronoun. Implicit causality is a property associated with a particular set of verbs which influences processing of the pronoun in constructions such as 'John blamed Bill because he...'. It is manifested as a bias to interpret the pronoun as consistent with the implied locus of cause underlying the described event; such as the action of 'blaming' in this example. 'Blame' is classed as an NP2 biasing verb as it biases toward the character occupying the second Noun Phrase as the causal locus. Similarly there are also verbs such as 'fascinate' which bias toward the first Noun Phrase.

The explicit cause information contained in the subordinate clause (e.g. 'broke the window') is an important disambiguating cue. In Example (B) the fragment 'didn't really like Bill' indicates that the pronoun should be interpreted in a manner inconsistent with the implicit causality bias. The causality congruency effect (Garvey & Caramazza, 1974; McDonald & MacWhinney, 1995) is the finding that it takes longer to read a sentence where the implicit cause and explicit cause conflict than when they are consistent with each other.

So then, the cues available to aid identification of a pronoun's referent include order of mention, implicit cause, gender and explicit cause. Given the restriction that gender and explicit cause must agree, the set of all possible combinations of cues has a cardinality of 8. This total set is shown in Table 1 with example sentences exhibiting those features and with the mean reading times associated with reading the disambiguating fragment, i.e. the explicit cause (Stewart, Pickering & Sanford, in press).

Compared to the large body of work proposing and investigating possible parsing mechanisms, there are relatively few formal theories of pronoun resolution.

## Centering Theory

An adequate explanation of a process requires reference to a possible formal mechanism underlying that process and, for pronoun resolution, must take into consideration factors such as gender agreement and implicit causality verb biases. Centering Theory (Gordon, Grosz & Gilliom, 1993) is the best articulated theory in the literature. Centering proposes that utterances have associated with them a set of forward and a set of backward looking centres. The forward looking centre contains as its members entities, one of which forms the referential link between one utterance and the next. Factors such as the grammatical role of the characters in a text influence the ordering of the prominence of each of these entities. The backward looking centre of an utterance contains one member; the entity used to maintain reference between that utterance and the one preceding. Centering theory is a descriptive theory, rather than a processing theory, in as much as it describes the nature of the referential cohesion between units of a text. Although it describes what

information might be used to facilitate pronominal reference resolution, it doesn't formalise how that information is used. This is hardly surprising as the theory originally grew out of work in Artificial Intelligence and so was never designed as a psychological model. How might a formal psychological model of pronoun resolution be arrived at? We propose that a possible way in which to arrive at a formal model of pronoun resolution is to make the explicit analogy between the problem faced by the processor in pronoun resolution and the problem faced by the processor in tasks of categorisation. In fact, at an important computational level we believe these problems are one and the same. There are many formal categorisation models and we believe that one in particular can be reinterpreted as a formal model of pronoun resolution.

## Mapping the problem of pronoun resolution onto that of categorisation

Let us return to Example (A), repeated below,

(A) John blamed Bill because he had damaged John's car.

The problem upon encountering the pronoun 'he' in this sentence can be understood as one of deciding of which category it is a member: should it be interpreted as a member of the set of expressions referring to the character 'John' or as a member of the set of expressions referring to the character 'Bill'? Furthermore, as we have discussed in above, this decision process is guided by explicit cause (and by gender, when it is relevant) and, to a lesser extent, by first mentioned character and by implicit causality information; these cues can be treated as features because they are discriminable parts of sentences that may be diagnostic with respect to the pronominal referent. Thus, a strong analogy can be made between problems of pronoun resolution and problems of categorisation. We shall study this parallel more thoroughly in the next section.

## A Categorisation Mechanism

SLIP (Strategy Length & Internal Practicability) was originally developed to model the results of experiments examining basic-levelness (Gosselin & Schyns, 1997, 1999). In this section we informally describe the SLIP framework and suggest how it can be used to model performance when faced with the type of categorisation problem required in identifying a pronominal referent. We provide a more complete treatment of this model in the Appendix.

We believe that pronoun resolution can be construed as a two-stage categorisation process. In the first stage, a hypothesis as to which referent is the most likely is generated. This is followed by the testing of this hypothesis. In the first stage, a SLIP categoriser extracts features randomly from the first half of the sentence. As soon as one critical feature is selected, a hypothesis is formulated. We believe that the first stage is informed by

Table 1. Total set of feature combinations with example sentences, reaction times reported in Stewart, Pickering & Sanford (in press), Experiment 4 and theoretical predictions of our categorisation model.

Sentence	Features							RT	Prediction
	F	NP1	NP2	G1	G2	CH1	CH2		
	1	1	0	1	0	1	0	1695	3.511
(1)	John fascinated Mary because he was very interesting.								
	1	1	0	0	1	0	1	1980	9.851
(2)	Mary fascinated John because he was easily interested.								
	1	1	0	1	1	1	0	1983	7.146
(3)	John fascinated Bill because he was very interesting.								
	1	1	0	1	1	0	1	2234	20.864
(4)	John fascinated Bill because he was easily interested.								
	1	0	1	1	0	1	0	1769	6.681
(5)	John blamed Mary because he was in a bad mood.								
	1	0	1	0	1	0	1	1641	6.681
(6)	Mary blamed John because he broke the window.								
	1	0	1	1	1	1	0	1893	14.005
(7)	John blamed Bill because he was in a bad mood.								
	1	0	1	1	1	0	1	1919	14.005

the first mentioned character and the implicit causality information. Order of mention is relatively salient and trivially recovered from the input. Au (1986) demonstrated that implicit causality information is also a very salient property. Both order of mention information and implicit causality contain some degree of uncertainty but they are also both useful predictors as to which way a sentence is going to continue (Garvey, Carmazza & Yates, 1975). The first mentioned character feature (F) can lead only to hypothesis\_1, i.e. the hypothesis that the first referent is the pronominal referent. The implicit causality information, however, favours hypothesis\_1 if the NP1 biasing implicit causality feature (NP1) is present in the sentence and hypothesis\_2 (the hypothesis that the second mentioned character is the pronominal referent) otherwise.

Consider again the first portion of our example sentences (1) and (5) in Table 1:

- (1) John fascinated Mary because he...  
 (5) John blamed Mary because he...

In the first case, the probability that hypothesis\_1 will win is 1 because the two diagnostic features (first mention and implicit causality) both suggest that hypothesis\_1 is appropriate. This is true of the first four example sentences in Table 1. For sentence (5) however, the probability that hypothesis\_1 will win is only .5 as the two features contradict each other. This is true of example sentences (5)-(8).

The hypothesis that was adopted in the first stage and the diagnosticity of gender both influence which verification strategy will be adopted in the second stage.

Suppose, for instance, that a categoriser is presented example sentence (1) from Table 1:

- (1) John fascinated Mary because he was very interesting.

At the end of stage one, the categoriser knows that gender information is relevant and it makes the hypothesis

that 'John' is the correct referent (i.e. hypothesis\_1). The extraction of either feature G1 or feature CH1 in the rest of the sentence verifies this hypothesis.

SLIP postulates a categoriser with a feature-extraction mechanism with a stochastic component. It is thus very likely that some features that are picked up by the categoriser are noninformative. For sentence (1), hypothesis\_1 will ultimately be verified but this can take time. In the SLIP framework it is simple to compute the number of features, on average, that will be needed to be picked up for the categoriser to reach a decision (see Appendix). This is the measure reported in the simulation. The predictions of our model for all the sentences are shown in Table 1 together with reading time data reported in Stewart, Pickering and Sanford (in press).

Let us contrast the treatment of sentence (1) with one identical on all points except for gender diagnosticity. A categoriser is presented with sentence (3) from Table 1:

- (3) John fascinated Bill because he was very interesting.

At the end of the first stage, hypothesis\_1 is generated and gender information is known to be nondiagnostic. We thus have one nondiagnostic gender feature and one diagnostic CH1 feature in this case (i.e. CH1). In the terminology of the SLIP framework, this sentence has less redundancy than sentence 1. After a while, hypothesis\_1 is also verified, but it takes longer to verify it in sentence (3) than in sentence (1) because of the lower redundancy of diagnostic information.

We now compare the first two situations with a third one in which the hypothesis formulated at the end of stage 1 is rejected in stage 2. A categoriser is shown example sentence (2) from Table 1:

- (2) Mary fascinated John because he was easily interested.

At the end of stage 1, hypothesis\_1 is proposed and gender is known to be diagnostic. This is similar to the outcome of stage 1 for sentence (1). Either G1 or CH1

would verify the hypothesis. Neither is present in the second portion of sentence (2) as the explicit cause information points to the second mentioned character (CH2). Thus, hypothesis\_1 needs to be rejected and hypothesis\_2 accepted. In the SLIP framework it is possible to compute a stop criterion based on an acceptable error rate so that if this criterion is reached, a revision of the hypothesis is made, i.e. the alternate hypothesis is adopted. In our simulation we have set the stop criteria at 11%, the error rate observed by Stewart, Pickering and Sanford (in press) (Experiment 4). Rejection of a hypothesis takes longer than verification of that hypothesis.

For sentences (5)-(8) from Table 1, the situation is slightly more complicated. Half the time hypothesis\_1 is selected in stage 1; half the time, hypothesis\_2 is selected. The average number of features that will be needed to be extracted before a decision can be made is the mean of that measure for the two possibilities. Take, for instance, example sentence (5) from Table 1:

(5) John blamed Mary because he was in a bad mood.

When hypothesis\_1 is proposed, the treatment of sentence (5) becomes equivalent to example sentence (1) already discussed; when hypothesis\_2 is elected, however, its treatment becomes equivalent to example sentence (2). So, the average number of features extracted before a decision is reached in sentence (5) is the mean of that in sentences (1) and (2). Arriving at a decision for sentence (5) is slower than (1) but faster than (2).

Stewart, Pickering and Sanford (in press) report the results of three further experiments examining the processing of anaphors in the context of sentences containing cues identical to the ones present in Experiment 4. The most important difference between those experiments and their Experiment 4 is that, while the anaphors in Experiment 4 are all pronouns, those in the remaining experiments are a mixture of ambiguous pronouns and unambiguous repeat names. In this paper we argue that the case of anaphor resolution can be reformulated as one of categorisation. Our main focus has been on the processing of anaphoric pronouns. To strengthen our argument, we need to show that our model also accounts for the processing of other types of anaphor. In addition to modelling Experiment 4 from Stewart, Pickering and Sanford (in press), we also modelled their Experiments 2 and 3 (deep processing condition). The raw Pearson correlations between the models' best predictions and the experimental data are .884 ( $p < .05$ ; best predictions: 1.12286, 1.11834, 1.12060, 1.12060, 1.15261, 1.28229, 1.25107, 1.25107 in the order of Stewart, Pickering, & Sanford's Table 1), .817 ( $p < .05$ ; best predictions: 1.20796, 1.64386, 1.42591, 1.42591, 1.38960, 2.02429, 1.69340, 1.69340 in the order of Stewart, Pickering, & Sanford's Table 1), and .816 ( $p < .05$ ), respectively, for Experiments 2, 3, (deep-processing condition), and 4. So, not only can our model correctly predict the reading time data associated with processing pronouns reported in Stewart, Pickering and Sanford (in

press), it can also correctly predict the reading times associated with the processing of more general anaphoric expressions.

## Discussion

Our categorisation function explains the first mention effect (Gernsbacher & Hargreaves, 1988; Gernsbacher, 1989), the causality congruency effect (Caramazza, Grober, Garvey & Yates, 1977; Ehrlich, 1980; Garnham, Oakhill & Cruttenden, 1992), and the effect of gender diagnosticity (Caramazza et al, 1977; Garnham et al, 1992) reported in the psycholinguistic literature. As outlined above, the first mention privilege is the finding that the first mentioned character is easy to later refer to within the sentence in which it appears. By considering the first mentioned character as 'special', and by associating a feature with it, SLIP performs more quickly when this character is the pronominal referent than when it is the second mentioned character. In other words, our model predicts that pronoun resolution is relatively straightforward when a pronoun refers to the first mentioned character. Our model also accounts for the causality congruency effect. It predicts that pronouns are more difficult to resolve when they occur in a sentence containing an NP1 implicit cause and an NP2 explicit cause. Our model predicts that the causality congruency effect will not be found for NP2 implicit cause verb conditions where the explicit cause is NP1. This is because the first mention privilege allows some difficulty that arises as a result of the implicit causality inconsistency to be overcome. In other words, our model predicts that, all other things being equal, the causality effect is asymmetrical. Although the causality congruency effect has been widely reported in the literature (McDonald & MacWhinney, 1995), possible accounts of its asymmetrical nature have never been provided. Finally, our model predicts that it should be easier to identify a pronoun's antecedent when gender information differentiates between possible referents (Caramazza et al, 1977; Garnham et al, 1992). Additionally, it also offers a computational explanation for why this is the case. In light of the close correspondence between our model's predictions and well-established psycholinguistic phenomena it is clear that not only does our categorisation function successfully characterise human performance on tasks of anaphor resolution, it also provides an explanation at the level of categorisation with respect to why this pattern of performance arises.

The success of SLIP on tasks as (apparently) diverse as anaphor resolution and basic level categorisation suggests that other types of cognitive tasks may also benefit from their reinterpretation as categorisation problems. Understanding the degree to which computational problems faced by the cognitive system in specific processing domains can be interpreted as specific instances of more general problems allows for the proposal of mechanisms of greater explanatory power than those currently suggested in (for example) the literature on anaphor resolution.

## References

- Au, T.K. (1986). A verb is worth a thousand words: The causes and consequences of interpersonal events implicit in language. *Journal of Memory and Language*, **25**, 104-122.
- Caramazza, Grober, Garvey, and Yates (1977). Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behaviour*, **16**, 601-609.
- Ehrlich, K. (1980). Comprehension of pronouns. *Quarterly Journal of Experimental Psychology*, **32**, 247-255.
- Garnham, A., Oakhill, J. and Cruttenden, H. (1992). The role of implicit causality and gender cue in the interpretation of pronouns. *Language and Cognitive Processes*, **7**, 231-255.
- Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry*, **5**, 459-464.
- Garvey, C., Caramazza, A. and Yates, J. (1976). Factors underlying assignment of pronoun antecedents. *Cognition*, **3**, 227-243.
- Gernsbacher, M.A. (1989). Mechanisms that improve referential access. *Cognition*, **32**, 99-156.
- Gernsbacher, M.A., & Hargreaves, D.J. (1988). Accessing Sentence Participants: The Advantage of First Mention. *Journal of Memory and Language*, **27**, 699-717.
- Gordon, P.C., Grosz, B.J., & Gilliom, L.A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, **17**, 311-347.
- Gosselin, F., & Schyns, P. G. (1997). Debunking the basic level. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the nineteenth annual conference of the Cognitive Science Society* (pp.277-282). New Jersey: Lawrence Erlbaum Associates, Publishers.
- Gosselin, F. & Schyns., P.G. (1999, submitted). A new formal model of basic-level categorization and recognition, and its testing.
- MacDonald, M.C., & MacWhinney, B. (1990). Measuring inhibition and facilitation from pronouns. *Journal of Memory and Language*, **29**, 469-492.
- McDonald, J.L., & MacWhinney, B. (1995). The time course of pronoun resolution: Effects of implicit verb causality and gender. *Journal of Memory and Language*, **34**, 543-566.
- Neath, I. (1993). Distinctiveness and serial position effects in recognition. *Memory & Cognition*, **21**, 689-698.
- Neath, I., & Knoedler, A.J. (1994). Distinctiveness and serial position effects in recognition and sentence processing. *Journal of Memory and Language*, **33**, 776-795.
- Sheldon, A.L. (1974). The role of parallel function in the acquisition of relative clauses in English. *Journal of Verbal Learning and Verbal Behaviour*, **13**, 272-281.
- Stevenson, R.J., & Vitkovitch, M. (1986). The comprehension of anaphoric relations. *Language and Speech*, **29**, 335-357.
- Stevenson, R.J., Nelson, A.W.R., & Stenning, K. (1995). The role of parallelism in strategies of pronoun comprehension. *Language and Speech*, **38**, 393-418.
- Stewart, A.J., Pickering, M.J., & Sanford, A.J. (in press). The time-course of the influence of implicit causality information: Focus versus integration accounts. *Journal of Memory and Language*.
- Tyler, L.K., & Marslen-Wilson, W. (1982). The resolution of discourse anaphors: Some on-line studies. *Text*, **2**, 263-291.

## Appendix

The gist of SLIP is both simple and intuitively appealing: a classifier with an imperfect pick-up mechanism serially cycles through one or many strategies test by test in an attempt to verify one of them. A strategy gives the procedure required to check whether an object is a member of a given category. More specifically, a strategy is a series of sets of redundant features. For instance, take example sentence (1) in Table 1 :

(1) John fascinated Mary because he was very interesting.

At the end of stage 1, hypothesis\_1 (i.e. the hypothesis according to which the first mentioned character is the pronominal referent) is made and gender is known to be diagnostic. This translates into the following strategy: S1 = [{G1, NP1}]. This is a length 1 strategy because it has only one set of redundant features. All the strategies required for pronoun resolution are of length 1 although for SLIP this does not have to be the case (see Gosselin & Schyns, 1997, 1999). For the sake of simplicity our formal discussion is confined to length 1 strategies here. The set of redundant features in S1 contains all the features which can decisively verify hypothesis\_1 in example sentence 1. Three other strategies are also used for the set of example sentences in Table 1: S2 = [{NP1}], S3 = [{G2, NP2}], and S4 = [{NP2}]. S2 is used when hypothesis\_1 is made and gender is nondiagnostic; S3 is employed when hypothesis\_2 is made and gender is diagnostic; and S4 is used when hypothesis\_2 is made and gender is nondiagnostic.

In the SLIP framework, a strategy as a whole is verified whenever all sets of redundant features have been individually verified in a specific order. A set of redundant features has been verified as soon as a one of its features has been verified. For example, S1 is verified as soon as either G1 or NP1 is verified. Given that a SLIP categoriser has a stochastic feature-pick-up mechanism, this verification habitually happens after a succession of misses. The probability of having t-1 successive misses is given by  $(P-PQ)^{(t-1)}$  where  $P$  is the probability of a random slip and  $Q$  is the probability of a diagnostic slip, i.e. the cardinality of the set of redundant features divided by the total number of features in the shown sentence. We assume in this article that 10 features are present in sentences for the verification stage: gender information (sometimes diagnostic and sometimes not), explicit cause (always diagnostic), and eight nondiagnostic features such as verb tense (this number was arbitrarily chosen, but a different one would make little difference). The probability of a hit is simply 1 minus the probability of a miss. Thus, the probability that a certain strategy will be verified after t tests is:

$$(P-PQ)^{(t-1)}[1-(P-PQ)].$$

This expression gives the Special Response Time Density Function (SRTDF) of a SLIP categoriser. It describes a geometric density function. The best fit between the data and our predictions is obtained with  $P = 1$ , meaning that features are gathered randomly.

The global measure reported in our simulations is the average number of features that have to be picked up before the categoriser reaches a decision (i.e. to verify or reject a strategy). We begin with the rejection case. If a categoriser has failed to verify a strategy after  $t_{stop}$  ( $t_{stop} = 1$ ) feature pick-ups either the strategy does not apply, or the categoriser's extraction mechanism has until then slipped onto nondiagnostic features. As  $t_{stop}$  increases the second possibility becomes less and less likely. A classifier could thus conclude quite confidently that a strategy does not apply if it has reached  $t_{stop}$  pick-ups if beyond this point the probability that the strategy applies to the pronoun is smaller than some small constant probability  $D$ . Given  $P$ ,  $Q$  and  $D$ ,  $t_{stop}$  can be calculated easily:

$$t_{stop} = \log D / \log(P-PQ).$$

This equation is known as the inverse survival function of probability  $D$ . A categoriser using this method errs with a probability of  $D$  on negative trials (i.e. it rejects the hypothesis when it is correct with a probability of  $D$ ). For the simulations  $D$  was set at .111, the subjects' mean error rate in Stewart, Pickering and Sanford (in press, Experiment 4). Note: this is not a free parameter. Consider example sentence (2).  $Q = 2/10$ . It thus takes our categoriser an average of 9.851 pick-ups before rejecting hypothesis\_1 and thus accepting the alternative hypothesis\_2.

Now that we know how to compute  $t_{stop}$ , we can calculate  $t_{mean}$ , the mean number of pick-ups required to verify positive trials (i.e. when a strategy is correct):

$$t_{mean} = \frac{\sum_{t=1}^{t_{stop}'} t \cdot SRTDF}{\sum_{t=1}^{t_{stop}'} SRTDF}$$

where  $t_{stop}'$  is simply  $t_{stop}$  rounded up to the next integer. Consider example sentence (1).  $Q = 2/10$ . We can thus use the  $t_{stop}$  calculated for example sentence (2); once rounded up it becomes 10. So,  $t_{mean}$  is equal to 3.511; it takes an average of 3.511 pick-ups for hypothesis\_1 to be accepted in this case.

## Category Induction for Ordinary Facts

Roman Taraban (r.taraban@ttu.edu)

Matt Hayes

Department of Psychology  
Texas Tech University  
Lubbock, TX 79409-2051

### Abstract

Typically, research on category learning has examined the acquisition of correct responses for explicitly identified categories. A connectionist model developed by McClelland (1981) used an interconnected network of factual elements to show that it was possible for a network to correctly infer connections between knowledge representations that were not explicitly coded into the network. Two experiments were conducted with adults using facts from the McClelland model. Clustering related facts, presenting the full set of transfer probes, and providing intermittent feedback during learning, did not reliably amplify the induction of implicit categories that was necessary for the transfer of learning tasks. The data in both experiments revealed a wide range of individual differences suggestive of graded levels of category induction. A series of simulations using backpropagation with recurrent connections showed that individual differences could be accounted for by manipulating feedback connections, the number of hidden units, and their connectivity. The discussion considers the relation of these findings to related research involving correlated features.

The notion of similarity has been very compelling in explanations of category acquisition. Intuitively it makes sense that we group things together because they are similar to each other. The details of how similarity should be computed have changed as theories have replaced one another over time (Taraban, 1993). However, one essential idea has remained, that category acquisition is driven by the identification or weighting of features that signal membership in one category or another. What is important to note for purposes of the present paper is that these theories of the relation of features to categorical distinctions have largely considered cases in which the features are all present and immediately available in instances of the object, and the possible classifications are made explicit to the learner (e.g., Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994). Typically, experiments have involved *supervised* learning, in which the category labels are part of what a participant learns. However, category learning is sometimes *unsupervised*, as when children learn to use linguistic elements that have an underlying categorical structure in their native language without ever labeling those categories (MacWhinney, Leinbach, Taraban, & McDonald, 1989), when they learn unlabeled categories in artificial languages (Billman, 1989; Brooks, Braine, Catalano, Brody & Sudhalter, 1993; Frigo & McDonald, 1998), or when they learn to classify events

(Kersten & Billman, 1997). In these cases, properties that are correlated form the basis for categories. The categories remain unlabeled and outside of direct instruction, but still influence individuals' classifications of novel instances.

Ordinary communication often carries correlated information, but it presents special difficulties because information is dispersed over time due to the serial and temporal nature of speech and print. A body of correlated information may be communicated, but the individual must construct or induce these correlations against temporal constraints. The question of how categories are formed and processed when the characteristic or defining features of the category are not simultaneously available has received very little attention. One of the few explicit models for the dynamic induction of categories based on co-occurring properties was proposed by McClelland (1981; see also McClelland & Rumelhart, 1988), who proposed a mechanism that could form generalizations from stored representations. The model was unique in that the "probe" or initiating information did not provide all the relevant cues simultaneously (e.g., *large, white, triangle*, as in Nosofsky et al., 1994). Instead, categories of co-occurring properties could be induced from encoded facts through a process of spreading activation and inhibition along connected pathways in a connectionist network. The network was capable of filling in missing information about an individual who was represented in the network (e.g., a gang member called *Lance*) due to co-occurring properties between that individual and other individuals represented in the network. The network could also provide the most likely features associated with groups (e.g., a gang called the *Jets*) based on properties that generally co-occurred with that group. The model is plausible in the sense that ordinary circumstances often expose us to disparate facts, and it has important implications in that it suggests we may still induce useful categorical generalizations based on co-occurrence patterns within those facts.

### Experiments 1 and 2

In Experiments 1 and 2, participants learned facts about individuals. Part of the learning was supervised, consisting of a name (e.g., *Lance*) and a category (e.g., *education*), with feedback indicating the correct response (*high school*). These facts are labeled the *base learning* items in Table 1. The responses to base learning items clustered into two im-

## Category Induction

PLICIT categories (i.e., all the individuals sharing the properties *Jets*, *high school*, and *drug dealer*, and all the individuals sharing the properties *Sharks*, *junior high*, and *car thief*). If participants induced these connections, it was through unsupervised learning, because those connections were never directly brought to participants' attention. Further, induction of these implicit categories drew heavily on memory, as the correlated properties never appeared together on a learning trial. Participants' induction of the implicit categories was tested using two types of transfer questions, *base transfer* and *novel transfer* items. Base transfer items asked participants about individuals' properties that were not among the base learning items. For instance, participants learned about Lance's gang and education, but not his occupation. For novel transfer items participants were given a "hint" about a totally new person (e.g., *Moe has a junior high education*) and were asked about the remaining properties (his *gang* and *occupation*).

Consonant with connectionist learning principles, performance on the transfer items would depend on weighted interconnections between elements. Relatedly, the amount of transfer would not be all-or-none but could occur in varying degrees. The discovery of individual differences would be consistent with learning in connectionist models, which occurs incrementally at different learning rates.

Category features that are activated together become associated (Billman, 1989; Kersten & Billman, 1997). As has already been described, learning trials in Experiments 1 and 2 presented a single property. Experiment 1 examined whether learning the underlying pattern of connections between properties could be facilitated by presenting related properties on contiguous trials. In Experiment 2, two additional manipulations were tested in an attempt to amplify category induction. One involved providing intermittent feedback during learning. The other involved presenting transfer probes during learning, without feedback. Both were meant to encourage participants to think about the connections between the properties they were learning.

## Method

**Participants** The 40 participants in Experiment 1 and the 60 participants in Experiment 2 were recruited from introductory psychology courses at Texas Tech University and participated for course credit.

**Materials** The learning and transfer materials were based on the table of information in McClelland (1981), although considerably reduced. Each of 16 individuals was described along three binary-valued dimensions: his gang membership, his education level, and his occupation. The assignment of dimension values to individuals was fully reliable in the following sense. If an individual belonged to the Jets, then that individual was also a drug dealer with high school education. Likewise, a member of the Sharks always had a junior high education and made his living as a car thief. Forty probes and eight hints were constructed from the matrix of information in Table 1. Each probe consisted of two parts: a name and the category of information about the named person. Example probes were *Art's gang*, *Art's education*, *Art's occupation*.

**Procedure** In Experiment 1, participants were randomly assigned to one of two conditions. In both conditions, learning trials were presented one at a time in blocks consisting of 18 probes. In the *unclustered* condition, the 18 base-learning probes were randomized. In the *clustered* condition, probes were also presented in random order with the exception that all probes about a particular person were presented in sequential trials (e.g., trial<sub>n</sub>: *Lance's education*; trial<sub>n+1</sub>: *Lance's gang*).

In Experiment 2, participants were randomly assigned to one of three conditions. The conditions differed in the number of base items that appeared in the learning sets and in the amount of feedback provided to participants during learning. The *control* condition was identical to the unclustered condition in Experiment 1. Participants learned the 18 base learning items shown in Table 1. The *intermittent-control* condition was identical to the control condition, except that feedback consisting of the correct response was provided at random two thirds of the time. The *intermittent-base* condition was identical to the intermittent-control condition, except that the six base transfer probes were mixed in randomly in each block of learning trials. Participants never received feedback on the base transfer items. The final test was identical in all three conditions.

Participants worked individually at a computer. One or two meetings were provided for learning the base items to

Table 1: Base Learning, Base Transfer, and Novel Transfer Items

Base Learning and Base Transfer Items				Novel Transfer Items			
Name	Gang	Education	Occupation	Name	Gang	Education	Occupation
Art	Jets	high school	drug dealer	Chuck	*	~ high school	*
Lance	Jets	high school	*	Jake	~ Jets	*	*
Greg	*	high school	drug dealer	Zane	*	~ high school	*
Pete	Jets	*	drug dealer	Ed	~ Jets	*	*
Nick	Sharks	junior high	car thief	Moe	*	~ junior high	*
Earl	Sharks	junior high	*	Gene	*	*	~ car thief
Karl	*	junior high	car thief	Vick	~ Sharks	*	*
Bill	Sharks	*	car thief	Ron	~ Sharks	*	*

\* Base and novel transfer items are marked with an asterisk. ~ The hints used for novel items are marked with a tilde.



criterion and one meeting was provided for taking the final test. Learning was self-paced. A learning trial consisted of a screen displaying a probe about one of the individuals in the experiment. The participant typed in a response. Feedback (when provided) indicated whether the response was correct or incorrect, and the correct response was displayed on the screen. Participants initiated the next trial by pressing a key on the keyboard. When participants reached the criterion of 17 or 18 correct, they were dismissed until the next day when the final test was administered.

Test trials were identical to learning trials in the way probes were presented and responses were made, except that participants did not receive feedback. Test trials were presented in two sets. The first set consisted of the presentation of the full set of 24 base-learning and base-transfer probes in random order. Immediately after responding to the 24 probes, participants were instructed by the computer that they would be presented with new items, that for each trial they would be given a “hint,” and that they should give their best response. For all trials, participants’ responses and accuracy were automatically recorded by the computer.

### Results for Experiment 1

Participants took an average of 215 trials to reach criterion in the unclustered condition and 189 trials in the clustered condition. Mean accuracy for all the learning trials in the unclustered condition was 63.6% and in the clustered condition 66.7%. Although these means favored mastering the base learning items in the clustered condition, an analysis of variance using number of trials as the dependent variable failed to show that the effect of condition was significant [ $F(1, 38) = 0.58, MSE = 11769, ns$ ], and an analysis using percent correct as a dependent measure failed to show that accuracy rates were significantly different [ $F(1, 38) = 1.60, MSE = 59.26, ns$ ].

Table 2: Mean Percent Accuracy for Final Test

Item Type	Experiment 1		Experiment 2		
	UN	CL	CO	IC	IB
Base Learning	83	84	85	83	82
Base Transfer	57	50	55	67	52
Novel Transfer	59	67	63	67	55

Note. UN: unclustered; CL: clustered; CO: control; IC: intermittent control; IB: intermittent base

Table 2 summarizes the final test data. Participants’ accuracy was high on base learning trials (84%), lower on novel transfer trials (63%), and lowest on base transfer trials (53%). A 2 (Condition: clustered, unclustered) X 3 (Item Type: base learning, base transfer, novel transfer) analysis of variance showed a main effect for item type [ $F(2, 76) = 23.85, MSE = 409.19, p < .001$ ] but not for condition [ $F(1, 38) = 0.03, MSE = 628.92, ns$ ] nor interaction of the two factors [ $F(2, 76) = 1.31, MSE = 409.19, ns$ ]. Tukey HSD (alpha = .05) tests showed that base learning accuracy differed significantly from base transfer and novel transfer ac-

curacy; base transfer and novel transfer did not differ from one another.

The possibility of individual differences in these data warranted a closer examination of individual test outcomes. Indeed, the rationale for a *gamma* parameter in the McClelland (1981) model was to allow for individual differences in induction. The incremental nature of connectionist learning is also suggestive of individual variation. Participants did quite well on the base learning items, but differed noticeably in their performance on the transfer items (See Appendix). Assuming a binomial distribution ( $n = 16, p = .50$ ) of individual responses for the novel transfer items, 35% ( $n = 7$ ) of the participants in the unclustered condition and 50% ( $n = 10$ ) in the clustered condition had accuracy rates that were not likely to be due to chance (accuracy  $\geq 75\%, p < .03$ ). For base transfer items, 10% ( $n = 2$ ) of the participants in the unclustered condition and 5% ( $n = 1$ ) in the clustered condition scored better than chance (binomial:  $n = 6, p = .50$ ; accuracy = 100%,  $p < .02$ ) (These three participants also scored better than chance on the novel transfer items).

### Results for Experiment 2

Participants took an average of 250 trials to reach criterion in the control condition, 284 trials in the intermittent-control condition, and 222 trials in the intermittent-base condition. These differences were not reliable [ $F(2, 57) = 1.10, MSE = 17756.41, ns$ ]. The mean percent correct was 61.4 in the control condition, 63.7 in the intermittent-control condition, and 58.4 in the intermittent-base condition. These differences were significant [ $F(2, 57) = 4.26, MSE = 33.87, p < .02$ ]. Tukey HSD tests showed that mean accuracy in the intermittent-control condition was significantly higher than in the intermittent-base condition. The control and intermittent control conditions were not significantly different. The small advantage for the intermittent-control condition was due in part to the additional blocks of trials these participants needed to reach criterion. These late trials tended to be error free.

The test data are summarized in Table 2. A 3 (Condition: control, intermittent-control, intermittent-base) X 3 (Item Type: base learning, base transfer, novel transfer) analysis of variance showed a main effect for item type [ $F(2, 114) = 31.93, MSE = 362.67, p < .001$ ]. Tukey HSD tests showed that base learning accuracy differed significantly from base transfer and novel transfer accuracy; base transfer and novel transfer did not differ from one another. The effect of condition was not significant [ $F(2, 57) = 1.56, MSE = 834.38, ns$ ], nor was the Condition X Item Type interaction [ $F(4, 114) = 0.98, MSE = 362.67, ns$ ].

A closer look at individual performance was again undertaken. For the novel transfer items, 40% ( $n = 8$ ) of the participants in the control condition, 55% ( $n = 11$ ) in the intermittent-control condition, and 30% ( $n = 6$ ) in the intermittent-base condition had accuracy rates that were not likely to be due to chance (binomial:  $n = 16, p = .50$ ; accuracy  $\geq 75\%, p < .03$ ). On base transfer items, 15% ( $n = 3$ ) of the participants in the control condition, 20% ( $n = 4$ ) in the intermittent-control condition, and 5% ( $n = 1$ ) in the intermit-

tent-base condition scored better than chance (binomial:  $n = 6$ ,  $p = .50$ ; accuracy = 100%,  $p < .02$ ); all but one of these also scored better than chance on the novel transfer items.

Because the experimental manipulations in Experiments 1 and 2 failed to produce reliable differences, combining the data was warranted in order to increase statistical power. Across the 100 participants, mean accuracy on base transfer items was 56%, and on novel transfer items it was 62%. A one-sample  $t$ -test showed that base transfer performance was significantly greater than chance (50%) [ $t(99) = 2.55$ ,  $p < .02$ ]. A paired  $t$ -test showed that performance on novel transfer items was significantly higher than on base transfer items [ $t(99) = -2.06$ ,  $p < .05$ ]. Overall, participants did better on novel transfer items. Performance for both types of transfer items exceeded chance.

## Discussion

A knowledge base was organized around individuals, like Art and Lance, about whom participants learned properties (e.g., *Lance's gang is Jets*; *Lance's education is high school*). In order to transfer knowledge of these facts to new instances, participants had to use the learned facts to induce the categorical relations between them. There is no specific way in which this had to be done. A person might notice that *Jets* and *high school* always co-occur without noting the link with *drug dealer*. A person might induce some other connections or the complete set of connections. An examination of individual performances (see Appendix) suggests large discrepancies in induction. Some participants appeared to perform at chance on the transfer trials, others scored perfectly, and yet others were in between. These data suggest that individuals can evoke the appropriate categories even if these were not explicitly taught. As the McClelland (1981) model suggests, it is not necessary to develop explicit connections in order to exploit the existing connections in useful ways. The data also support the supposition that the level of induction is graded. This is consistent with connectionist models, which do not encode rules or *If-Then* productions, but which develop interconnections and internal representations (on a hidden layer) incrementally, or alternatively, which control the spread of activation and inhibition parametrically within a storage and retrieval mechanism like McClelland's (1981).

## Connectionist Simulations

In a preliminary set of simulations, the base learning items in Table 1 were interconnected as described in McClelland (1981) and McClelland and Rumelhart (1988). There were no direct connections for base transfer items (e.g., a link from *Lance* to *drug dealer*). For novel transfer items, the "hint" (e.g., *Jets*) was activated. The model was tested at three settings of  $\gamma$ , a parameter that controls the level of inhibition between units in the same pool of units in the network, and thereby changes the level of generalization (McClelland, 1981). With  $\gamma$  set to .1000, .1249, or .1500, the mean probability of a correct response to base learning items was .98. Probabilities for base transfer items were .97, .67, .50, respectively. For all three  $\gamma$  values

the probability of a correct response to novel transfer items was .98.<sup>1</sup> The first two findings are consistent with the findings in Experiments 1 and 2, that is, a range of individual differences on base transfer items when performance was high on base learning items. The third outcome of the simulation, uniformly high performance on novel transfer items, did not fit the data, which showed a wide range of individual differences on these items. It is generally impossible to salvage the McClelland (1981) model. When given a hint like *Jets*, the network needs to find only one member with that feature and will generalize from that member. The highest levels of inhibition leave at least one member to generalize from. Another shortcoming of the McClelland model for present purposes is that it uses fixed weights and thus does not account for individual differences based on learning.

The binomial analyses of individual performance presented earlier suggested that there were three major patterns of behavior (See the Appendix for representative data). The most predominant pattern, characteristic of 54% of the 100 participants, was above chance performance on the base learning items and chance performance on base transfer and novel transfer items, which will be labeled the HLL pattern. The next most predominant pattern, representing 29% of the participants, was above chance performance on the base learning, chance performance on the base transfer items, and above chance performance on novel transfer items (the HLH pattern). Ten percent of the participants achieved above chance performance on all items (the HHH pattern). Only one participant had a HHL pattern. Three participants had LLH patterns and three had LLL patterns. The connectionist solution presented next required the simultaneous manipulation of multiple connectionist factors, including network architecture, hidden unit resources, connectivity, and internal feedback. The model replicated all the patterns above except the LLL and the LLH patterns. Adding a decay factor would be necessary to account for human participants who reached the criterion on the day prior to the test but who scored at chance (L - -) on base learning items on the day of the test.

The model is depicted in Figure 1. The inputs, corresponding to probes, activated exemplars (*local* representations). Each input (e.g., *Art*) was linked to a single exemplar unit (e.g., *Art*). All other units between pools were fully interconnected. The output units had recurrent connections back to a hidden layer of units (*distributed* representation). The hidden layer also received connections from the exemplar units. A learning trial occurred in two

<sup>1</sup> All connections in the network were set to 1. Default values were used for the parameters  $\alpha$ ,  $\text{decay}$ , and  $\text{estr}$ . The remaining parameters were set to  $\text{max} = 1.10$ ,  $\text{min} = .01$ ,  $\text{rest} = .01$  for ease of interpreting the output. The probabilities were based on an application of the Luce (1959) choice rule to the activations from the relevant pool of units; e.g.,  $P(\text{Jets}) = \text{activation}(\text{Jets}) / \text{activation}(\text{Jets}) + \text{activation}(\text{Sharks})$ .

Table 3: Mean Probabilities (X100) for the Connectionist Simulations

Item Type	Feedback Connections Only										Feedback and Exemplar Connections										
	Number of Hidden Units										Number of Hidden Units										
	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	
Base Learning	90	92	91	91	91	91	91	91	91	91	90	94	96	97	97	97	97	97	97	97	97
Base Transfer	51	53	44	49	48	49	52	52	51	51	53	73	88	92	92	89	94	93	91	91	91
Novel Transfer	50	55	78	88	82	81	52	48	51	49	51	54	89	95	93	94	51	50	55	52	52

passes. During the first pass, the output units were activated by the inputs via the exemplar and hidden units. During the second pass, activation was passed back to the hidden units via the recurrent connections and only these units fed activation forward to the output units. Error on each pass was calculated and weight adjustment took place through the application of the backpropagation learning rule (Rumelhart, Hinton, & Williams, 1986) at the end of each epoch of training. An epoch of training consisted of one exposure to each base learning item.

For base learning and base transfer test trials, a name (e.g., *Art*) and category (e.g., *occupation*) were presented on the input layer, and the probability of a correct response was computed using the Luce (1959) choice rule for the relevant activations (see footnote 1). For novel transfer items, the “hint” (e.g., *gang*, *Jet*) was activated on the output layer, which functioned as the “teacher,” and fed back to the hidden units; the probe category (e.g., *occupation*) was activated on the input layer. Activation was fed forward to the output layer and probabilities were computed as described above. On test trials, there was a single pass through the network and no weight adjustments.

Activation and weight adjustment roughly corresponded to a trial in the human experiments. An explicit probe evoked an output, feedback was provided and all the weights were adjusted in order to improve performance on the base learning items. During the second pass, there was an implicit recirculation of the outputs through the hidden layer (cf., McClelland, McNaughton, & O’Reilly, 1995, for a discussion of consolidation in memory) and related weight adjustment. The first pass in learning corresponded to elements that could be observed in the experimental manipulation (e.g., probes, response, feedback). The second pass corresponded to unobservable processes for which some justification will be provided in the course of describing the simulation manipulations and results.

Each simulation outcome presented in Table 3 is the mean of 10 independent simulations. All the parameters in the simulations were fixed, except whether or not the exemplar units were connected to the hidden units, and the number of hidden units, which varied from one to ten. The learning rate for all trials was .01, momentum was .90, and each simulation consisted of 4000 epochs of training. Eleven input units were used to code base learning trials. One input was allocated to each of eight gang member names and one to each category of information (i.e., gang, education, occupation). Thus, the input layer mimicked a participant who was asked *Art’s gang*, for instance. Each of these input units connected to a single exemplar unit. These internal exemplar representations were necessary to guarantee high learning of the base learning items across all manipulations of the hid-

den units. The exemplar units connected fully to all the output units. The output units coded the same elements as the input layer and, additionally, the responses to the probes (e.g., *Jets*, *drug dealer*). These feedforward connections from input to exemplar units to output units were all that was required to learn base learning items.

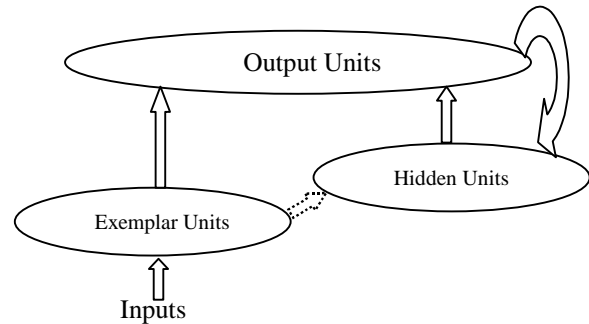


Figure 1: The Simulation Model

Feedback connections from all the output units were fully connected to the hidden units. A practical benefit of the feedback connections is that they allowed the “hint” that was used for the novel transfer trials to originate from the “teacher” (output) units, where the responses were coded. Feedback connections like these, or an equivalent, were necessary to cross reference the probes and responses in the experimental manipulation. These feedback connections were primarily responsible for the gradations of performance, up to 88% correct, on novel transfer items depending on the number of hidden units (see Table 3). By varying the number of hidden units, base learning items were high (90%-92%) and base transfer items were near chance (48%-53%). In half of the simulations, the exemplar units also connected to the hidden units. These additional connections, along with the feedback connections from output units, allowed for gradations of performance on base transfer (53%-94%) and novel transfer items (50%-95%), depending on the number of hidden units. Near perfect performance on all three types of items occurred with four hidden units: base learning (97%), base transfer (92%), and novel transfer (95%).

The simulations were not intended to account for all the individual differences in the human data, but rather to uncover network organizations that produced variations comparable to those observed in the human performance. Doing well on base transfer items requires connections between the exemplar units and the hidden units. One account of participants’ better performance on novel versus base transfer items is that connections from exemplar units to hidden units were formed less readily than feedback (recurrent) connec-

tions. The pattern in Table 3 for the transfer trials also suggests that too few or too many hidden units is not ideal. This suggests that the hidden units control the dimensionality of the solution (cf., Hinton, 1992). The dimensionality of the solution determines how much induction takes place (cf., Landauer & Dumais, 1997).

### General Discussion

The experiments and simulations presented here were inspired by McClelland's (1981) connectionist model that was able to infer connections between stored "facts," even though the network was not explicitly trained to make those connections. Correctly making these inferences depends on uncovering the correlational structure between the facts. Further research is necessary to confirm that differences in the conditions of learning and retrieval are crucial to explaining the strong individual differences found in the human performance here and elsewhere. Two bodies of research currently suggest somewhat different conclusions on these points. Billman (1989) and Kersten and Billman (1997) contrasted learning and generalization for stimuli with many correlated features to those with only few correlated features and found that participants readily generalized from the former but not the latter. The features of the facts in Experiments 1 and 2 were also highly correlated, but participants did not readily form generalizations from them. The crucial difference, which remains to be tested more fully, is that Billman and Kersten explicitly displayed the correlations as part of their learning phase, whereas the connections in our experiments were implicit across learning trials. For instance, *Jets* and *drug dealer* are perfectly correlated, but participants never viewed those two features together. The present manipulations are more comparable to the "control" language in the artificial language experiments of Brooks et al. (1993). Their control language also had a rich correlational structure comparable to the ways in which noun gender in languages like Russian affects morphological and inflectional differences. In spite of the underlying correlations, participants found the control language difficult to learn, we surmise, because the correlational structure was implicit across trials as in the present study. However, the generally low performance for the control language presumably included a range of individual differences, as presented here, meaning that some individuals discover the correlational structure in spite of its dispersion over trials. Therefore, we believe that the present research begins to bridge several lines of existing research, that it uncovers the broad individual differences in performance, and offers a preliminary connectionist explanation for those differences. The counterpart to the "experimental" language in Brooks et al. and "high systematicity" in Billman needs to be tested for the current stimuli and integrated into the present connectionist architecture.

### Appendix

Each triple, separated by semicolons, is mean percent accuracy for each participant, for base learning, base transfer, and novel transfer items, respectively.

**Experiment 1, Unclustered** 100, 83, 69; 100, 100, 94; 94, 67, 94; 94, 50, 88; 94, 33, 50; 89, 50, 25; 89, 100, 100; 89, 50, 31; 89, 83, 44; 89, 67, 44; 89, 50, 25; 83, 33, 100; 83, 50, 69; 78, 17, 25; 78, 67, 25; 72, 33, 76; 72, 67, 25; 72, 50, 47; 61, 33, 88; 50, 50, 63 **Clustered** 100, 100, 100; 94, 67, 31; 94, 67, 38; 94, 33, 41; 94, 33, 18; 89, 33, 56; 89, 50, 18; 89, 33, 44; 89, 67, 94; 89, 67, 100; 83, 50, 94; 83, 17, 44; 78, 17, 100; 78, 67, 50; 78, 67, 100; 78, 50, 81; 78, 50, 100; 72, 50, 94; 72, 33, 50; 67, 50, 88.

### References

Billman, D. (1989). Systems of correlations in rule and category learning: Use of structured input in learning syntactic categories. *Language and Cognitive Processes, 4*, 127-155.

Brooks, P., Braine, M., Catalano, L., & Brody, R. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language, 32*, 76-95.

Frigo, L., & McDonald, J. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language, 39*, 218-245.

Hinton, G. (1992). How neural networks learn from experience. *Scientific American, 145*-151.

Kersten, A., & Billman, D. (1997). Event category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 638-658.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Luce, R. (1959). *Individual choice behavior*. New York: Wiley.

MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules. *Journal of Memory and Language, 28*, 255-277.

McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of specifics. *Proceedings of the Third Annual Meeting of the Cognitive Science Society*, 170-172.

McClelland, J. L., McNaughton, B., & O'Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*, 419-457.

McClelland, J. L., & Rumelhart, D. (1988). *Explorations in parallel distributed processing*. Cambridge, MA: MIT Press.

Nosofsky, R., Gluck, M., Palmeri, T., McKinley, S., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22*, 352-369.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In D. Rumelhart, J. McClelland, & the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Taraban, R. (1993). Introduction: A coupling of disciplines in categorization research. In G. Nakamura, R. Taraban, and D. Medin (Eds.), *The psychology of learning and motivation, Vol. 29: Categorization by humans and machines*. San Diego, CA: Academic Press.

# Learning and Generalizing New Concepts

Jean-Pierre Thibaut (jpthibaut@ulg.ac.be)

Department of Psychology

Université de Liège

Boulevard du Rectorat, 5

4000 Liège BELGIUM

## Abstract

When subjects learn to categorize new stimuli adequately, they have to segment these stimuli into relevant features for categorization. In the experiments reported here, children had to discover a rule for categorization. Preliminary experiments have shown that depending on the nature of the irrelevant features, children could find the relevant features from age four or could not find them before the age of eleven or twelve. A central question is whether children aged four or six who have discovered the rule in a simplified version of the relevant features would generalize to a "complex" version (i.e., in which there is more background noise) of the relevant features, i.e., a version that they would be unable to learn before twelve without pre-training. Conditions promoting the generalization from the simple version to the complex version were also investigated. Two conditions were compared: relearning with or without feedback. Results showed that children aged 4 and 6 could generalize the "simple" version of the target concept to a more complex version of the same concept, either with and without feedback in the generalization phase.

## Introduction

Children have to learn to categorize stimuli according to adults' standards. In order to achieve this correctly, they have to find the relevant features for categorization. If the particular task is to learn to categorize a set of new stimuli into two new categories, they will have to find the features that characterize stimuli of each category and that distinguish them from stimuli of the other category. Imagine a traditional concept learning experiment in which participants have to discover one relevant feature that allows for perfect categorization. Stimuli are constituted of a number of dimensions, either relevant or irrelevant. Subjects are presumed to formulate and test simple hypotheses concerning the rule that define membership (Nosofsky, Palmeri, & McKinley, 1994). This means that participants will analyze stimuli into their dimensions and test whether each dimension partitions the set of stimuli. A number of characteristics of the stimuli contribute to the task difficulty. The salience of dimensions: a non salient relevant dimension among salient irrelevant dimensions presumably requires more systematic analyses of the stimuli than a salient relevant dimension among non salient irrelevant dimensions.

Variability in the perceptual manifestation of a relevant feature can hinder this relevant feature and impede its discovery. For example, compare Figure 1A stimuli with

Figure 1C stimuli which define two experimental conditions. In the two conditions, the stimuli come from two categories defined by the same relevant features. Each stimulus has four "legs", with one category being defined as "1 isolated leg and 3 connected legs" (1+3), the other category being defined as "two sets of two connected legs" (2+2). In Figure 1C the length, shape, size of the legs were made more variable than in Figure 1A. Preliminary results obtained by Thibaut (1999) indicate that the rule (1+3 vs. 2+2) could be discovered from the age of four in the case of Figure 1A stimuli whereas children under thirteen could not find the equivalent rule for Figure 1C stimuli. Figure 1B stimuli elicited intermediary results: most children aged ten discovered the rule.

Thibaut (1999) suggested that young children had problems either in screening the stimuli, or inhibiting irrelevant features, or plan systematic comparisons between stimuli. The purpose of the present contribution is to assess to what extent young children (four- or six-year olds) who discovered the relevant features for categorization 1+3 vs. 2+2 in the simplified version (Figure 1A) would be able to generalize to more complex versions of the same features (Figure 1B and 1C). In other words, once he/she has learned to apply a classification rule in a low variability context (such as Figure 1A stimuli), is a child able to apply it in a high variability context ?

It has been emphasized in the developmental literature that there are differences between adults' and children's in processing abilities. According to Kemler (1989), children are more holistic processors than adults. She suggested that holistic processors would run into more difficulties when only one of many attributes is relevant for categorization than when categories are defined by overall similarity relationships, i.e., when stimuli share many characteristic features. Other authors consider that property-specific information is accessible to young children, even those aged 4 or 5 years. This means that children can analyze stimuli in terms of their constituent features, even if they do not analyze the stimuli in the same way older children and adults do. Ward (1989), Ward and Scott (1987) have argued that the difference between young learners and older learners is that younger learners may have rigid attribute preferences.

Following the holistic view, one can hypothesize that if young children perceive stimuli holistically, they should be unable to analyze the complex stimuli into their constituents and, thus, should also be unable to isolate specific aspects of the legs in order to generalize the simple version of the rule to the complex version. In the same

way, if young children have rigid attribute preferences it might be that, when confronted to the complex stimuli, they will focus their attention on the salient irrelevant properties and be unable to analyze the legs in terms of less salient properties.

Studies on generalization generally take a different perspective from the one followed here. Usually, children first learn a given concept, then they are presented with a set of new stimuli, the purpose being to analyze to which among these new stimuli they generalize the concept. Here the issue is to analyze to what extent children who discovered a rule for categorization in a simplified context will be able to generalize it to more complex objects for which they would be unable to discover the rule if they had to discover it without being first presented with the simple version. This is important because a positive answer would mean that an appropriate learning sequence can lead to an understanding of concepts which, otherwise, would remain out of the conceptual world of the child. Two generalization conditions will be compared. In the first one, children will be given feedback when they will learn to apply the simple rule to the complex stimuli. In the second condition, there will be no such feedback. It is believed that feedback will promote the understanding of the equivalence between the known simple version of the rule and its complex version. This is because, if young children do not perceive this equivalence at first glance, they can test different translations of the simple rule in terms of the complex rule and get feedback at each trial. In the no feedback condition, successive trials do not bring any information about children's successive hypotheses. If a child does not find the correct way to generalize the simple version of the rule after a limited number of trials, the absence of feedback increases the probability that his/her attention will be caught by salient irrelevant features.

### Experimental Design

Preliminary results (Thibaut, 1997) have shown that children under thirteen could not parse Figure 1C stimuli adequately. In the same way, most of children under eight could not find the relevant feature for categorization in the stimuli displayed on Figure 1B. On the other hand, the majority of children aged four could find the relevant features 1+3 and 2+2 in stimuli such as the ones displayed in Figure 1A. The purpose of the experiment was to assess whether children aged four and six who are able to find the relevant features for categorization for the simple stimuli (Figure 1A) would be able to generalize them to the stimuli displayed in Figures 1B or 1C.

The design of the experiment is summarized in Table 1.

### Methods

**Participants.** Fourteen 6-6.11-year-olds participated in the complex transfer items with feedback condition, eleven 6-

6.11-year-olds participated in the complex transfer items with NO-feedback condition, fourteen 6-6.11-year-olds participated in the semi-complex transfer items with feedback condition, fifteen 6-6.11-year-olds participated in the semi-complex transfer items with NO feedback condition, eleven 4-4.11-year-olds participated in the complex transfer items with feedback condition, twelve 4-4.11-year-olds participated in the semi-complex transfer items with feedback condition and nine 4-4.1-year-olds participated in the complex transfer items with NO feedback condition. All children were tested individually.

Table 1 : design of the experiment.

Age	Aged 4	Aged 6
Conditions		
Training condition and transfer with complex stimuli, NO feedback	x	
Training condition and transfer with complex stimuli, with feedback		
Training condition and transfer with semi-complex stimuli, with feedback		
Training condition and transfer with semi-complex stimuli, NO feedback		

Note. Cell marked "x" was not run.

**Materials.** The two categories (1+3 and 2+2) of eight stimuli were the ones used by Thibaut (1997). The learning stimuli (simple version) are presented on Figure 1A. The 16 stimuli were composed of four legs which were thin and vertical. There were eight 1-3 stimuli and eight 2-2. In this condition, the purpose was to remove salient irrelevant features for categorization. There were two sets of transfer stimuli, complex and semi-complex. The complex transfer stimuli were outlines of unknown shapes composed of two parts, the upper part (the body) and the lower part (four legs). The two categories had the same structure. In five out of the eight stimuli, the body had a mushroom-like shape that was slightly distorted over the stimuli in the case

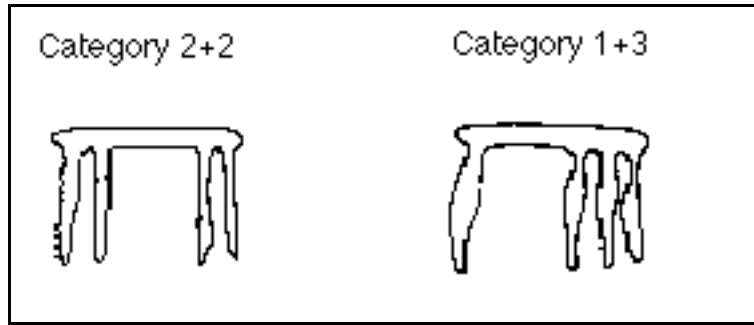


Figure 1A: two "simple" stimuli used in the training phase.

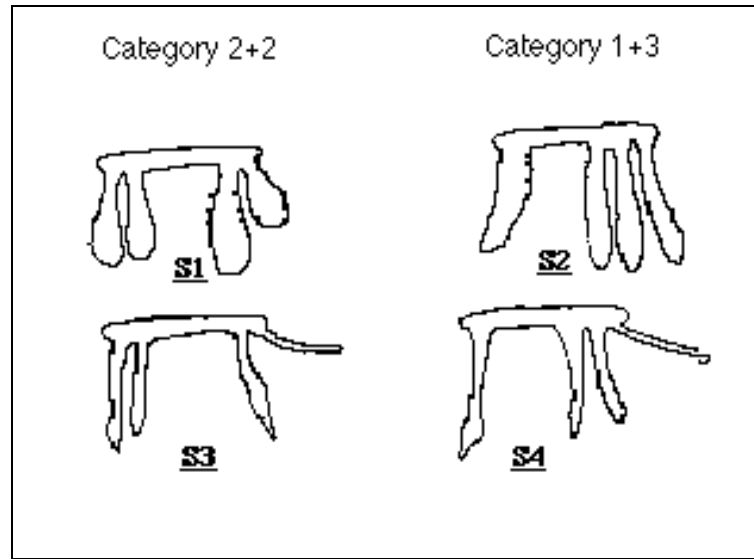


Figure 1B. Four semi-complex stimuli. Both categories (2+2 and 1+3) contain an equivalent proportion of thin and large stimuli.

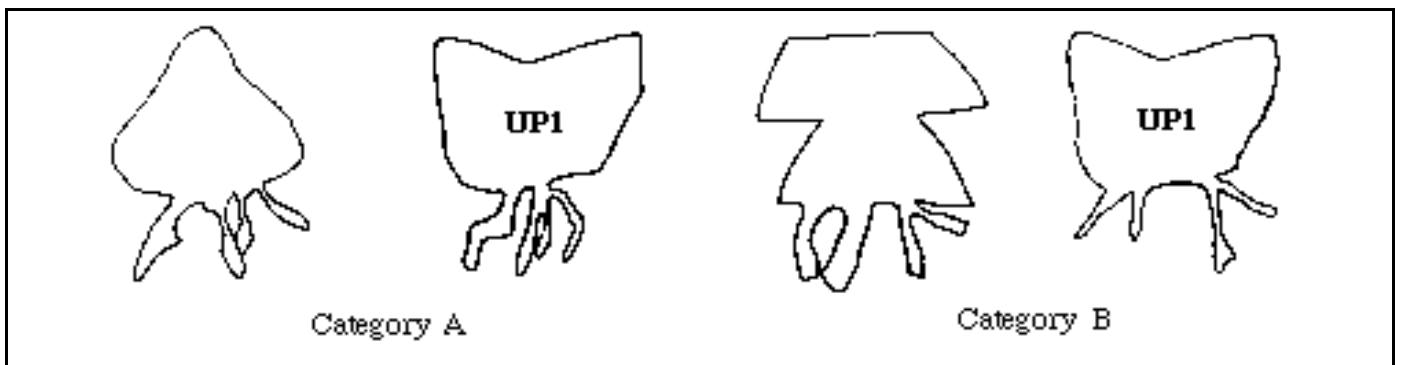


Figure 1C. Four complex stimuli from categories 1+3 and 2+2. The first stimulus has the body (upper part) characteristic of category 1+3 and the third stimulus has the body characteristic of category 2+2. The UP1 stimuli are neutral stimuli.

of category 1+3, and an angular shape in the case of category 2+2 stimuli. These two shapes were selected for their distinctiveness and perceptual saliency. The three remaining stimuli from each of the two categories were constructed with three different bodies (UP1, UP2, UP3). Since UP1, UP2, and UP3 were present in both categories

they could not be considered as cues for categorization (see Figure 1C). For each stimulus, the lower part consisted of four legs which were spatially grouped either as one leg on the left and three legs on the right in category 1+3, or two pairs of legs in category 2+2 (see Figure 1C). These

distinctive features (1-3 vs. 2-2) were the only ones available in order to categorize all the stimuli correctly.

For the semi-complex transfer stimuli, a set of 16 stimuli was constructed. The irrelevant cues "thin" "vertical", "the rightmost leg pointing to the right", and "large" were crossed with the cues "one leg plus three legs" (1+3) and "two pairs of legs" (2+2) according to four types of stimuli. There were four 1-3 stimuli and four 2-2 stimuli with "thin" legs and "the rightmost leg pointing to the right", and four 1-3 stimuli and four 2-2 stimuli composed of "broad and vertical legs" (see Figure 1B for examples of the 4 types of stimuli).

### Procedure

**Familiarization phase.** The entire set of training stimuli (Figure 1A) was presented once to the subject. Each stimulus was shown for five seconds. Then, it was removed and followed by a new stimulus. There was no feedback during this phase, and when it was over, participants were then told that they would have to learn to sort the stimuli into two categories, the name of which was provided, "bollo" for the 1-3 category, "tipi" for the 2-2 stimuli.

**Learning phase.** A first stimulus (simple version, Figure 1A) was presented for approximately five seconds

and the subject had to guess its name. The experimenter gave the appropriate feedback and presented the second stimulus in the same way, followed by the other stimuli. Feedback was provided after each answer. The order of presentation of the stimuli was random. Once the entire set of stimuli had been presented to the subject, it was presented a second time. The learning phase was stopped when children made no mistake during two successive presentations of the set of stimuli or if they were still making errors after the ninth presentation of the set. Subjects were tested individually. A session lasted for 10 to 25 minutes, depending on the number of trials necessary to complete the task.

**Transfer phase.** Children who had learned the rule for categorization had to categorize the transfer stimuli. Children were told that they would have to classify new "tipi" and "bollos" different from the ones they had seen before. In the complex stimuli with feedback condition, children were presented with the complex stimuli (Figure 1C) in the same way as in the learning phase. They received a feedback after each trial. In the semi-complex with feedback condition, children were presented with the semi-complex stimuli and received a feedback after each trial. In

Table 2. Number of subjects who reached the criterion in the two age groups and the various experimental conditions: with or without training with simple stimuli and with or without feedback in the transfer phase

Condition	Four-year-olds		Six-year-olds	
	Correct	Failure	Correct	Failure
Complex stimuli (no training with simple stimuli)	0	10	0	10
Semi complex stimuli (no training with simple stimuli)	0	10	6	8
Training condition and transfer with complex stimuli and no feedback	x	x	6	5
Training condition and transfer with complex stimuli and feedback	4	7	9	5
Training condition and transfer with semi-complex stimuli and with feedback	9	3	12	2
Training condition and transfer with semi-complex stimuli and NO feedback	8	6	11	4

Note. Cells marked "x" were not run.



the complex with NO feedback condition, complex stimuli were presented, and children never received a feedback after their classification. In the semi-complex with NO feedback condition, semi-complex stimuli were presented, and children never received a feedback after their classification. In all these experiments, the learning criterion was the same as in the learning phase.

## Results and discussion

The purpose of the experiment was to assess whether children who had first learn the rule for categorization with simple stimuli would be able to generalize it to semi-complex or complex stimuli when a feedback was provided or not. Results are summarized in Table 1. Khi square comparing data obtained in the control condition (no training with simple stimuli, Thibaut, 1997) with the new data (training with simple stimuli) revealed a significant difference in the majority of cases ( $p < .05$ ). The only exception was the case of the "generalization to complex stimuli with feedback" condition with children aged four. In this condition, a majority of children failed to generalize correctly. In sum, in a majority of conditions, training with simple stimuli influenced generalization positively. This is important because it suggests that people can generalize what they have learned to new situations that would have been beyond their understanding without this pre-training. The results obtained in conditions with feedback were compared with the equivalent results in conditions with no feedback. Comparisons revealed no significant difference (Khi square,  $p > .05$ ).

A number of authors have described children's concept learning in terms of attentional capacities (capacity to focus on specific dimensions) or of sensitivity towards dimensions (see introduction). The present results indicate that one has to include other dimensions in any model of concept learning. First, provided that exemplars of a given dimension can be highly variable (compare the simple and the complex versions of the rule), the notion of a "sensitivity to a dimension" cannot be assessed independently of the variability across instances of this dimension. This means that the probability that a relevant dimension will be discovered also depends on the presence and the structure of the other dimensions (irrelevant) that compose the stimuli. Second, in order to understand whether or not a particular instance of a dimension will be discovered by children, one has to include participants' history of categorization. By history of categorization, I mean the categorizations already performed by an individual (see Schyns, Goldstone, & Thibaut, 1998; Thibaut & Schyns, 1995). The present data suggest that the history of categorization influenced positively the way children generalized the rule. To summarize, a model of categorization and generalization has to take selective sensitivity to a particular dimension into account, provided that this notion incorporates the notion of variability in the instantiation of the dimension across stimuli. It must also

incorporates the history of categorization with a particular category in order to understand whether or not children are able to generalize a given dimension to new instances of this dimension. The present data show that knowing the history of categorization, one can predict whether a set of new stimuli is learnable. Complementarily, one can predict which history of categorization is necessary to promote generalization to subsets of highly variable stimuli. This is particularly important given that, in a majority of cases, we do not encounter identical instances of the same category.

The results presented here are important because the status of the transfer stimuli is controlled *a priori* more systematically than in traditional category learning experiments. In these latter studies, participants are confronted with transfer items of which the "intrinsic complexity" is not known. Here, the stimuli complexity in terms of learnability was independently assessed before the experiment. This is important for the control of the "paths of generalization". Following the learning strategy used here, one can bypass the role of the salient irrelevant features that would mask the relevant features for categorization whereas starting with the complex stimuli would lead to the incorrect conclusion that young children are unable to abstract the rule for categorization.

## Acknowledgements

The author would like to thank Evelyne Artus, Sophie Bylyna, Nathalie Stoffe, Marie Wiart and Julia Wronowski for their help in running the experiments.

## References

- Nosofsky, R.M., Palmeri, T.J., & McKinley, S.C. (1994). Rule-Plus-Exception Model of Classification Learning. *Psychological Review*, *101*, 53-79.
- Schyns, P.G., Goldstone, R. & Thibaut, J.-P. (1998). The development of object features. *Behavioral and Brain Sciences*, *21*, 1-53.
- Smith, L. B. (1989). A model of perceptual classification in children and adults. *Psychological Review*, *96*, 125-144.
- Thibaut, J.P. (1999) The abstraction of relevant features by children and adults in the case of visual stimuli. Manuscript in preparation.
- Thibaut, J.P., & Schyns, P.G. (1995). The development of feature spaces for similarity and categorization. *Psychologica Belgica*, *35*, 167-185.
- Ward, T.B. (1989). Analytic and holistic modes of processing in category learning. In B.E. Shepp S. Ballesteros (Eds.) *Object perception: structure and process* (pp. 387-419). Hillsdale, NJ: Lawrence Erlbaum.
- Ward, T.B. & Scott, J.G. (1987). Analytic and holistic modes of learning family-resemblance concepts. *Memory & Cognition*, *15*, 42-54.

# Rules versus Statistics in Biconditional Grammar Learning: A Simulation based on Shanks et al. (1997)

**Bert Timmermans** (Bert.Timmermans@vub.ac.be)  
Dienst Persoonlijkeids- en Sociale Psychologie  
Vrije Universiteit Brussel  
Pleinlaan 2  
1050 Brussels – Belgium

**Axel Cleeremans** (Axel.Cleeremans@ulb.ac.be)  
Séminaire de Recherche en Sciences Cognitives  
Université Libre de Bruxelles,  
Avenue F.-D. Roosevelt, 50 – CP 122  
1050 Brussels – Belgium

## Abstract

A significant part of everyday learning occurs incidentally — a process typically described as implicit learning. A central issue in this and germane domains such as language acquisition is the extent to which performance depends on the acquisition and deployment of abstract rules. In an attempt to address this question, we show that the apparent use of such rules in a simple categorisation task of artificial grammar strings, as reported by Shanks, Johnstone, and Staggs (1997), can be simulated by means of a simple recurrent network, and may thus turn out not be incompatible with the acquisition of statistical regularities rooted in the processing of exemplars of the presented material.

## Introduction

Over development and learning, we acquire a considerable amount of information incidentally. Natural language offers perhaps the most striking example of such incidental learning: Infants do not need to be explained grammar rules in order to be able to communicate effectively and are presumably unaware of the fact that they are learning something at all. Adult speakers likewise “know” whether expressions of their native language are grammatically correct but can seldom explain why.

## Implicit Learning

The notion of “*implicit learning*” (IL) usually designates cases in which a person learns about the structure of a fairly complex stimulus environment, without necessarily intending to do so, and in such a way that the resulting knowledge is difficult to express (Berry and Dienes, 1993). In short, IL is the ability to learn without awareness (Cleeremans, Destrebecqz, and Boyer, 1998), as opposed to explicit learning, which is strategy- and/or hypothesis-driven, and of which one tends to be consciously aware.

IL can produce *implicit knowledge*. According to Cleeremans (1997), “at a given time, knowledge is implicit when it can influence processing without possessing in and of itself the properties that would enable it to be an object of representation, and implicit learning is the process by which we acquire such knowledge.” (p.199) As for the notion of “representation”, we agree with Perruchet and Vinter (submitted), who state that a representation has to *represent* an entity in the real world and has to be in and of itself manipulable *as* that entity (Perruchet and Vinter talk about

its “function within a causal system”). Therefore, an entity that is an *object of representation* has to exist independently from the “hardware” of the system by which it is represented, making it available for information-processing operations in a variety of contexts (Cleeremans, 1997) — such as a rule that is applicable to different instances of a certain problem.

Inherent to this issue is the question of whether the mechanisms through which implicit and explicit knowledge are acquired are best viewed as being subtended by separate processing systems. This is exactly what has been suggested by Shanks and colleagues (Shanks and St John, 1994; Shanks, Johnstone, and Staggs, 1997; St John and Shanks, 1997), who proposed to abandon the distinction between Implicit and Explicit Learning in terms of conscious awareness being present or not, and instead suggested that the distinction is one of rule-based versus memory-based learning processes. Before going any deeper into this matter, let us consider two different ways of looking at learning in general, to illustrate how they can possibly account for Implicit Learning.

## Computational Modelling of IL

Two views come forth when considering the mind in general, and implicit learning in particular: the symbolic and the connectionist approach. Each has its own view on how knowledge is represented and how it might be manipulated. The symbolic metaphor is usually associated with rule-based learning, while the connectionist metaphor is associated with memory-based learning based on the statistical characteristics of the stimuli.

**The Symbolic Metaphor.** Cleeremans & Jiménez (submitted) point out that a symbol system leaves no room for a concept like IL. In a symbol system, expressions that are formed are static representations of (real-world) entities or relations, stored in the system’s memory. These symbols, be it of objects or of rules, have to be interpreted by something — a processor — when they are to be used by the system to augment its knowledge base (memory), that is, to learn. From this perspective, IL can only exist if one assumes the existence of a *cognitive unconscious*, i.e. a subset of the mind that can basically process all the information that the conscious system can process, only minus consciousness. Consequently, consciousness is purely

epiphenomenal in this framework. It is exactly the fact that all symbols have in and of themselves the property of being an accessible representation, independent of the processor, which makes them unsuitable as a metaphor for implicit knowledge. For it is impossible to conceive of any knowledge that could influence processing while remaining unavailable to outside inspection. Importantly, this perspective also makes it possible to assume the existence of *abstract* knowledge that remains inaccessible to conscious inspection.

**The Connectionist Metaphor.** By contrast, in a connectionist network, there is no external processor engaged in learning, that is, learning does not consist of augmenting a distinct knowledge base. Instead, learning in a connectionist network is the *result* of changes that occur in the network (weight-change between units). These changes are themselves *caused by* information processing, i.e. the coupling of a certain input with a desired output. Thus, this processing also changes the process of learning (through for example back-propagation of the error between the actual and the desired output). Furthermore, as transient knowledge in a connectionist network consists of activation patterns, instead of symbols, a piece of knowledge does not have to be "interpreted" by the central processor before it can influence processing. These properties make it possible for a connectionist network to possess knowledge that can influence behaviour despite failing to be represented as such. It makes it possible to consider implicitness as something more than simply a property of the database or a property of the processor.

From the connectionist point of view, subjects are said to base their judgements on the basis of exemplar information, without explicitly extracting abstract generalities, or rules – the abstract processing is performed online during the test, when necessary. The episodic account provides a refined version of mere instance-based processing (e.g. Neal & Hesketh, 1997). One of the most popular instances of traditional connectionist networks is the Simple Recurrent Network (SRN), as proposed by Elman (1990). Here, judgements are no longer based on instances, but on instances *within their context*. Learning is nothing more than a byproduct of the processing itself (weight-change), while retrieval results from the overlap between processes operating during study- and test-phases. Several variations on this basic principle have been proposed, but the main point remains as stated: no abstract rules in implicit learning. Instead, more fragmentary knowledge is used to gradually and dynamically build up representations of the stimulus environment. This leaves room for implicitness, not in the way of equating the existence of representations with accessibility to consciousness (as do for example Perruchet and Vinter, submitted; O'Brien and Opie, 1999), but in virtue of the dynamical aspects of representation building. For example, it might be possible to conceive of conscious representations as being structured differently than unconscious ones, or as being of lower quality.

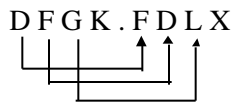
## Experimental Research on IL

Recently, some of the processes involved in word segmentation have been described as rooted in the same mechanisms as implicit learning and frequency estimation. For instance, Saffran et al. (1997) conducted an experiment on word segmentation in artificial speech. They exposed children (6-7 years old) and adult subjects to a continuous speech flow such as *bupadapatubitutibudutabapidabu*. Subjects were told that the experiment was about the influence of auditory stimuli on creativity (to make sure learning was incidental and not intentional). The only cues to word boundaries were the transitional probabilities between pairs of syllables (e.g., *bu-pa*), which were higher within words than between words. Afterwards, subjects heard two sets of sounds, each consisting of three syllable pairs, and were told to decide which one sounded more like the tape they had heard. Both adult and child subjects managed to perform well above chance, suggesting that learning might proceed in the absence of attention and the intention to do so, even despite the brevity of the exposure (one or two times a 21' tape). The fact that children did as well as adults suggests a robust phenomenon that might play a role in natural language acquisition.

In another interesting artificial language experiment, Marcus et al. (1999) claim to have showed that 7-month-old infants can "represent, extract, and generalise abstract algebraic rules." In short, the infants were exposed to artificial "sentences" during a training phase, and subsequently were presented with a few test items, some of them belonging to the same language, while others introduced some structural novelty. For example, when an infant had been habituated to *gatiti* or *linana* (both having an ABB structure), it was subsequently presented with test sentences such as *wofefe* or *wofewo* (the last one being of ABA structure). The basic set-up is similar to the Saffran et al. (1997) experiment, with the important difference that there where the Saffran et al. test items were composed of the same material as the training items, Marcus et al. introduced a change in the sensory content of the material. That is, prior to hearing the above illustrated test items, the infants had never heard */wo/*, or */fel/*. Still, infants tended to listen more to the sentences containing a structural novelty. As a result, since this task could not be performed on the basis of mere transitional probabilities, Marcus et al. concluded that infants had the capacity to represent algebraic rules. However, Marcus et al.'s claim that an SRN could not model the observed effect was disputed by Elman (Seidenberg & Elman, 1999; Elman, 1999) and McClelland and Plaut (1999), basically on the account that an overlap need not be present in the "raw input" itself. Instead "the relevant overlap of representations required for generalisation [...] can arise over internal representations that are subject to learning." (McClelland & Plaut, 1999, p.2) Transfer and generalization remain precarious issues, however, when it comes to computational modelling in a connectionist network. An experiment by Shanks et al. (1997) clearly illustrates this point.

## Biconditional AGL: Shanks et al. (1997)

As mentioned before, Shanks and St John (1994) proposed to abandon the idea of the conscious/unconscious dichotomy in favour of a rule-based/instance-based dichotomy. The basic idea is that humans possess two learning systems capable of creating distinct forms of mental representation, one system consisting of symbolic rule-abstraction mechanisms and the other involving subsymbolic, memory-based, connectionist mechanisms (see Shanks, 1998, for a discussion). In this context, Shanks et al. considered transfer in AGL tasks to be at least to some extent mediated by abstract (rule-) knowledge and claimed that people systematically become aware of the relevant regularities in AGL tasks where only rule learning is possible. To demonstrate, Shanks et al. exposed subjects to artificial grammar strings generated by a biconditional grammar (see also Mathews et al., 1989). Biconditional grammars involve cross-dependency recursion (see Christiansen & Chater, 1999) such that letters that appear at each position before and after a central dot depend on each other. An example is given in Figure 1, where letter D is paired with F, G with L, and so on.



**Figure 1.** A biconditional grammar string as used by Shanks et al. (1997). Possible letters in each position before the dot are linked biconditionally with the letters that may appear after the dot.

Shanks et al. constructed biconditional grammar training strings as well as a set of grammatical and nongrammatical and test strings, in such a way that grammatical and nongrammatical test items could not be distinguished on the basis of their overlap with the training strings in terms of bigrams or trigrams (or any other  $n$ -gram). During training, two groups of subjects were shown strings one a time on a computer screen and had to perform one of two tasks on each trial. One group (the match group) had been told that the task was about memory, and had to select the correct string among five strings presented on screen. The other group (the edit group) was told that the strings had been constructed according to rules and that their task was to find them. On each trial, edit subjects had to indicate which letters they thought violated or confirmed to the rules, and were subsequently given feedback. All subjects then performed a classification test in which they were asked to decide which strings were grammatical or not. Shanks et al. showed a dissociation between the two groups: While the edit group performed well and most subjects extracted the rules, the match group performed at a chance level, thus suggesting that "instance-memorisation and hypothesis-testing instructions recruit partially separate learning processes." (Shanks et al., 1997, p.243)

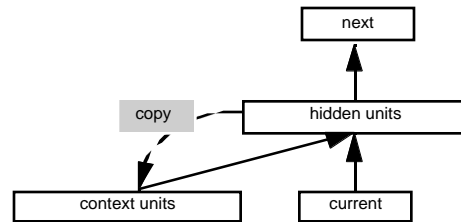
The basic claim is thus that, in order to perform the biconditional grammar task, it is necessary to conceive of some abstract (symbolic) rule-like knowledge of the

grammatical structure, and that, subsequently, the distinction made between grammatical and nongrammatical strings cannot be simulated by a connectionist network making use of simple frequency statistics. The goal of this paper is to demonstrate that in fact no such abstract rules are necessary and that, at least under some conditions, biconditional grammar learning can be accomplished by a network developing representations based on frequency statistics.

## A Simulation of Shanks et al.

### Simulation Parameters and Procedure

**The Simple Recurrent Network (SRN)** is a connectionist network especially designed to predict the next step in a sequence. Its design allows it to "keep in memory" the earlier steps in that sequence, by using what preceded as a *context*. This context is a copy of the learning-state at time  $t-1$ , which is fed back into the network at time  $t$ , together with the new input. In this way, the network is able to integrate the new input with what it has already learned in earlier stages, and will predict on this basis the sequence step at  $t+1$ . A typical example of an SRN is given in Figure 2.



**Figure 2.** The Simple Recurrent Network as conceptualised by Elman (1990).

Importantly, on each time step the context units contain a copy of the *patterns of activation* that existed over the *hidden units* at  $t-1$ . As described in Servan-Schreiber, Cleeremans and McClelland (1988, 1989; see Cleeremans, 1993), learning progresses in a continuous fashion through three stages, during which more and more temporal contingency information is incorporated in the context, and hence in the hidden unit representations. The statistical regularities the SRN uses to predict the next letter are thus gradually "stored" in the hidden unit representations of the network. As a consequence, the network becomes able to behave in a rule-like manner and to predict the next element in the sequence *as if* it knew the grammar rules.

**Network Architecture and Parameters.** The SRN had 9 input and output units, necessary for representing the information that was available to the subjects in the Shanks et al. experiment. (The six letters of which the strings were composed, D, F, G, L, K and X, as well as the beginning and end of each string, together with the dot in between the first and the last four letters of a string.) The number of hidden units (and hence context units) was 100, which made use of logistic adjustment. The learning algorithm was error backpropagation, with a learning rate of .15 and the context

being reset to zero after each complete string presentation. Weight adjustment was not applied on the connections from context to hidden units (1 on 1 relation). Momentum was set at .9.

**Training Material.** The basic training material consisted of a set of 18 strings as designed by Shanks et al. (List 1). Based on these strings, they created 18 grammatical and 18 nongrammatical strings.

The items were to meet four objectives: (1) Grammatical strings had to conform to the biconditional grammar: Letter position 1 is linked to 5, 2 to 6 and so on, with the linked letters being D-F, G-L, and K-X. (2) The use of the 6 letters was balanced, so that each letter appeared 3 times in each of the 8 letter locations. (3) Each training string differed from all other training strings by at least 4 letter locations. (4) Each training item had a grammatical similar item and an ungrammatical similar item that each differed from the training item by only 2 letter positions. Each training item was different from all other test items by at least 3 letter locations. The basic simulation was carried out on exactly these strings. A training epoch consisted of all 18 strings being presented once to the network, in a random fashion.

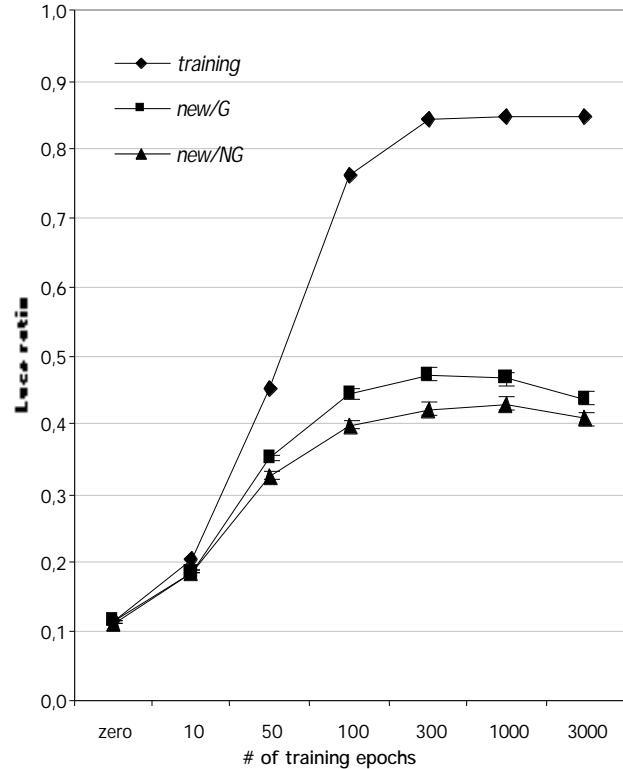
**Measurement of Accuracy.** Different measurements of accuracy exist, of which we used the *Luce ratio* (Luce, 1969) — a simple measure of relative strength in which the activation of the target output unit is divided by the sum of the activations of all output units. To assess network performance, we considered the average Luce ratios for all strings. In addition, we also considered the Luce ratio on a letter-by-letter basis for more detailed analyses.

### Simulation Results

**Learning.** In order to assess learning, the network was tested before and during training on seven occasions. On each test, the network was tested on the 18 grammatical training strings, the 18 new grammaticals, and 18 nongrammaticals. Results were obtained over 9 simulations and averaged. As described before, the Luce ratio of the output was computed for each element of each string. Subsequently, the ratios were compared over the two conditions of interest (grammatical test/nongrammatical test) by means of an ANOVA, for each learning step.

As can be seen in Figure 3, the SRN was indeed able to discriminate between grammatical and nongrammatical strings. Original training strings were learned almost perfectly from 100 epochs onwards. Further, the network clearly discriminates between novel grammatical and nongrammatical strings (i.e., better predictions for grammatical strings), even *before* it is completely successful in mastering the training strings. ANOVA measures are, at 50 epochs,  $F(1,161)=24.1$ ,  $p<.001$ ; at 100 epochs,  $F(1,161)=36.3$ ,  $p<.001$ ; at 300 epochs,  $F(1,161)=33.5$ ,  $p<.001$ . From 1000 epochs onwards, the network gets a little 'overtrained' on the original strings, causing it to do somewhat less well on the unseen strings; at 1000 epochs  $F(1,161)=13.3$ ,  $p<.001$ ; at 3000 epochs  $F(1,161)=8.34$ ,

$p<.005$ . The figure also makes it clear that the main effect is not due to some initial biasing since initial performance is identical for the three types of strings (prior to training,  $F(1,161)=1.13$ , ns; at 10 epochs,  $F(1,161)=.048$ , ns).



**Figure 3.** Network learning, measured with the Luce ratio. Error bars are shown for novel G and nonG strings.

Based on these findings we can therefore conclude that contrary to what Shanks et al. claimed, the SRN can in fact distinguish between grammatical and nongrammatical strings generated by a biconditional grammar without making use of explicit rules. In order to rule out the possibility of the SRN merely having learned to predict the dot and/or the end of a string, we computed the mean Luce ratios on a letter-by-letter basis, as presented below in Table 1. Shown are the important ratios, belonging to the letters after the dot (ratios for training strings had value 1). When the difference exceeds .05, the highest ratio is in bold.

Table 1 clearly shows the mean Luce ratios on a letter-by-letter basis to be higher in grammatical than in nongrammatical strings. This indicates that the network has learned something other than merely the dot or the end of the string.

**Table 1.** Mean Luce ratios on a letter-by-letter basis, in each position, after 3000 epochs, for grammatical and nongrammatical test strings (included is the frequency of occurrence of the letter in each position).

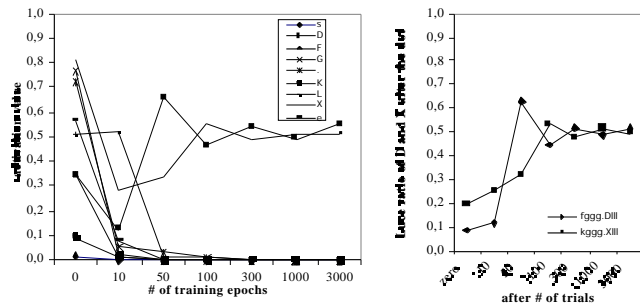
## GRAM

	5th	#	6th	#	7th	#	8th	#
D	.25	(2)	.89	(2)	.01	(4)	.69	(4)
F	.70	(4)	.50	(4)	.55	(1)	.18	(3)
G	.71	(2)	.49	(2)	.26	(5)	.38	(3)
K	.41	(2)	.72	(5)	.09	(3)	.99	(2)
L	.58	(4)	.99	(3)	.11	(3)	.68	(3)
X	.43	(4)	.56	(2)	.33	(2)	.01	(3)

## NGRAM

	5th	#	6th	#	7th	#	8th	#
D	.19	(2)	.70	(2)	.72	(4)	.13	(4)
F	.39	(4)	.87	(4)	.20	(1)	.75	(3)
G	.37	(2)	.37	(2)	.17	(5)	.01	(3)
K	.46	(2)	.32	(5)	.19	(3)	.33	(2)
L	.74	(4)	1.00	(3)	.00	(3)	.50	(3)
X	.50	(4)	.38	(2)	.50	(2)	.00	(3)

**When the network fails to learn.** In order to illustrate exactly when a network can learn, we include a simulation of a situation in which it *fails to learn*. We created two (grammatical) strings with a high degree of similarity, FG GG.DLLL and KG GG.XLLL, and presented them to the network in the same way as was done in the main simulation. Figure 4 shows the activation values of the 9 output nodes for one string (the other showed the same evolution in activation values), as well as the evolution of the Luce ratios.



**Figure 4.** Left Panel: Evolution of output unit activations after presentation of the dot in the FG GG.DLLL string. Right Panel: Evolution of the Luce ratio after presentation of the dot for the two strings FG GG.DLLL / KG GG.XLLL.

Here, the network fails to reach a decision: it gets stuck at a "post-dot" activation value of 0.5 for both D and X (exactly the same plot is produced for the other string). The reason why learning fails in this case is addressed in the discussion.

## Discussion

What had to be shown *was* shown, namely that a connectionist network, more precisely a Simple Recurrent Network, is able to make a distinction between grammatical and nongrammatical letter strings, generated from a biconditional grammar as used by Shanks et al. (1997). These strings were designed so that, according to them, subjects *had* to make use of abstract rules in order to

accomplish the categorisation task. This paper clearly demonstrates that this is not the case, and that judgements of grammaticality using biconditional grammars can be made by extracting statistical features out of the material.

One of the major challenges in working with connectionist networks is how to probe the hidden units in order to "unfold" the complex representation of the stimulus material. Cluster analysis or principal component analysis performed on the hidden unit activations are standard ways of doing so, but may not always provide insight into how the representations enable the network to solve the task. The fact that cluster analysis does not reveal a clear structure does not necessarily imply that there *is* no structure. It may simply mean that the representational aspect needed to accomplish the most important aspect of the task, is *not* the most important aspect. Thus, clustering will not be carried out on that aspect — which, importantly, does not necessarily entail that the network is unable to use the relevant information successfully (see Cleeremans, 1993).

Biconditional grammars are difficult to master because they require maintaining information across intervening irrelevant items. Servan-Schreiber et al. (1991) explored the conditions under which the network can carry information about distant sequential contingencies (e.g. 1–5) across intervening elements, to distant, to-be-predicted elements. It appeared that this information is retained as long as it is in some way relevant to predicting each intervening item (the *prediction-relevance* criterion). When it is not, the relevant information tends to be lost as training progresses, as a consequence of the way in which representations of the temporal context are only gradually built up. Indeed, for different predictions to be achieved at any point in a sequence, the network needs to have developed different internal representations of the sequence so far. When two sequences are identical for a number of time steps so that the relevant information for making different predictions has to be retained over these intervening elements, each training trial actually induces the development of increasingly similar internal representations of the two sequences (because they require similar predictions)— exactly the opposite of what would be required for the network to master the material. Hence, the network fails to predict the fifth letter in the example above because the first letter of each string fails to be prediction-relevant when processing the intermediate Gs and ends up, as a result, with internal representations that fail to be sufficiently distinctive of each string to enable it to make different predictions about the fifth letter when presented with the dot.

Shanks et al. however, could not present the extremely simple (and for the network, extremely difficult) material to their subjects, for everyone would have discovered the rule in that case. Importantly however, the way in which their material is constructed results, for instance, in all the training strings to be determined by their *first two* elements — something that enables the network to learn the construction paths of each training string very quickly. In addition, in most cases, sequential information was in fact prediction-relevant on each step, which makes it easy for the network to distinguish between grammatical and nongrammatical strings. These findings suggest that the

Shanks et al. material was in fact inadequate to test for the rule based versus memory based distinction. As mentioned before however, it is clearly impossible to conceive of easy strings like KGGG.XLLL for which the rules are not discovered by subjects.

Insofar as simulations are concerned, while the SRN fails on such degenerate cases (unlike human subjects), the issue of whether this failure reflects a principled limitation of connectionist networks in general remains an open issue. Servan-Schreiber et al. showed that even very slight adjustments to the statistical structure of otherwise identical sequences could greatly enhance the prediction accuracy of the SRN. Thus, embedded information, as in recursive structures, need only be prediction-relevant in terms of the statistical distribution of the embedded elements for such structures to be successfully mastered by an SRN. There is also accumulating evidence that the pattern of failures observed with models like the SRN closely mimic that observed with human subjects (e.g., Christiansen & Chater, 1999) in the domain of natural language learning.

Empirically, we would like to suggest that experiments be carried out on a slightly different basis than used in Shanks et al., since their 'match' group showed no sign at all of having learned the material. One possibility would consist of changing the instructions of the match group so that attention is not *drawn away* from certain properties that might allow subjects to become sensitive to the structural properties of the material.

To conclude, we have demonstrated that a simple connectionist network can in fact master material previously considered to *require* the acquisition of rule-based knowledge for mastery of novel instances to occur. This outcome does not entail that rule-based learning never occurs (as it obviously does for some subjects in Shanks et al.'s experiments), but simply that biconditional grammars might not address all the issues involved in efforts to dissociate rule-based vs. memory-based learning processes in the implicit learning literature. Further simulation work will attempt to explore these issues in greater depth.

### Acknowledgments

Axel Cleeremans is a Research Associate of the National Fund for Scientific Research (Belgium). This work was supported by a grant from the Université Libre de Bruxelles in support of IUAP program #P/4-19.

### References

Berry, D.C. and Dienes, Z. (1993). *Implicit learning: Theoretical and empirical issues*. Lawrence Erlbaum Associates, Hove.

Christiansen, M., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance, *Cognitive Science*, 23, 157-205.

Cleeremans, A. (1993). *Mechanisms of implicit learning*. MIT Press, Cambridge, MA.

Cleeremans, A. (1997). Principles for implicit learning. In D.C. Berry (Ed.), *How implicit is implicit learning?*, pp. 195–234. Oxford University Press.

Cleeremans, A., Destrebecqz, A., and Boyer, M. (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences*, 2, 406–416.

Cleeremans, A. and Jiménez, L. (submitted). Implicit cognition with the symbolic metaphor of mind: Theories and methodological issues.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, 14, 179–211.

Elman, J.L. (1999). Generalization, rules and neural networks: A simulation of Marcus et al. (1999).

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol.1). New York: Wiley.

Marcus, G.F., Vijayan, S., Bandi Rao, S., and Vishton, P.M. (1999). Rule learning by seven-month-old infants. *Science*, 283, 77–80.

Mathews, R.C., Buss, R.R., Stanley, W.B., Blanchard-Fields, F., Cho, J.R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 1083-1100.

McClelland, J.L. and Plaut, D. (1999). Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences*, 3, 166-168.

Neal, A. and Hesketh, B. (1997). Episodic knowledge and implicit learning. *Psychonomic Bulletin and Review*, 4, 24–37.

O'Brien, G. and Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences*, 22, 175-196.

Perruchet, P. and Vinter, A. (submitted). The self-organizing consciousness.

Saffran, J.R., Newport, E.L., Aslin, R.N., Tunick, R.A., and Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, 8, 101–105.

Seidenberg, M.S. & Elman, J.L. (1999). Do infants learn grammar with algebra or statistics? Letter in *Science*, 284.

Servan-Schreiber, D., Cleeremans, A., & McClelland, J.L. (1991). Graded State Machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning*, 7, 161–193.

Shanks, D.R. (1998). Distributed representations and implicit knowledge. In K. Lamberts and D.R. Shanks (Eds.), *Knowledge, concepts & categories*, pp.197–214. Psychology Press, Hove.

Shanks, D.R., Johnstone, T., and Staggs, L. (1997). Abstraction processes in artificial grammar learning. *The quarterly Journal of Experimental Psychology*, 50A, 216–252.

Shanks, D.R. and St John, M.F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, 17, 367–447.

St John, M.F. and Shanks, D.R. (1997). Implicit learning from an information processing standpoint. In D.C. Berry (Ed.), *How implicit is implicit learning?*, pp. 124–161. Oxford University Press.

# Representational Scaffolding During Scientific Inquiry: Interpretive and Expressive use of Inscriptions in Classroom Learning

Eva Erdosne Toth (etoth+@andrew.cmu.edu)  
Department of Psychology; 5000 Forbes Avenue  
Pittsburgh, PA 15213

## Abstract

External representations (inscriptions) such as tables, visualizations, graphs and diagrams have been widely studied to determine their cognitive effects. However, a research-based pedagogy for their classroom use is yet to be offered. The two independent studies presented here are conceived from the same standpoint: the cognitive effects of inscriptions are influenced by the mode of their use (interpretive or expressive). The results of these studies indicate that the interpretive and expressive use of different external representations during scientific inquiry influenced students' (1) understanding the logic of designing experiments and their (2) ability to coordinate experimental data with theories. Representational scaffolding with collaboratively shared evidential-consistency maps helped students overcome traditional "inquiry traps" such as confirmation bias. These results show that representational scaffolding can provide an effective pedagogy for cognitively-based instructional interventions to teach scientific inquiry skills.

## Introduction

Representing knowledge with inscriptions such as pictures, diagrams and maps is a pivotal aspect of scientific practice. These external representations can be tools to think with for the clarification of ideas while designing experiments, analyzing data, and formulating theories. They can also be tools to talk with, aiding communication between a community of scientist-peers. Increasingly, everyday decisions are based on information presented with external representations found in textbooks, newspapers and even on breakfast cereal boxes. Today it is considered part of basic education to be able to understand and communicate with external representations. Arguably, external representations are at the center of both scientific and everyday reasoning. Thus a study of how one learns to make sense of and reason with representations has great significance.

For the purpose of research on learning, external representations should be differentiated from the internal products of one's thinking, often also described with the terminology "representations" (Kotovsky & Simon, 1990). I use the word "inscriptions" (Latour & Woolgar, 1979; Lehrer & Schauble, 1998) interchangeably with "external representations" and with both I refer to artifacts of thinking existing outside of one's head. My primary aim with this article is to describe the potential use of these external representations of thinking to scaffold learning. First, I describe the theoretical grounding of my overall approach to the understanding and use of inscriptions in classroom

learning. Next, I detail two studies, which describe the effects of inscriptions on two crucial aspects of scientific inquiry: (a) the logical process of setting up informative (unconfounded) experiments and (b) the reasoning associated with making sense of empirical observations by building evidential-consistency relationships between experimental data and theories. Finally, I describe the implications of these results for learning and teaching scientific inquiry skills and outline future research to further develop a cognitively-based pedagogy built on representational scaffolding.

## Theoretical grounding

My analysis of the educational benefit of various forms of external representations starts with a framework outlining the cognitive value of inscriptions. Developed by Collins and Ferguson (1993), this framework states that each form of external representation carries with it a methodology or heuristic for its use. External representations function through two mechanisms. They can (a) narrow the space of information search by localizing the most important message into perceptually salient, jointly displayed chunks (Larkin & Simon, 1987). They can also (b) provide a way for previously obscured information to become available (to "emerge") during the development and interpretation of inscriptions (Koedinger, 1992). That is, external representations scaffold activity by making certain aspects of inquiry salient (Stenning & Oberlander, 1995) and by constraining the user to certain activities (Suthers, 1999). With such mechanisms external representations provide representational scaffolding during classroom learning.

For example, students provided with the table shown in Figure 1a can clearly see that the most important variables to consider when determining what makes balls roll farther down ramps are the steepness of the ramp, the length of run, the surface of the ramp and the type of ball. No other variable is deemed important by the developer of this table, thus students' thinking is constrained to these essentials. One salient piece of information from this table is that all four of these variables should be considered for each apparatus of a simultaneous comparison (for both ramp A and ramp B).

Another salient component of experimental design emerges when students fill this table out. If students neglect to pay attention to any of the variables on one of the ramps the missing information becomes salient – as indicated by



the empty cells in Figure 1b. Thus, representational scaffolding can provide an innovative instructional methodology to teach students how to design informative (un-confounded) experiments.

VARIABLES	RAMP A	RAMP B
Surface		
Steepness		
Length of run		
Type of ball		

Figure 1a. Representational scaffolding by constraining students' thinking to the focal elements of experimentation.

VARIABLES	RAMP A	RAMP B
Surface	Smooth	Rough
Steepness	High	High
Length of run	Long	
Type of ball	Golf	

Figure 1b. Representational scaffolding by the salience of empty cells drawing students' attention to variables that may have been ignored during experimental testing.

With few exceptions (Lehrer & Schauble,1998) however, the full potential of the use of inscriptions in classroom science learning environments has not been examined.

There are two modes of using inscriptions: interpretive and expressive. These two modes are inherently combined during the development and the use of external representations in scientific practice. In classroom environments, however, students usually use inscriptions interpretively: they are given a teacher-developed external representation to make sense of by observing parts or by completing, "filling it out" – as shown in Figure 1. Expressive use of inscriptions entails the active generation of a form of external representation with the aim of communicating an idea. While students' active generation and manipulation of their own knowledge is considered important under the currently dominant constructivist pedagogy, inscriptions are rarely used expressively by students during classroom learning.

The study of inscription use in classroom learning environments is best approached by the examination of students' difficulties with each of these modes of using inscriptions. The first study described here shows how students used inscriptions both interpretively and expressively while designing and conducting scientific experiments and recording data results. The second study describes the effects of a software tool that eases students' way into working with external representations. Both of these studies illustrate the value of learning with inscriptions and indicate which student difficulties should be considered

in the further refinement of an instructional methodology that is built on representational scaffolding.

## Representational scaffolding during scientific experimentation

### Subjects and Procedures

Two classrooms of 28, 4<sup>th</sup> grade students experimented in small groups of 3-4 to determine what makes balls roll farther down on ramps. The instructional goal was to teach valid experimentation skills, specifically the control of variables strategy or CVS. The task of each student group was to design experiments with a pair of physical ramps and record their experiments in laboratory notebooks. The outcome of the research-based methodology to teach valid experimentation skills from this study was detailed in Toth, Klahr & Chen (in press). The present paper focuses on the effects of using inscriptions during scientific inquiry in the classroom. Two slightly different procedures were employed by the classroom teacher, each allowing a focused view at either the interpretive or the expressive use of inscriptions.

During the first week of instruction student groups learned how to create controlled comparisons with pairs of ramps and filled out a teacher-specified table of dependent and independent variables to indicate their experimental setup. This table representation was similar to that in Figure 1, with the exception that the two possible values for each of the four variables were already provided in the cells of this table. The students' role was simply to circle the variable value of their choice to map their ramp setup onto the table. Prior to experimentation, students were specifically taught how to conduct this mapping activity. This interpretive use of inscriptions was intended to help students as well as the teacher keep track of the experimental designs used over time. The researcher's classroom observation notes recorded student's activities and their difficulties.

During the second week of the study - after two weeks of spring break – the students applied the previously learned skill of designing controlled experiments to learn more about ramps. This time they designed experiments using one ramp at a time. They were asked to record each of their experiments into any external representations of their choice (expressive use). Data sources included videotapes of classroom experimentation, laboratory notebooks recorded by students and the researchers' observation notes. Students' expressive use of inscriptions was scored using a modified version of a quantitative coding scheme suggested by Kosslyn (1989). The scoring included attention to how the inscription communicated the logical design of experiments and the clarity of reasoning inherent in it.

### Results and Discussion

**Interpretive use of inscriptions** Classroom observations of the interpretive use of inscriptions (first week) indicated that

as soon as students looked at the teacher-developed table and tried to fill it out, they wanted to go back to the physical ramps and make changes to the setup of the variables. They indicated that they wished to set their tests up "better" or "differently." While this intent of the students created a slight problem for the experimenters – who at the time were interested in documenting students' developing knowledge of experimentation strictly adhering to a prior protocol developed during laboratory studies (Chen & Klahr, 1999) – this observation soon led to the realization that the interpretive use of the pre-developed table representation may have helped students abstract the overall structure of the experiment and thus aided their understanding of the design of un-confounded experiments. I hypothesize that as students filled out their teacher-defined table an important characteristic of scientific inquiry may have become salient to them: the criteria that all important variables of an experimental setup should be considered during experimental design. As outlined above (Figure 1) this table constrained students' thinking to the most important variables, but also made student errors in designing informative (controlled) experiments salient to them. If one of the variables was not attended, this omission became obvious ("emerged") during the work with the representation. Similarly, if two variables were changed instead of one between experiments (a confounded experiment was created) this oversight was made perceptually salient. Thus information previously not available to students became obvious through the representational scaffolding provided by the use of this inscription.

After a week of focused experimentation and instruction, students learned the strategy of creating controlled experiments from which they could tell with certainty the effect of any focal variable under investigation (Toth, Klahr & Chen, in press). Having learned to overcome systemic error with the use of the control of variables strategy (CVS) during experimentation, students were presented with a new challenge: to record their experiments with inscriptions of their own choice (expressive use) while they continued applying CVS to learn more about variables associated with ramps. The subsequent analysis of the student-developed inscriptions from this second week of study revealed various student difficulties.

**Expressive use of inscriptions** Two characteristics of the expressive use of inscription were noticeably difficult for students: (1) using the common techniques of developing inscriptions (using labels and data correctly in a coordinated way) and (2) reasoning scientifically with inscriptions. Various problems resulting from the lack of experience with a common representational technique were identified. Common problems included missing labels, missing data content and insufficient alignment of data with labels.

Students' laboratory notebooks fell into three specific patterns in terms of reasoning scientifically about

experiments through inscriptions: (a) showing incorrect CVS only, (b) showing a combination of correct and incorrect CVS over time, (c) indicating a possible search for interaction of variables without clear CVS design

These effects were found even after students were documented to have learned the control of variables strategy (Toth, Klahr and Chen, in press). That is the effect found here can be attributed to either students' inability to use external representations or the lack of transfer of the CVS strategy to situations slightly changed from the condition of learning.

Overall it appears that the interpretive use of a well-selected external representation (Figure 1) can positively influence student's understanding of skills associated with scientific inquiry. This effect is due to the representational scaffolding inherent in any form of inscription. This characteristic makes certain inscriptions especially fitting for a learning task while not appropriate for others. In this example, a table representation appears to be fitting for the interpretive task of abstracting and combining the logical components of scientific experimentation. However, the expressive use of inscriptions is more problematic as it should consider the structure of the domain, the goal of the activity as well as the cognitive state of the interpreter. It is hypothesized that innovative pedagogies such as collaborative reflection and discussion conducted through external representations may help students learn the skills of developing effective inscriptions.

The next study describes a software tool called Belvedere (Suthers et al., 1997) that eases students' way into working with external representations. It also details the effects of representational scaffolding by different forms of representations and suggests a reflective methodology to support the use of inscriptions in classroom learning environments.

## **Representational scaffolding while coordinating data with theories**

### **Subjects and Procedures**

Four classrooms of 9<sup>th</sup> grade students (N = 73) participated in a 2X2 research design in which the effects of two different external representations (evidence mapping vs. prose writing) was studied. As part of their science class—taught by their regular science teacher—the students participated in problem-based-learning. They were presented with a set of scientific challenges to which no known solutions existed at the time. Their task was to explore web-based information sources – a set of researcher developed<sup>1</sup>, hypertext materials – and to find a solution to a scientific challenge such as mass extinctions, the evolution of marine iguanas or the sudden appearance of a mysterious

---

<sup>1</sup> The materials used in this study were developed primarily by Arlene Weiner with assistance from the author.

disease. Evidence mapping entailed using a shared, whiteboarding software tool, BELVEDERE, to diagram evidential consistency relationships between hypotheses and data. BELVEDERE's main menu provided epistemological categories that included object-shapes for hypotheses (rounded rectangle shapes in Figure 2), data (rectangle shapes in Figure 2) and links to indicate consistency ("for" links), inconsistency ("against" links) and conjunction ("and" links) between data and hypotheses (Figure 2).

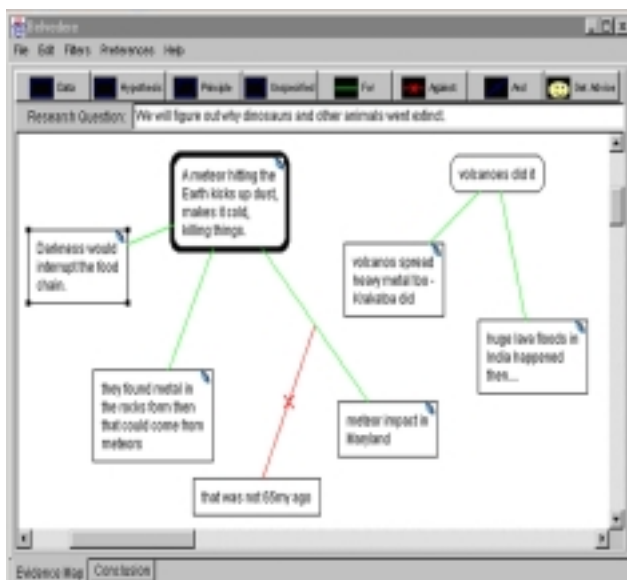


Figure 2. Example evidence map created with the BELVEDERE software tool.

The diagramming activity started with exploring the web-based materials for information and continued with recording ideas using the software tool. In order to make this record with the mapping tool students needed to categorize the currently considered information as data or hypothesis, select from the appropriate shape from the menu pallet and copy the information being considered into the selected object-shape. Similarly students could choose links from the menu to indicate evidential consistency relationships between data and hypotheses (Figure 2.)

Students not using the software tool for mapping used prose writing to record their thinking during inquiry. Prose writing consisted of using a word-processor to write a prose-based account of exploring the web-based materials and solving the challenge problem. Students were instructed to record the main aspects of their inquiry: both data and hypotheses and detail how they decided on a conclusion. In addition, one of the classrooms in each condition (mapping and prose) was given a method of explicitly reflecting on the process of their inquiry by using a paper-based handout of specific inquiry criteria detailing optimal performance during inquiry. These so-called “reflective assessment rubrics” were used from the beginning of inquiry for

reflection as well as during final assessment of performance. Students who did not use the explicit reflection were only implicitly prompted about the criteria by which their work will be evaluated. It was expected that this implicit prompting would result from the structure of the challenge materials and the nature of the inquiry activity.

Thus each of the four classrooms was randomly assigned to one of the following treatments: Map&Reflect, Map-NoReflect, Prose&Reflect, Prose-NoReflect (Figure 3).

Treatment	Materials	Representation	Reflection
Map & Reflect	Hyper-Media	Mapping	Reflection
Map - NoReflect	Hyper-Media	Mapping	No-Reflection
Prose & Reflect	Hyper-Media	Prose	Reflection
Prose - NoReflect	Hyper-Media	Prose	No-Reflection

Figure 3. Elements of experimental design for study two.

The effects of representational scaffolding on students' inquiry skills was analyzed from students' inquiry artifacts - the products of their work (maps or prose). Though students conducted many units of problem-based-learning, the results discussed here are from the first unit only. This unit was chosen with the expectation that the use of inscriptions in the early stages of inquiry would provide intricate details of representational scaffolding inherent in these inscriptions. It was expected that this software tool will ease students' into reasoning about data and hypotheses with inscriptions, as it scaffolds the development of evidence-map representations with pre-defined categories that can be flexibly used to allow students' to actively generate and manipulate their own scientific knowledge.

As an indicator of students' effectiveness during inquiry three scores were analyzed: (1) the number of information units (hypotheses and data) recorded and correctly categorized, the (2) number of inferences recorded (indicating correct relationships between data and hypotheses) and the (3) the quality of final conclusions drawn by students in each condition.

## Results and Discussion

Groups in the prose writing condition recorded about the same number of hypotheses and data as groups in the evidence mapping condition. However, analysis of the amount of information that was correctly categorized by students revealed that the mapping groups categorized significantly more of their recorded information as hypothesis and data compared to the prose writing groups. Since in either condition (mapping or prose) the percentage of incorrectly categorized information units was minimal

(<10%) it is reasonable to argue that it was the effect of the mapping representation that scaffolded students' categorization efforts. That is, the evidential consistency mapping, with its pre-defined epistemological categories, prompted students to consider the meaning of these categories and to organize the outcome of their investigations based on these categories. Unlike the mapping activity, the prose writing was a familiar mode of communication for students. However, the prose representation did not make the categories of scientific inquiry perceptually salient throughout students' investigation, resulting in the lower number of information pieces categorized by the prose writing groups. There was no significant effect of the type of reflection on either the number of information units recorded nor on the categorization of these records.

Interesting representational scaffolding effects were found during the analysis of the inferences recorded by student groups in the different conditions. The mapping groups recorded significantly higher number of inferences describing relationships between data and hypotheses compared to the prose writing groups (Figure 4).

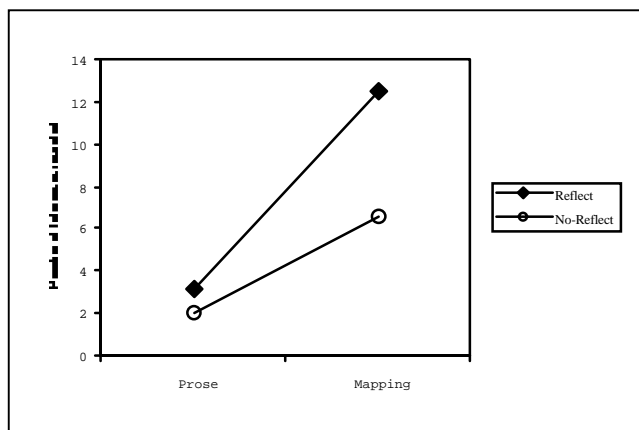


Figure 4. The effects of representational scaffolding and explicit reflection on the number of inferences between data and hypotheses recorded by students

The use of explicit reflection also significantly influenced student's ability to express inferences indicating the relationship of their data and hypotheses. The Map&Reflect groups performed significantly higher than any other group, including the Map-NoReflect group on this measure.

Analysis of the types of inferences (consistency, inconsistency and conjunction) revealed that the difference between the Map&Reflect groups compared to the Map-NoReflect groups was in the frequency of inconsistency ("against") relationships recorded (Figure 5). This is a crucial finding that indicates the value of both mapping and explicit reflection and their combined effect helping students overcome confirmation bias during the evaluation of scientific hypotheses based on empirical data.

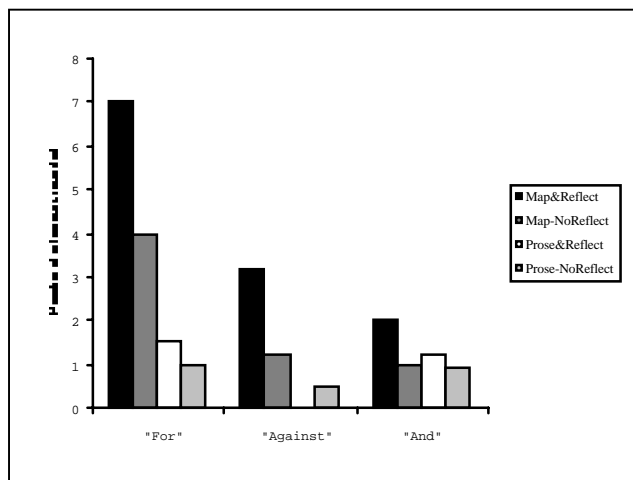


Figure 5. Sub-scores of information evaluation by groups in the four treatment condition.

Furthermore, when students' final reasoning with prose conclusions were analyzed there were no significant differences between prose and mapping groups in the quality of the final conclusions. Since these conclusions were written in prose by students in both conditions (mapping or prose), this finding indicates the lack of efficiency in transferring the inquiry skills learned by the mapping groups from the mapping activity to prose writing. Further instructional interventions are necessary to ensure a more effective transfer.

Overall it appears that the evidence mapping provide better scaffolds for students during scientific inquiry when the goal of activity is to categorize information and evaluate scientific hypotheses based on evidence. Explicit reflection on the specific criteria of scientific inquiry was found to support the evidence mapping activity, but not traditional prose writing.

## Conclusion and Educational Significance

External representations of thinking can play a pivotal role in the learning of scientific inquiry skills. The studies presented here detailed the effects of inscriptions during two important processes of scientific inquiry: (1) designing and conducting experiments and (2) coordinating experimental evidence with domain theories. The classroom studies described here yielded evidence for two methods of using inscriptions: interpretive and expressive. Interpretive use of a teacher-developed table representations was found to scaffold students' progress of inquiry by making the variables of an experiment salient and by perceptually constraining the students' attention to abstract the characteristics of correct experimentation. However, during the expressive use of inscriptions students were found to have difficulty with using the common techniques of

developing inscriptions and with indicating their reasoning during experimentation. While the lack of transfer of CVS skills has been clearly documented since this study (Toth & Klahr, 2000) further instructional interventions and studies can reveal how a reflective pedagogy based on representational scaffolding may help with this transfer difficulty. The results of study two indicate that such methodology may be very effective.

While evaluating experimental data against theories, the representational scaffolding effect provided by evidential consistency mapping (compared to prose writing) was confirmed. Evidence mapping was found to be a successful instructional methodology to teach how to categorize and label scientific information and to teach students how to evaluate hypotheses based on empirical data. The findings also suggested that reflective assessment (by the use of the explicit criteria for maximum performance in rubrics format) was an effective instructional manipulation to support the scaffolding effect of external representations.

In collaborative classroom learning environments there has been a need for a methodology that combines cognitive effectiveness with the social circumstances of collaborative learning. The effects of collaborative reflection over shared representations seem to be promising and should be further refined. One such study is currently under way by the author to explore how the social circumstances of the anticipated peer-interpretation of inscriptions influence students' expressive use. Further research should also consider the use of software tools in the shared activity of expressively and collaboratively using various forms of inscriptions.

### Acknowledgements

Funding for study one was provided by the James S. McDonnell Foundation to David Klahr and Zhe Chen of the Department of Psychology, Carnegie Mellon University. Funding for study two was provided by the Presidential Technology Initiative to Dan Suthers, Eva Toth and Arlene Weiner who worked under the leadership of Alan Lesgold at the University of Pittsburgh Learning Research and Development Center (LRDC). Many thanks to colleagues David Klahr, Jennifer Schnackenberg, Anne Siegel, Rose Russo and Jolene Watson for their contributions.

### References

Collins, A. and Ferguson, W. (1993). Epistemic forms and epistemic games: Structures and strategies to guide inquiry. *Educational Psychologist*, 28(1) pp 25-42.

Chen, Z., and Klahr, D. (1999) All other things being equal: Children's acquisition of the control of variables strategy. *Child Development* 70(5), 1098 – 1120.

Koedinger, K. R. (1992). Emergent properties and structural constraints: Advantages of diagrammatic representations for reasoning and learning. *Symposium of the American Association for Artificial Intelligence*. Stanford University.

Kotovskiy, K., and Simon, H.A. (1990). What makes some problems really hard: Explorations in the problem space of difficulty. *Cognitive Psychology*, 22, 143-183.

Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, 3, 185-226.

Latour, J., and Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*. Beverly Hills, CA: Sage.

Larkin, J. H., and Simon, H. A. (1987). Why a picture (sometimes) is worth ten thousand words. *Cognitive Science*, v. 11 pp 65-99.

Lehrer, R., and Schauble, L. (1998). Inventing data structures for representational purposes: Elementary grade students' classification models. *Paper presented at the annual meeting of the American Educational Research Association annual convention, San Diego, CA*.

Roth, M.W. and McGinn, M. K. (1998). Inscriptions: Towards a theory of representing as social practice. *Review of Educational Research*, 68 (1), 35-59.

Stenning, K. and Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science* 19(1), pp 97-140.

Suthers, D., Toth, E., and Weiner, A. (1997). An Integrated Approach to Implementing Collaborative Inquiry in the Classroom. *Proc. 2nd Int. Conf. on Computer Supported Collaborative Learning*, pp. 272-279.

Suthers, D. D. (1999). Representational support for collaborative inquiry. *Proceedings of the 32nd Hawaii International Conference on the System Sciences* January 5-8, 1999, Maui, Hawaii

Toth, E. E., Klahr, D., and Chen, Z. (in press). Bridging research and practice: A cognitively-based classroom intervention for teaching experimentation skills to elementary school children. *Cognition and Instruction*.

# From Dipsy-Doodles to Streaming Motions: Changes in Representation in the Analysis of Visual Scientific Data

**Susan B. Trickett**  
([stricket@gmu.edu](mailto:stricket@gmu.edu))  
Department of Psychology  
George Mason University  
Fairfax, VA 22030

**Wai-Tat Fu**  
([wfu@gmu.edu](mailto:wfu@gmu.edu))  
Department of Psychology  
George Mason University  
Fairfax, VA 22030

**Christian D. Schunn**  
([schunn@gmu.edu](mailto:schunn@gmu.edu))  
Department of Psychology  
George Mason University  
Fairfax, VA 22030

**J. Gregory Trafton**  
([trafton@itd.nrl.navy.mil](mailto:trafton@itd.nrl.navy.mil))  
Naval Research Laboratory  
NRL Code 5513  
Washington, DC 20375

## Abstract

This paper investigates the change in scientists' representation of phenomena of interest during the exploratory analysis of visual data. The scientists initially represented expected findings in formal, scientific terms, whereas they represented anomalies in informal terms. Over time, these representations shifted from informal to formal. We propose that this shift in representation is the result of an increased understanding of the individual phenomena, rather than of greater understanding of the data at a global level.

## Introduction

A strong and perhaps foundational theme in cognitive science is the issue of representation. From both empirical and computational perspectives, performance has been found to depend heavily on how information is internally represented (Kotovsky, Hayes, & Simon, 1985; Larkin & Simon, 1987; Newell & Simon, 1972; Zhang & Norman, 1994).

One area in which representation is likely to be especially important is scientific discovery (Schunn & Klahr, 1995). There are many formal and informal methods for representing data, even within the same discipline and narrow sub-area. The choice of representation of the data is likely to have a large impact on what can and will be discovered.

An additional twist on the issue of data representations in science is the difference between goals of scientific discovery and goals of communication of the discoveries. The external representations that are best for discovery are not necessarily the representations that are best for communication of the discovery to others. For example, issues of historical convention are likely to be more important in communication, whereas issues of ease of generation and manipulation are going to be more important for the original discovery.

The goal of this paper is to examine how scientists represent data internally to themselves while they are analyzing their data. In particular, do they tend to think of their data in formal, discipline-specific terms, or do they rely on more informal and simple perceptual terms? One might expect them to use formal terms because of their expertise and extensive domain knowledge. On the other hand, they may use perceptual terms because in many areas of science, the data are presented in fairly complex visual displays that make heavy use of spatial metaphors—or indeed represent spatial dimensions directly (Trafton et al, under review).

One dimension that we hypothesize would influence the choice of internal representation is the degree to which the

data are as the scientist expects. That is, perhaps scientists are more likely to represent apparently anomalous data in informal, perceptual terms and expected data in formal, conceptual terms.

Another related dimension that we investigated was time: How do scientists' representations of their data change over time as they explore their data? One might imagine that the representations become more formal as scientists develop an understanding of the dataset as a whole. Alternatively, the changes in representation may occur at a more item-specific level—the representation of each item changes separately as understanding of the item changes.

A wide variety of methodologies has been used to study scientific reasoning and scientific discovery, each with their advantages and disadvantages (see Klahr & Simon, 1999, for a review). For this research project, we adopted a modified form of Kevin Dunbar's "in vivo" methodology (Dunbar, 1995, 1997, in press). The "in vivo" methodology involves observing scientists as they are doing their research. Dunbar focused on the activities that occur in lab group meetings. Because we were interested in the processes of data analysis, we focused, instead, on pairs of scientists working at their computers, analyzing their data. Like Dunbar, we perform a form of protocol analysis (Ericsson & Simon, 1993), analyzing the speech produced by the scientists to make inferences about the underlying cognitive processes.

The reason for focusing on pairs of scientists rather than on an individual scientist is that dyads produce speech naturally as part of their data analysis activities. By contrast, forcing an individual scientist to give a think-aloud protocol may change the very representations that we seek to study. For example, the individual scientist may change her focus to aspects of the data that are more easily verbalized, or she may change her representations from visio-spatial representations to more verbal representations.

Our methodology also contrasts with the retrospective analyses of historical cases from science (e.g., Gentner et al., 1997; Nersessian, 1985; Thagard, 1999). By focusing on the activities of non-famous (albeit expert) scientists working on a problem that may or may not lead to an important discovery, we may obtain a more representative view of how scientists reason.<sup>1</sup>

---

<sup>1</sup> Of course, if one's goal is to understand how large conceptual leaps are made in science, the historical case-study approach may be more fruitful.

Because our methodology is extremely labor-intensive, it lends itself most readily to case studies. However, the use of case studies always raises the question of generalizability: does the pattern found with these scientists at this particular time in this particular domain generalize to other scientists in other domains? To address this issue, we gathered data from two sets of scientists working in different disciplines on different kinds of problems. The first set of scientists was a pair of astronomers examining radio and optical data of distant galaxies. The second set of scientists was a pair of neuropsychologists examining fMRI imaging data of brain functioning under different experimental conditions. Thus we included both observational and experimental research from disciplines differing widely in types of training and age of the discipline. One should note, however, that both situations involved preliminary examinations of complex data visualizations presented on computer screens.

## Method

### Participants

The participants in the first domain were two expert astronomers, one a tenured professor at a university, the other a fellow at a research institute. The astronomers had earned their Ph.D.s six years and ten years respectively before this study; one has approximately 20 journal publications and the other approximately 10 in this area. One of the astronomers, hereafter referred to as A1, focuses on conducting and analyzing astronomical observations, and has an expertise in ring galaxies; the other, hereafter referred to as A2, combines teaching with primarily theoretical astronomical research and model construction. The astronomers have been collaborating for some years, although they do not frequently work physically alongside one another (i.e., work simultaneously at the same computer screen to examine data).

The participants in the second domain were two scientists in neuropsychology, one a postdoctoral researcher (B1) who has been in the field over 3 years, the other a graduate researcher (B2) who has been in the field for 1 year. The scientists work in a renowned national US research institute and are involved in developing a new methodology for analyzing fMRI brain data. They frequently work simultaneously at the same computer screen to examine data.

### Procedure

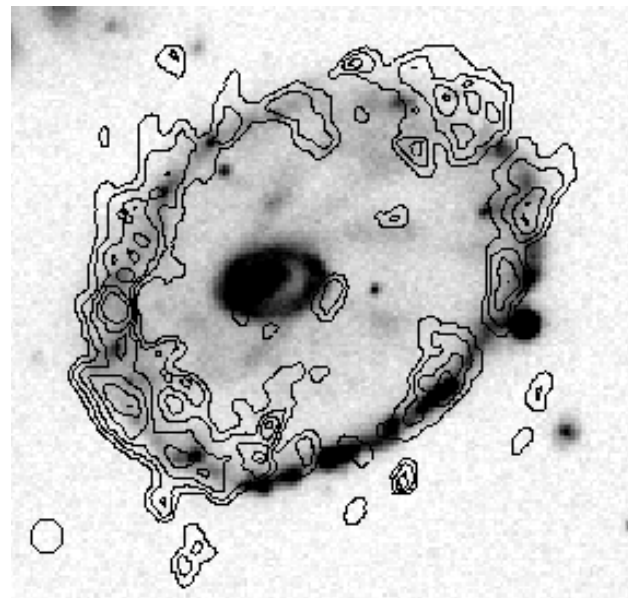
In both studies, the scientists were video- and audio-taped as they explored computer-generated visual representations of a new set of data. For the first study, A1 was in charge of the keyboard and mouse and sat directly in front of the screen; A2 sat slightly to his left. For the second study, B2 was in charge of the keyboard and mouse and sat directly in front of the screen while B1 sat slightly to her right. In both studies, all scientists had the shared monitor in their clear line of sight. They were instructed not to explain or interpret their comments to the researchers, but to carry out their work as though no camera were present. For each study, the relevant part of the session lasted about 1 hour. The scientists' interactions were transcribed and coded as described below. At a later date, we interviewed the scientists in both domains in order to obtain clarification of some domain-related issues.

### The Tasks and the Data

The astronomical data under analysis were optical and radio data of a ring galaxy. A ring galaxy forms as the result of a collision between two galaxies, and such collisions are relatively frequent cosmic events; consequently, ring galaxies *per se* are not uncommon. Both astronomers had conducted research and published scholarly articles on other ring galaxies, but this particular galaxy was relatively new to them. Nor had they examined this data set before; consequently, they considered this session exploratory.

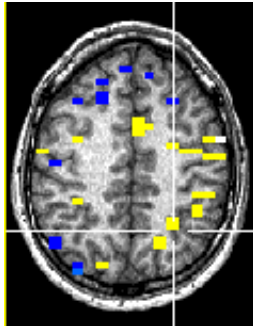
The astronomers' high-level goal was to understand the evolution and structure of the ring galaxy, by a complex sequence of inferences that began with interpreting contour lines on the display in terms of the 3-dimensional flow of gas in the galaxy. The astronomers' task was made difficult by two characteristics of the data: First, the data were one- or at best two-dimensional, whereas the structure they were attempting to understand was three-dimensional. Second, the data were noisy, and there was no easy way to distinguish between noise and real phenomena. Figure 1 shows a screen snapshot of the type of data they were examining.

Figure 1. Example of data examined by astronomers.  
Radio data (contour lines) are laid over optical data.



The fMRI data were obtained to understand how activation patterns inside the brain would change when people are anticipating some events to happen. There were two experimental conditions and one control condition. The scientists had to examine the data and compare them across the three conditions. This was the first time they had conducted the experiment and examined the data. The session was considered exploratory. Figure 2 shows an example of the fMRI data that they were analyzing. Similar to the astronomical data, fMRI data are inherently noisy and can only be displayed in two dimensions although the activation patterns under analysis were mostly three dimensional.

Figure 2. Example of fMRI data (color removed).



### Coding Scheme

The protocols were divided into 829 (astronomy) and 370 (fMRI) segments. As each scientist spoke in turn, a new segment was established. Then the scientists' individual utterances were further segmented by complete thought.

A coding scheme was developed to examine how the scientists explored the data. The entire astronomy protocol was coded independently by 2 different coders in order to establish the reliability of this scheme. Inter-rater reliabilities for each code are reported below. Because we found high agreement in coding the astronomy protocol, we expect the agreement in the neuropsychology protocol to be high also.

**On/Off Task** In order to allow us to focus our analysis only on those utterances that were relevant to the scientists' task of data analysis, we coded each segment as on-task or off-task. All segments that addressed matters external to the data analysis were coded as off-task; these segments included external interruptions (e.g., the telephone ringing), extraneous comments by the scientists (e.g., jokes or banter between them), comments relating to the software, specific details about plans for future observations, and so on. All segments that addressed issues of data analysis were coded as on-task. These included comments relating to the selection of a display type (as opposed to comments about how to implement that display) as well as decisions about obtaining additional data in the future (as opposed to details about how to obtain those data). Initial agreement between the coders was 90%. All disagreements were resolved by discussion.

**Noticings** In order to establish which phenomena the sci-

entists attended to, we first coded for the scientists' *noticing* phenomena in the data or features of the display. A noticing could involve merely some surface feature of the display, such as a line, shape, or color, or it could involve some interpretation by the scientists, for example, identifying an area of star formation or concentration of gas for the astronomers or activation in a particular area of the brain (e.g. thalamus) for the neuropsychologists. Only the first reference to a phenomenon was coded as a noticing; coding of subsequent references to the same phenomenon is discussed below. Agreement between the coders was 95%. Disagreements were resolved by discussion.

Because our investigation focused on the change in representation of anomalies in the data, we further coded these noticings as either "anomalous" or "expected," according to one or more of the following criteria: a) in some cases the scientists made explicit verbal reference to the fact that something was anomalous or expected; b) if there was no explicit reference, domain knowledge was used to determine whether a noticing was anomalous or not; c) a phenomenon might be associated with (i.e., identified as either like) another phenomenon that had already been established as anomalous or not; d) a phenomenon might be contrasted with (i.e., identified as unlike) a phenomenon that had already been established as anomalous or not; e) the scientists might question a feature, thus implying that it is unexpected. Table 1 illustrates these codes. Agreement between the coders was 87%. Those noticings for which disagreement could not be resolved were excluded from further analysis.

**Subsequent References** Our investigation focused on the astronomers' representation of phenomena over time. Whereas the coding of the noticings captured the first reference the astronomers made to a phenomenon of interest, we also needed to establish how they made subsequent reference to each noticing. Consequently, all subsequent references to each phenomenon were also identified.

Because the scientists were sharing a computer monitor, frequently the first interaction between them after a noticing was to establish that they were both looking at the same thing. Subsequent references that served purely to establish identity were *not* included in the analyses.

Not all subsequent references immediately followed a noticing; frequently, the scientists returned to a phenomenon of interest after investigating other features of the data. The

**Table 1.** Noticings (italicized) coded as unusual or expected.

Criterion	Code	Example - Astronomy	Example - fMRI
Explicit	Anomalous	What's that <i>funky thing</i> ...That's odd	Bunch of <i>stuff here</i> ... Yeah, that's weird
Domain Knowledge	Expected	You can see that all the <i>H1</i> is concentrated in the ring	So there is a <i>subcortical activation</i> that is probably caudate.
Association	Anomalous	You see similar kinds of <i>intrusions</i> along here	So there's the <i>thing</i> we've been seeing consistently.
Contrast	Expected	That's odd...As opposed to <i>these things</i> , which are just the lower contours down here	So <i>it's lateral</i> , which means its not in the midline... on our incentive task we see midline, but not lateral, so that's why that's not a spot.
Question	Anomalous	I still wonder why we don't see any <i>H1</i> up here in this sort of northern ring segment?	[None found]



scientists made frequent gestures to the feature of the image under discussion; by constructing a map of the noticings, and cross-referencing it with these gestures, the coders were able to determine the specific noticing to which a subsequent reference referred. Tables 2a and 2b illustrate the coding scheme for subsequent references in each domain.

**Entity Coding** To investigate the initial and changing representations of the phenomena the scientists noticed, we first identified what characteristics of each noticing (anoma-

Table 2a. Subsequent references in astronomy domain.

Noticing: First reference to phenomenon  
 Establish identity: Reference excluded from analysis  
 SR: Subsequent reference included in analysis

Code	Utterance
Noticing (N9)	A1: What's that funky thing...
Establish identity	A2: Left center, you mean...
Establish identity	A2: This stuff? [points to screen]
Establish identity	A1: Yeah
Establish identity	A2: Yeah
SR to N9	A1: What is that? A2: You can see there is some gas here [points to different area] inside the ring, but not much...
Noticing (N10)	
SR to N9	A1: Except for that little knot there.

Table 2b. Subsequent references in neuropsychology domain.

Code	Utterance
Noticing (N23)	B1: There, did you see that?
Establish identity	B2: Yah did you see that? [points to screen]
SR to N23	B1: That was near the thalamus.
SR to N23	B1: That might be spurious.
Noticing (N24)	B2: So the z-score of that one is 4.22.
SR to N24	B1: It's right up there [points to the threshold on screen]

lous and expected) first caught the scientists' attention. We then noted what characteristics the scientists attended to in their subsequent references to each noticing. We coded each noticing and subsequent reference as either "formal" or "informal" as follows. Formal references are those for which the scientists referred to the underlying phenomenon, using the terminology of the domain—for example, to a specific gas, star formation, the stellar continuum, or the like in the astronomy domain, and to the thalamus or caudate nucleus, for example, in the neuropsychology domain. Informal references include references to some generic feature of the display, such as a blob, a bulge, or a "dipsy-doodle" in the astronomy domain, and a neuron "lighting up" in the neuropsychology domain. They also include references to a phenomenon by its location (e.g., "lower right," "northwest") and anaphoric references, (e.g., pronouns). A few references combined characteristics of more than one code (e.g., "big blob of H1" combined the informal reference to a "blob" with the formal reference to H1 gas). Such references were coded as "mixed" references, and were excluded from subse-

quent analysis. Coder agreement on this coding was 100%.

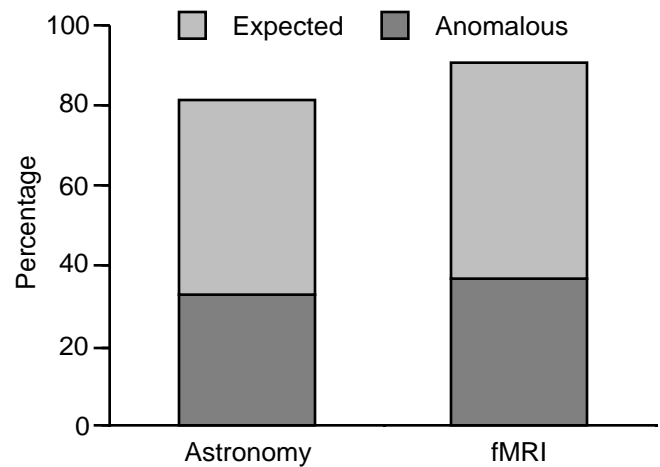
## Results and Discussion

There were 619 (75%) (astronomy) and 317 (85%) (neuropsychology) on-task segments. Subsequent analyses do not include off-task segments.

### Noticing Anomalies and Expected Phenomena

There were 27 (astronomy) and 35 (fMRI) noticings. In the astronomy data, 9 (33%) were anomalous, 13 (48%) were expected, and 5 (19%) were uncoded because either the astronomers or the coders disagreed. In the fMRI data, 13 (37%) were anomalous, 19 (54%) were expected, and 3 (9%) were uncoded. Figure 3 shows that the proportion of anomalous and expected noticings was similar in each dataset. Uncoded noticings were excluded from subsequent analysis.

Figure 3. Percentage of anomalous and expected noticings.



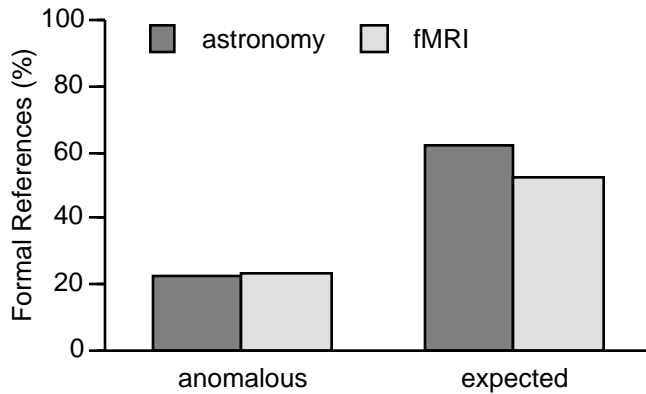
### Representation of Noticings

Our first question concerned how the scientists initially represented the phenomena they investigated. In the astronomy domain, 8 of the 13 (62%) expected phenomena were first identified by formal references and the remaining 5 (38%) by informal references. In contrast, most of the anomalies (78%) were initially identified by informal references, with only 2 of the 9 anomalies (22%) identified formally. Interestingly, both formal references were negative—they referred to the *absence* of the astronomical phenomenon (e.g., "I still wonder why we don't see any H1 up here.") A similar pattern was observed in the neuropsychology domain. Ten of the 19 (53%) expected phenomena were first identified by formal references, and 9 (47%) by informal references. Ten of the 13 (77%) anomalies were identified informally, with 3 (23%) identified by formal references. Again, 2 of the 3 formal references were negative (e.g. "There's nothing on the thalamus either, that's surprising"). Figure 4 shows the percentage of formal references to these initial noticings.

Thus it appears that in general, the scientists initially represented the expected phenomena in the formal, scientific terminology of that domain. However, their initial representations of unexpected or anomalous features of the data were

highly informal. Recall that these informal references were based primarily on irregular features of the display rather than the underlying phenomena that these features represented. Occasionally, it was the *absence* of a phenomenon that first drew the scientists' attention to these anomalies.

Figure 4. Percentage of formal references to initial noticings (anomalous and expected) in two domains.



### Local Changes in Representation

Next, we examined whether the scientists' representation of these phenomena changed as their investigation of the data progressed. The analyses that follow depend on the subsequent references to the noticings. In order to ensure a sufficient basis on which to judge change, we include only those noticings that received more than the mean number of subsequent references (i.e., more than 8 subsequent references for the astronomy and more than 3 for the fMRI data).

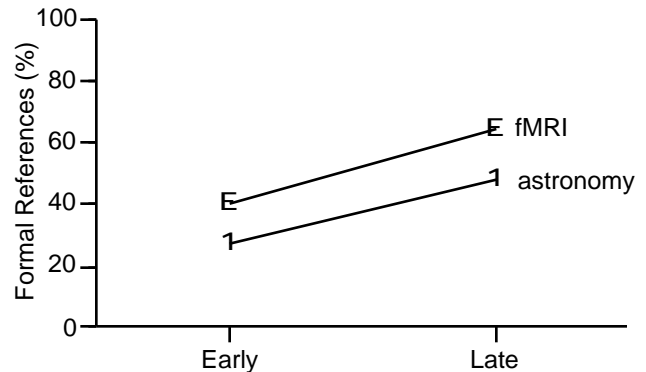
Five of the noticings in the astronomy data and 15 in the fMRI data received more than the mean number of subsequent references. In the astronomy data, the subsequent references to these 5 noticings account for 66% of all segments that made any reference to a phenomenon noticed by the astronomers. In the fMRI data, the subsequent references to these 15 noticings account for 69% of all such segments. Thus by confining our analyses to these 20 noticings, we focus on the majority of the data. It should also be noted that, because in general expected phenomena received little further attention, especially in the astronomy domain (Trickett, Trafton, & Schunn, 2000), most noticings included in these analyses are anomalies.

In order to examine change over time, we divided the period of attention to each individual noticed object into two phases, early and late. We tallied the total number of subsequent references for each and divided it by 2. For noticed objects with an odd number of subsequent references, we discarded the midpoint reference, to insure an even split. We then compared the numbers of formal and informal references in the early and late phases of the scientists' investigation.

In the astronomy data, in the early phase of investigation, 17 of the 63 (27%) subsequent references were formal compared with 30 (48%) in the later phase. By contrast, 39 (61%) of the subsequent references were informal in the early phase compared with 25 (39%) in the later phase,  $\chi^2(1) = 6.65, p < .01$ . (These percentages do not sum to 100% because of the mixed references excluded from the analysis.) In

the fMRI data, in the early phase of investigation, 18 of the 45 (40%) subsequent references were formal compared with 29 (64%) in the later phase. By contrast, 27 (60%) subsequent references were informal in the early phase compared with 16 (36%) in the later phase,  $\chi^2(1) = 5.39, p < .05$ . Thus, in both domains, the number of formal representations increased, while the number of informal representations decreased. Figure 5 shows the increase in formal references in the later phase.

Figure 5. Changes in representation of noticed objects.



These results show that the scientists' representations changed significantly over time, as they investigated these anomalies. In the early phase of analysis, their representations were informal and display-based, most likely because they did not have a precise understanding of the phenomenon under investigation. The scientists needed a label by which they could identify, discuss, and refer to the phenomenon, and this label tended to be based on the visual appearance of the feature. As their investigation proceeded, however, these visually-based labels decreased. The reduction in display-based and anaphoric references suggests that the scientists became more specific, and points to an increased understanding of these anomalous phenomena.

### Global Changes in Representation

It is possible that the shift toward a formal representation occurred not because the scientists' understanding of individual anomalies increased, but because their global understanding of the data increased over time. In order to investigate this possibility, we divided each entire analysis session into early and late phases, based on overall time spent. Thus, in this analysis, there were unequal numbers of reference in each phase, but the time spent on each phase was the same.

We counted the number of formal and informal references to these well-referenced phenomena in each phase. In the astronomy protocol, 63% of the references in the early phase were informal, compared with 45% in the late phase; 35% of the early references were formal, compared with 39% in the late phase. This difference was not significant,  $\chi^2(1) = 1.4, p > .2$ . Although the proportion of informal references did drop off, the number of formal references remained constant. In the fMRI protocol, 53% of the references in the early phase were informal, compared with 52% in the late phase; 47% of the early references were formal, compared with 48%

in the late phase. This difference was also non-significant,  $\chi^2(1) < 1$ . Thus, it does not appear that the shift toward a more formal representation occurred as a result of a more general, global understanding of the data.

### General Discussion

Our results show that both groups of scientists initially represented expected and anomalous phenomena quite differently. Whereas they represented the expected phenomena in the formal terms appropriate to their domain of expertise, they represented the anomalous phenomena in highly informal terms that referred to salient features of the visual data. These results also show that these internal representations changed over time, shifting from informal to formal representations. However, this shift in representation did not appear to be caused by a global increase in understanding of the data under analysis, but was instead local, and associated with the individual phenomena under investigation. This shift in representation appears to have affected primarily the scientists' representation of anomalous or unexpected findings in their data. We have investigated elsewhere the key role of anomalies in the exploratory stages of data analysis (Trickett, Trafton & Schunn, 2000).

Our focus in this study has been on the exploratory stages of data analysis. We believe that including two independent data analysis sessions in quite different scientific domains strengthens our claims about these changes in representation. However, clearly we need to ascertain whether our results generalize to other situations and scientific domains.

In this paper we examined fairly small changes in representation at the item-specific level. Much research in cognitive science on the topic of conceptual change has focused on relatively larger scale changes in representation (e.g., Chi, 1997; Thagard, 1999). One may wonder what the relationship is between the micro-level changes that we have reported in this paper and the more macro-level changes reported in the conceptual change literature. Some researchers (e.g., Chi, 1997) have speculated that some macro-level changes are not the result of many small changes; instead Chi has argued that some macro-level changes are the result of a complete conceptual reorganization. Similarly, some historians of science have noted that some scientific changes appear to be more radical or revolutionary than others appear to be (e.g., Kuhn, 1967). We believe that the relationship between the micro-level changes in representation and the macro-level changes that are thought to constitute conceptual change remains an open question, and that this question could be fruitfully studied by observing the same scientists over a much longer time scale than we have done so far. We are currently planning such longitudinal studies.

### Acknowledgments

This research was supported in part by student fellowships from George Mason University to the first and second authors and by grant number 55-7850-00 from the Office of Naval Research to the Naval Research Laboratory. We thank Georgia Seeley and Audrey Lipps for assistance with coding.

### References

- Chi, M. T. H. (1997). Creativity: Shifting across ontological categories flexibly. In T. B. Ward & S. M. Smith (Eds.), *Creative thought: An investigation of conceptual structures and processes*. Washington, DC, USA: American Psychological Association.
- Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, 17(3), 397-434.
- Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. E. Sternberg & J. E. Davidson, (Eds.), *The nature of insight*. Cambridge, MA, USA: MIT Press.
- Dunbar, K. (in press). What scientific thinking reveals about the nature of cognition. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implications from everyday, classroom, and professional settings*. Mahwah, NJ: Erlbaum.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.
- Gentner, D., Brem, S., Ferguson, R. W., Markman, A. B., Levidow, B. B., Wolff, P., & Forbus, K. D. (1997). Analogical reasoning and conceptual change: A case study of Johannes Kepler. *Journal of the Learning Sciences*, 6(1), 3-40.
- Klahr, D. & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin*, 125(5), 524-543.
- Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology*, 17(2), 248-294.
- Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
- Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth 10,000 words. *Cognitive Science*, 4, 317-345.
- Nersessian, N. J. (1985). Faraday's field concept. In D. Gooding & F. James (Eds.), *Faraday rediscovered*. London: Macmillan.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Schunn, C. D., & Klahr, D. (1995). *A 4-space model of scientific discovery*. In the proceedings of the 17th Annual Conference of the Cognitive Science Society.
- Thagard, P. (1999). *How scientists explain disease*. Princeton, NJ: Princeton University Press.
- Trafton, J. G., Kirschenbaum, S. S., Tsui, T. L., Miyamoto, R. T., Ballas, J. A., & Raymond, P. D. (under review). Turning Pictures into Numbers: Use of Complex Visualizations.
- Trickett, S. B., Trafton, J. G., & Schunn, C. D. (2000). *Blobs, dippy-doodles, and other funky things: Framework anomalies in exploratory data analysis*. In the proceedings of the 22nd Annual Conference of the Cognitive Science Society.
- Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science*, 18(1), 87-122.

# Blobs, Dipsy-Doodles and Other Funky Things: Framework Anomalies in Exploratory Data Analysis

**Susan B. Trickett**  
([stricket@gmu.edu](mailto:stricket@gmu.edu))  
Department of Psychology  
George Mason University  
Fairfax, VA 22030 USA

**J. Gregory Trafton**  
([trafton@itd.nrl.navy.mil](mailto:trafton@itd.nrl.navy.mil))  
Naval Research Laboratory  
NRL Code 5513  
Washington, DC 20375

**Christian D. Schunn**  
([schunn@gmu.edu](mailto:schunn@gmu.edu))  
Department of Psychology  
George Mason University  
Fairfax, VA 22030

## Abstract

This study investigates the role of anomalies in the exploratory analysis of visual scientific data. We found that anomalies played a crucial role as two experts analyzed astronomical data. Not only did they pay significantly more attention to anomalies than expected phenomena, both immediately and over time, but also anomalies provided a framework within which they investigated the data.

## Introduction

Attention to the unexpected may be an important component of scientific discovery. Exploring anomalies can lead to theory development and even conceptual change. Philosophers of science (e.g., Kuhn, 1962) have argued that unusual findings play a key role in scientific revolutions, and scientists themselves have claimed that investigating anomalies lies at the heart of scientific innovation (e.g., Knorr, 1980).

Within cognitive psychology, response to anomalous data during scientific inquiry has been noted in a variety of studies, including historical reconstructions of actual scientific discoveries (e.g., Kulkarni & Simon, 1988), on-line studies of scientists (e.g., Dunbar, 1997), laboratory studies in which participants “rediscover” a scientific phenomenon (Dunbar 1993), and studies of those with little scientific training as they perform abstract scientific reasoning tasks (e.g., Tweney, Dowerty, & Mynatt, 1982; Klahr & Dunbar, 1988). These studies have not yielded a consistent pattern of response to unexpected data, possibly because of the range of scientific training and knowledge among the participants.

Recognizing this variety of responses to anomalous data, Chinn and Brewer (1992, 1993), propose a taxonomy of seven reactions to unusual findings, from ignoring the data and upholding the theory to accepting the data and changing the theory. This taxonomy is derived from anecdotal examples from the history of science and from empirical studies of scientific reasoning in the psychological literature. Although Chinn and Brewer propose that this taxonomy applies to scientists and non-scientists alike, they have tested it only among undergraduates with little scientific training.

Thus, despite the general belief that anomalous data is important in scientific discovery, no clear picture has emerged of how scientists (as opposed to laypersons performing scaled-down scientific discovery tasks) respond to unexpected findings. On one hand, there is a well-established tradition in studies of scientific thinking that shows people overlook data inconsistent with their hypothesis, looking

only for support for their theories (e.g., Wason, 1960). Within this tradition, scientists have been found to be as susceptible to this confirmation bias as laypeople (e.g., Mahoney & DeMonbreun, 1977; Mitroff, 1974.) Similarly, studies of complex visualization usage have shown that expert meteorologists do not pay much attention to unusual or anomalous features. Instead, they seem to extract information in a very goal directed manner, rarely following up on features that are not directly relevant to their immediate task (Trafton et al., under review). This evidence—of confirmation bias, even among scientists, and of the goal-directed nature of complex visualization usage—suggests that scientists may overlook unexpected results or anomalies.

On the other hand, however, Dunbar has recently questioned the validity of the studies of confirmation bias on the grounds that they employ arbitrary experimental tasks that involve no scientific knowledge and therefore bear little relationship to tasks that real scientists perform (Dunbar, 1997). Dunbar has argued that in order to investigate how scientists reason, one must observe scientists as they perform their scientific tasks.

Using an “in vivo” methodology that involves observing actual scientists at work, Dunbar has suggested that scientists do attend to unusual results (Dunbar, 1997). He found not only that scientists attended to unexpected results more than they did to expected findings, but also that individual scientists were quick to discard a hypothesis when faced with results that were inconsistent with it. Furthermore, he noted that in lab meetings, the group of scientists tended to focus on a surprising result until they had constructed a plausible hypothesis to account for it. Dunbar concluded that attending to anomalous findings is an important strategy that contributes to successful scientific inquiry (Dunbar 1997). Similarly, Kulkarni and Simon (1988) identified an “attend to surprising result” heuristic as crucial to Hans Krebs' discovery of the urea cycle.

Both Chinn and Brewer's and Dunbar's studies have involved participants, whether trained scientists or not, who were evaluating data to test a specific theory. However, there are many phases of scientific inquiry, and response to anomalous data might be quite different during an exploratory phase from when a theory is firmly established. During exploratory data analysis, theories may be only partially defined. Nonetheless, given their extensive domain knowledge, scientists doubtless have general frameworks which lead to expectations that may or may not be met by the data. They may therefore pay more attention to unusual results,

because such framework anomalies may provide insights for interpreting data and developing theories.

Similarly, there are many forms of data, but previous studies have focused on data that were either presented textually or required direct, relatively simple perceptual judgments. However, scientists in many domains employ complex visualization techniques in order to inspect their data. Little is known about the role of unusual or unexpected findings in either exploratory or scientific visualization.

Our goal is to investigate the role of anomalies during early, primarily exploratory phases of visual data analysis. Specifically, we investigate whether scientists notice anomalies in this type of data and, if so, the extent to which they attend to them, both immediately and over time. We also investigate how the visual nature of the data affects the detection of and attention to anomalies.

There are many methodologies available by which to examine the processes of scientific inquiry, and there are strengths and weaknesses associated with each (see Klahr & Simon, 1999 for a review). Our approach has been to combine features of several methodologies in order to take advantage of their respective strengths.

First, we have chosen to conduct a case study of actual scientists at work because this methodology offers an extraordinarily rich set of observations of high face validity. Most case studies of scientific inquiry have focused on famous scientists who have made discoveries of great historical importance (e.g., Gentner et al, 1997; Kulkarni & Simon, 1988). We have chosen instead to focus our investigation on more “ordinary”—albeit expert—scientists, the ultimate significance of whose work is currently unknown. We believe this focus on the more mundane aspects of scientific inquiry may yield results that are more representative of scientists’ everyday activities.

Second, we collected verbal and visualization data of the scientists working together and conducted a verbal protocol analysis of these data in order to gain insight into the scientists’ concurrent thought processes (Ericsson & Simon, 1993). Verbal protocols have frequently been collected in laboratory studies of non-scientists performing scientific discovery tasks; however, this methodology has rarely been used with practicing scientists. Furthermore, because we collected a protocol of a work session involving two scientists, there was no need for an experimenter to prompt the participants to keep talking. By focusing on a dyad, we were able to obtain a more natural account of the scientists’ thinking than is possible with an individual.

Finally, we have adopted Dunbar’s (1995, 1997) “in vivo” methodology because, as Dunbar points out, it affords a unique opportunity to observe “how scientists really reason.” Instead of observing a lab group as Dunbar did, however, we chose to study a dyad, for two reasons. First, the two scientists we observed were of equal professional status, thus we avoid social issues that might make junior scientists reluctant to question the interpretations of a senior colleague. Second, we think that the verbal protocols of a dyad might represent each scientist’s thinking more completely than those of a group. In a group setting, with more people “jumping into” a discussion, individuals may be less likely to pursue lines of thought in significant depth.

## Method

### Participants

The participants in this study were two expert astronomers, one a tenured professor at a university, the other a fellow at a research institute. The astronomers had earned their Ph.D.s six years and ten years respectively before this study; one has approximately 20 publications in this general area and the other approximately 10. One of the astronomers, hereafter referred to as A1, focuses on conducting and analyzing astronomical observations, and has an expertise in ring galaxies; the other astronomer, hereafter referred to as A2, combines teaching with primarily theoretical astronomical research and model construction. The astronomers have been collaborating for some years, although they do not frequently work physically alongside one another (i.e., work at the same computer screen at the same time to examine data).

### Procedure

The astronomers were video- and audio-taped as they explored computer-generated visual representations of a new set of observational data. They were working in one astronomer’s office at a shared computer monitor. One astronomer was in charge of the keyboard and mouse and sat directly in front of the screen; the other astronomer sat to his left, with the monitor clearly in view. They were instructed not to explain their comments to the researchers, but to carry out their work as though no camera were present. The relevant part of the session lasted about 53 minutes and generated 7676 words. The astronomers’ interactions were later transcribed and coded as described below. At a later date, we interviewed A2 to obtain clarification of domain-related issues.

### The Task and the Data

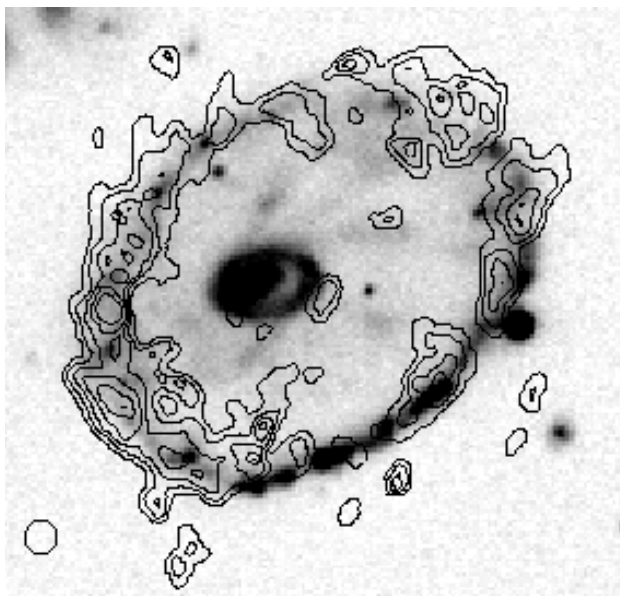
The astronomical data under analysis were optical and radio data of a ring galaxy. A ring galaxy forms as the result of a collision between two galaxies: one galaxy is thought to have passed through another, leaving both a doughnut-shaped ring of stars and gas (the ring galaxy) and a smaller galaxy nearby. Such galactic collisions are relatively frequent cosmic events; consequently, ring galaxies *per se* are not uncommon. Both astronomers had conducted research and published scholarly articles on other ring galaxies, but this particular galaxy was relatively new to them. Nor had they examined this data set before; consequently, they considered this session exploratory.

The astronomers’ high-level goal was to understand the evolution and structure of the ring galaxy. This understanding emerges from an understanding of where, how, and why star formation occurs within the galaxy, which rests on an understanding of the flow of gas in the galaxy. In order to understand the flow of gas, the astronomers must understand the kinematics (the velocity and position) of the system, by inferring the 3-dimensional streaming motions of the gas. They make inferences about streaming motions by interpreting the velocity field, represented by contour lines on the 2-dimensional display. Examining the velocity contours is thus the lowest level task in this chain of inferences.

The astronomers’ task was made difficult by two characteristics of their data. First, the data were one- or at best two-

dimensional, whereas the structure they were attempting to understand is three-dimensional. Second, the data were noisy, and there was no easy way to distinguish between noise and real phenomena. Figure 1 shows a screen snapshot of the type of data the astronomers were examining. In order to make their inferences, the astronomers used different types of image, representing different phenomena (e.g., different forms of gas), which represent different information about the structure and dynamics of the galaxy. Some of these images could be overlaid on each other. In addition, the astronomers could choose from images created by different processing algorithms that result in different weightings of the data, each with advantages and disadvantages (e.g., more or less resolution). Finally, they could adjust different features of the display, such as contrast or false color.

Figure 1: Example of data examined by astronomers. Radio data (contour lines) are laid over optical data.



### Coding Scheme

The protocol was divided into 829 segments: as each astronomer spoke, a new segment was coded; then their utterances were further segmented by complete thought.

A coding scheme was developed to examine the astronomers' attention to anomalous phenomena in the ring galaxy. The protocol was coded independently by two different coders. Inter-rater reliabilities for each code are reported below.

**On/Off Task** In order to allow us to focus our analysis only on those utterances relevant to the scientists' task of data analysis, we coded each segment as on- or off-task. All segments that addressed matters external to the data analysis were coded as off-task; these segments included external interruptions (e.g., the telephone ringing), extraneous comments by the astronomers (e.g., jokes or banter between them), comments relating to the software, specific details about plans for future observations, and so on. All segments

that addressed issues of data analysis were coded as on-task. These segments included comments relating to the selection of a display type (as opposed to comments about how to implement that display) as well as decisions about obtaining additional data in the future (as opposed to details about how to obtain those data). Initial agreement between the coders was 90%. All disagreements were resolved by discussion.

**Episodes** Next, we divided the protocol into discrete, non-overlapping episodes that would allow us to study the astronomers' shifting focus of attention. The protocol was segmented into 19 exhaustive episodes. An episode began with the astronomers' focus on a feature or point of discussion and lasted until their attention switched to another phenomenon or theoretical point; at this switch of attention, a new episode was coded. Although the focus of most episodes was a feature of the galaxy, this was not necessarily the case; for example, one episode consisted of a discussion about a future observation session and the data to be obtained from it. Agreement between coders was 98%.

A new episode frequently coincided with a display change, but did not necessarily do so. Sometimes the astronomers switched their focus of attention to another galactic feature visible on the same display, thus beginning a new episode without changing the display. At other times, they changed the display in order to explore another representation of a feature, thus changing the display within the same episode.

**Noticings** In order to establish which phenomena—unusual or not—the astronomers attended to, we first coded for the astronomers' *noticing* phenomena in the data. A noticing could involve merely some surface feature of the display, such as a line, shape, or color, or it could involve some interpretation by the astronomer, for example, identifying an area of star formation or concentration of gas. Only the first reference to a phenomenon was coded as a noticing; coding of subsequent references to the same phenomenon is discussed below. Agreement between the coders was 95%. Disagreements were resolved by discussion.

Table 1: Noticings (in italics) coded as unusual or expected

Criterion	Code	Example
a) Explicit	Anomalous	What's <i>that funky thing...</i> That's odd
b) Domain Knowledge	Expected	You can see that <i>all the HI</i> is concentrated in the ring
c) Association	Anomalous	You see <i>similar kinds of intrusions</i> along here
d) Contrast	Expected	That's odd...As opposed to <i>these things</i> , which are just the lower contours down here
e) Question	Anomalous	I still wonder why <i>we don't see any HI up here</i> in this sort of northern ring segment?

**Subsequent References** One of our questions was the

extent to which the astronomers continued to investigate anomalies. Whereas the coding of the noticings captured the first reference the astronomers made to a phenomenon of interest, we needed to establish how frequently they made subsequent reference to each noticing. Consequently, all subsequent references were also identified and coded.

Because the astronomers were sharing a computer monitor, frequently the first interaction between them after a noticing was to establish that they were both looking at the same thing. Subsequent references that served purely to establish identity were *not* included in the analyses.

Table 2: Coding of subsequent references  
 Noticing: First reference to phenomenon  
 Establish identity: Reference excluded from analysis  
 SR: Subsequent reference included in analysis

Code	Utterance
Noticing (N9)	A1: What's that funky thing...
Establish identity	A2: Left center, you mean...
Establish identity	A2: This stuff? [points to screen]
Establish identity	A1: Yeah
Establish identity	A2: Yeah
SR to N9	A1: What is that? A2: You can see there is some gas here [points to different area] inside the ring, but not much...
Noticing (N10)	
SR to N9	A1: Except for that little knot there.

Not all subsequent references immediately followed a noticing; frequently, the astronomers returned to a phenomenon of interest after investigating other features of the galaxy. The astronomers made frequent gestures to the feature of the image under discussion; by constructing a map of the noticings on the galaxy, and cross-referencing it with these gestures, the coders were able to determine the specific noticing to which a subsequent reference referred. Table 2 illustrates the coding scheme for (sequential) subsequent references.

## Results and Discussion

There were 619 on-task segments (75%). Subsequent analyses do not include off-task segments.

### Noticing Framework Anomalies

Our first question was did the astronomers notice anomalies in the data? Recall that a "noticing" is a first-time reference to a phenomenon of interest. There were 27 noticings during this session. Of these, 9 (33%) were anomalous, 13 (48%) were expected, and 5 (19%) were uncoded, because the astronomers themselves or the coders disagreed. This analysis shows that at least one-third of the phenomena the astronomers identified were unusual in some way. It appears then that the astronomers *did* notice anomalies in this dataset.

Interestingly, most of the anomalies (78%) were identified in highly informal terms or by features of the display, rather than by underlying astronomical phenomena. Thus, the astronomers usually identified anomalous phenomena as "blobs," "bulges," or "dipsy-doodles" rather than in formal astronomical terms (such as a specific type of gas). Not only

were anomalies important to the astronomers, but their attention to these anomalies appears to be drawn primarily by visual features of the data. We investigate the relationship between representation and anomalous/expected results elsewhere (Trickett, Fu, Schunn, & Trafton, 2000).

### Relationship between Episodes and Noticings

Next, we investigated whether the anomalies played any part in guiding or structuring this exploratory session, that is, whether there was any relationship between the noticings and the episodes, and if so, whether this relationship was different for the anomalies than for the expected phenomena.

In order to investigate this relationship, we noted how each episode began. Nine of the 19 episodes began with a noticing, 7 began with a subsequent reference, and 3 episodes began with something other than a noticing or subsequent reference to a noticing. Thus, out of 19 episodes, only 3 were initiated by theoretical or other similar considerations. Noticing and subsequent references, while common, only account for 61% of the segments. Thus episodes are more likely to start with a data-driven event (noticing or SR) than one would expect from the base rates,  $X^2(1) = 3.88$ ,  $p < .05$ . This result suggests that the most likely focus of attention was some feature of the data rather than some theoretical or other matter. This exploratory session analyzing visual data appears to have been driven primarily by the data themselves rather than theoretical considerations.

What features of the data were likely to attract the astronomers' attention? Of the noticings that sparked an episode, an equal number (3) were anomalous and expected. But whereas *no* episodes began with a subsequent reference to an expected phenomenon, 6 subsequent references to anomalies launched a new episode. This analysis suggests that at a first pass, the astronomers were equally likely to attend to expected as to anomalous phenomena in the data. However, as they explored the data further, it was the anomalies, not the expected phenomena, that directed their investigations. Table 3 summarizes these results.

Table 3: Noticings and subsequent references beginning an episode

	Notice	SR
Anomalous	3	6
Expected	3	0

### Initial Attention to Anomalies

Once the astronomers had identified something unusual in the data, what did they do with this observation? There are several possible reactions: they could pursue the anomaly in order to try to account for it, they might temporarily disregard it but return to it later, or they might move on to explore some other, better understood, aspect of the data. A related question is whether their response to anomalies was different from their response to expected phenomena.

First, we investigated this issue by considering the extent to which the astronomers made subsequent reference to a noticing immediately upon identifying it. If anomalies and expected phenomena are of equal interest, we would expect

them to make a similar number of subsequent references to both the anomalous and expected patterns. However, if anomalies play a more important role in their efforts to understand the structure of the galaxy, we would expect them to pay more attention (measured by the number of subsequent references) to anomalies than to expected observations.

Although there were fewer anomalies identified in this session, collectively these anomalies received over 3 times as many subsequent references within the same episode as the expected phenomena. The total number of subsequent references to anomalies was 68 (mean = 7.6), compared with 19 (mean = 1.5) for expected phenomena. A t-test on these data was significant,  $t(20) = 2.27$ ,  $SE = 2.69$ ,  $p < .05$ . These results show that the astronomers did pay more attention to the anomalies immediately upon noticing them, or soon thereafter, than they did to the expected phenomena. This in turn suggests that the anomalies were more important to the astronomers than those phenomena they expected to find. Table 4 summarizes the results of this analysis.

Table 4: Subsequent references (SRs) within episodes

	Total SRs	Mean SRs	Range
Anomalous (N=9)	68	7.6	1 - 30
Expected (N=13)	19	1.5	0 - 4

Furthermore, as Table 4 shows, the range of subsequent references was also much greater for the anomalies than for the expected phenomena. *All* the anomalies received at least one subsequent reference soon after the astronomers noticed it. In contrast, 5 of the 13 expected phenomena (38%) received no subsequent references, i.e., no immediate further attention. In addition, 5 of the 9 anomalies (56%) received more than 5 subsequent references; none of the expected phenomena was referred to so frequently. This analysis provides further support for our claim that overall, the anomalies were more important to the astronomers' goals than the expected phenomena. In addition, it suggests that the anomalies themselves were not of equal importance, with some anomalies receiving much more attention than others.

### Long-Term Attention to Anomalies

It appears, then, that as the astronomers explored the data about the ring galaxy, they paid more attention immediately to the anomalies in this data than they did to expected phenomena. These results say nothing, however, about the continuing role of the anomalies in the astronomers' analysis. Possibly, having explored an anomaly, the astronomers might "consider the matter closed" and switch their attention to another phenomenon. As with the astronomers' immediate attention to phenomena, we compare their treatment of anomalies with their response to the expected phenomena.

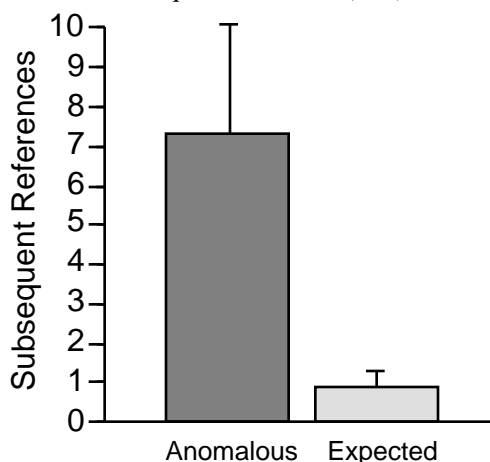
In order to investigate the extent to which the astronomers revisited the phenomena they noticed, we examined the number of subsequent references to both anomalies and expected findings *across* episodes. Recall that an episode ended when the astronomers switched attention to another feature or point of discussion. Thus, a reference to a feature across an episode indicates a switch of attention *back* to that fea-

ture, after having focused attention on something else.

One of the expected phenomena was first noticed in the last episode; because it was not possible for it to be referenced in a later episode. Thus the number of expected phenomena for these analyses is reduced from 13 to 12.

Seven of the 9 anomalies (78%) were referenced across episodes, compared with 6 (50%) of the expected phenomena. Overall, the total number of subsequent references across episodes to anomalies was 66 (mean = 7.3), compared with 11 (mean = 0.9) to expected phenomena. A t-test on these data was significant,  $t(19) = 2.66$ ,  $SE = 2.41$ ,  $p < .05$ . This result shows that the astronomers continued to pay more attention to anomalies than to expected phenomena, revisiting them even after switching their attention to other features of the data. Figure 2 summarizes these results.

Figure 2: Mean subsequent references (SRs) across episodes



Five of the 9 anomalies (56%) received more than 5 subsequent references across episodes. None of the expected phenomena was referenced so frequently. Furthermore, the astronomers persisted in returning to some anomalies, in 3, 4, 5, or even 6 episodes. The spread of episodes during which subsequent references were made was quite extensive and in several cases spanned almost the entire session. For example, Noticing 2 was first identified in episode 1 and was further referred to in episodes 2, 5, 6, 13, 15, and 17. Noticing 11 was first identified later in the session, in episode 9, and was further referenced in episodes 11, 13, 15, and 17. These results suggest that some anomalies were very puzzling to the astronomers and that they were sufficiently important to the exploration of the data that they returned to them repeatedly, even long after they had first noticed them.

### General Discussion and Conclusion

This study was conducted to investigate the role of anomalies in the exploratory stages of visual data analysis. We found that the astronomers did notice and pay attention to anomalies. They paid significantly greater attention to the anomalies in the data than to the expected phenomena. Furthermore, they found some anomalies sufficiently intriguing that they returned to them later in their exploration, in some cases repeatedly and over relatively long stretches of time. None of the expected phenomena received this type of prolonged attention. We conclude, therefore, that anomalies



played an important role in the exploration of these data.

In addition, we found that the astronomers' attention was initially drawn by features of the data rather than theoretical considerations. Although at first an expected phenomenon was as likely as an anomaly to become the focus of attention, as the analysis proceeded, the anomalies were more likely to hold the astronomers' attention. Furthermore, attention to the anomalies was initially drawn by irregular features of the visual representation rather than by the underlying phenomenon itself. This suggests that their approach was highly perceptual, because they identified anomalies primarily on the basis of unusual curves, lines, etc.

It is possible that anomalies played a significant role in this data analysis session *because* of the visual nature of the data. The anomaly was visible on the display at all times; consequently, it is possible that the astronomers were cued primarily by the display rather than memory to revisit the anomaly. However, this does not seem to be the case. If the display were the only means by which the astronomers were cued to make subsequent reference to the anomaly, we would expect them to make subsequent references to *all* anomalies. As our results indicate, though, they were selective in the anomalies they continued to investigate. Although visibility on the display may have helped to keep a particular anomaly activated in the astronomers' memory, this alone does not seem to have been sufficient to prompt them to revisit it. Rather, it appears that some anomalies were "tagged" as worthy of further investigation, and that the astronomers continued to search for a satisfactory way to explain them.

In contrast to the widely-held belief that scientists are susceptible to confirmation bias and seek chiefly to confirm what they already expect, our results present a picture in which investigating framework anomalies is a central activity in exploratory data analysis. We propose that the anomalies were instrumental in guiding the structure and content of the data analysis session.

We acknowledge that this is a case study of particular scientists in one domain, working at a specific phase of their research. However, our results are part of a growing body of evidence that attention to anomalies may be an important component of scientific inquiry (cf. Dunbar, 1997). Moreover, the scientists in this study were engaged in a task—exploratory data analysis—that is undertaken in all scientific domains. They neither employed unusual techniques nor used specialized equipment unique to their domain. We therefore expect our results to generalize to scientists in other domains. Whether or not they apply to later stages of data analysis (such as hypothesis-testing) remains an open question. We are currently extending this research by applying our methodology to a variety of scientists working with scientific visualizations in several domains. We are also planning to conduct longitudinal observations of scientists, in order to investigate the role of anomalies in their work at different stages of data analysis.

### Acknowledgments

This research was supported in part by a student fellowship from George Mason University to the first author and by grant number 55-7850-00 from the Office of Naval Research to the Naval Research Laboratory. We thank Georgia Seeley

and Audrey Lipps for coding assistance and Erik Altmann, Melanie Diez, Anthony Harrison, William Liles, and Lelyn Saner for comments.

### References

- Chinn, C. A., & Brewer, W. F. (1992). Psychological responses to anomalous data. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Chinn, C. A., & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching*, 35(6), 623-654.
- Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, 17(3), 397-434.
- Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & J. Vaid, (Eds.), *Creative thought: An investigation of conceptual structures and processes*. Washington, DC, USA: American Psychological Association.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.
- Gentner, D., Brem, S., Ferguson, R. W., Markman, A. B., Levidow, B. B., Wolff, P., & Forbus, K. D. (1997). Analogical reasoning and conceptual change: A case study of Johannes Kepler. *Journal of the Learning Sciences*, 6(1), 3-40.
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, 12(1), 1-48.
- Klahr, D. & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin*, 125(5), 524-543.
- Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94, 211-228.
- Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
- Kulkarni, D., & Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. *Cognitive Science*, 12(2), 139-175.
- Mahoney, M. J., & DeMonbreun, B. G. (1977). Psychology of the scientist: An analysis of problem-solving bias. *Cognitive Therapy and Research*, 3, 229-238.
- Mitroff, I. (1974). *The subjective side of science: A philosophical inquiry into the psychology of the Apollo moon scientists*. Amsterdam: Elsevier.
- Trafton, J. G., Kirschenbaum, S. S., Tsui, T. L., Miyamoto, R. T., Ballas, J. A., & Raymond, P. D. (under review). Turning Pictures into Numbers: Use of Complex Visualizations.
- Trickett, S. B., Fu, W-T., Schunn, C. D., & Trafton, J. G. (2000). From dippy-doodles to streaming motions: Changes in representation in the analysis of visual scientific data.
- Tweney, R. D., Doherty, M. E., & Mynatt, C. R. (Eds.). (1982). *On scientific thinking*. New York: Columbia University Press.
- Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12, 129-140.

# Reaction Times and Predictions in Sequence Learning: A Comparison

Ingmar Visser and Maartje E.J. Raijmakers and Peter C.M. Molenaar<sup>1</sup>  
{op\_visser, op\_raijmakers, op\_molenaar}@macmail.psy.uva.nl  
Developmental Processes Research Group  
Department of Psychology, University of Amsterdam  
Roetersstraat 15, 1018 WB Amsterdam  
The Netherlands

## Abstract

In the simple recurrent network (SRN) model, proposed by Cleeremans and McClelland (1991) to describe implicit sequence learning, the distinction between reaction time and prediction of the next trial is somewhat blurred. That is, the reaction time of the network is taken to be inversely proportional to the activation value of the corresponding node. In a prediction task the prediction would also be directly derived from the activities of the output nodes. In order to investigate the difference between ability to predict following stimuli and reaction times, we study implicit sequence learning in a similar vein as done by Cleeremans and McClelland (1991), using a slightly less complex grammar than they did. In addition we ask subjects to guess where the next stimulus will be at randomly chosen trials during the learning process. Results show a direct correspondence between fast reaction times and correct predictions.

## Introduction

Implicit learning has been studied for over thirty years starting with Reber (1967). Only recently attention has been given to modeling this kind of learning behavior in detail, mainly using neural networks. Specifically simple recurrent networks have been used successfully by Cleeremans and McClelland (1991) to model subjects' behavior on learning sequences that are generated by a finite state automaton, in fact the very same automaton that was used by Reber (1967).

Many different paradigms have been developed for studying implicit learning behavior. One characteristic that divides those paradigms is the way in which they assess the possession of implicit knowledge. In this paper two such measures, reaction times and predictions, are studied. In implicit learning research the sequential implicit learning paradigm has become increasingly popular and with that the use of reaction time as the primary measure of performance (see for example Nissen & Bullemer, 1987; Cleeremans & McClelland, 1991; Seger, 1997). We used an augmented sequence learning paradigm in which a direct comparison between reaction times and predictions was possible.

---

<sup>1</sup>The authors wish to thank their students Sander van Duyn, Wanda Toxopeus, Stijn Gooskens, Thijs de Jongh & Edibe Tali for valuable help in setting up this experiment and collecting and analyzing the data.

## Sequence learning

One of the more recent paradigms to study implicit learning is so-called sequence learning. Subjects are typically offered sequences of stimuli that are formed according to some (formal) rule(s). The only thing subjects have to do is press some key that corresponds to the current stimulus. For example when the stimuli are just zeros and ones, the current stimulus could be formed by taking the xor of the preceding two stimuli. It is now interesting to see if subjects implicitly learn this rule. This is measured by comparing RTs on correct trials, that is trials on which the current stimulus is in fact the xor of the two preceding trials, with RTs on incorrect trials, where the current stimulus is *not* the xor of the two preceding trials.

Cleeremans and McClelland (1991), using this paradigm, had their subjects learn an endless sequence of stimuli generated by a finite state grammar. To determine the effects of implicit learning, they assessed reaction times, and found these to be decreasing as subjects got more training. Similar studies have been done where, instead of measuring reaction times, performance was assessed by asking subjects to predict the next stimulus after having seen an initial segment of a string. However, few studies have investigated the exact relation between RTs and prediction performance in implicit learning. The present study aims to gain insight into this relation by analyzing RTs and prediction performance simultaneously.

In this context the work of Cleeremans and McClelland (1991) on the SRN model for implicit learning, is of interest. They use the SRN model to predict RT performance of subjects by taking the reaction time of the network to be inversely proportional to the activity of the output unit corresponding with the correct response<sup>2</sup>. The activity of the 'correct' output unit can thus be interpreted as a measure of anticipation of the position of the next stimulus. This anticipation in turn can be used to make *predictions* of the next stimulus as well; in this case the position corresponding with the output unit with the highest activity has the highest probability of being predicted. This means that the

---

<sup>2</sup>Note that this doesn't leave the possibility for incorrect responses. This is not a big problem, however, since typically incorrect responses are very seldom because of the simplicity of the task.

SRN model predicts a negative relation between prediction performance and RTs, with the RTs decreasing as prediction performance gets better. The aim of the present study is to test this hypothesis empirically.

## Experiment

To assess the relation between RTs and prediction of stimuli directly we did a sequence learning experiment in which the standard series of RT trials was interspersed with prediction trials at which subjects had to guess where the next stimulus would come. A similar procedure is proposed by Jimenez, Mendez, and Cleeremans (1996) which they named the *continuous* generation task. The main difference between this procedure and other generation tasks is that no feedback is given on the correctness of the prediction; rather, after subjects have made their prediction the next stimulus of the sequence is presented with the same response-stimulus interval as between consecutive RT trials.

Subjects were given a four-choice RT task, consisting of a total of 4800 trials divided in twenty blocks of 240 trials each. The blocks were split into two sessions that were presented on two consecutive days. Unknown to subjects the sequence of stimuli followed a pattern that was generated using the finite state grammar which is described below. Because of the rather complex structure of the sequences generated with such a grammar subjects were presented with 4800 trials. There were two types of stimuli: RT trials and prediction trials. At the RT trials subjects were asked merely to reproduce the current stimulus by pressing the appropriate key. At the prediction trials subjects were asked to predict the next stimulus by pressing the appropriate key. Each block of 240 trials was divided into subblocks of four types: grammatical RT, random RT, grammatical prediction and random prediction. The switch from one subblock to the next was not marked so subjects were unaware of the existence of these subblocks. The sequence of stimuli in the random subblocks was unrestricted but for the fact that no two consecutive stimuli could be the same, which would lead to undesired speed-up of responses due to priming.

The random trials are used as a control condition, accommodating for possible effects of motor training, as well as for additional effects of subjects gaining implicit knowledge of the grammar. This design provides the possibility to assess the effects of implicit learning, by comparing RTs and prediction performance in the grammatical trials to those obtained in the random trials. Note that this is a within subjects design, so that each subject is his own control group (i.e., the performance of each subject on the grammatical trials is compared with that same subject's performance on random trials). The prediction of an inversely proportional relation between RTs and prediction performance, as derived from the SRN model (Cleeremans & McClelland, 1991), translates into three statistical hypotheses. The first is an interaction effect of condition and time on the RTs: If implicit learning occurs, RTs should decrease more for the grammatical trials than for the random trials. The

second is an interaction effect of condition and time on prediction performance: over time, prediction should improve for the grammatical trials, but not for the random trials. Finally, on trials leading to correct predictions, RTs

## Method

**Subjects** Twenty-four subjects, undergraduates at the Department of Psychology of the University of Amsterdam, participated in this experiment. They received both course credits and money for participation. On top of that they could earn bonuses for fast and accurate responses.

**Procedure** At the start of the experiment subjects were told that in this task both accuracy and speed were important. The experiment started with two small blocks of trials that were not recorded to familiarize the subjects with the task. Each block consisted of four subblocks: 20 random RT trials, 100 grammatical RT trials, 100 grammatical prediction trials and 20 random prediction trials. In the RT subblocks only reproduction of the stimuli was asked of the subjects; in the prediction subblocks RT trials were interspersed with prediction trials. At the end an extra block was added in which the order of the random and grammatical trials was reversed to test whether the order of the subblocks could influence the results.

To enable a more direct comparison between prediction and reaction times, an extra block was added in which the series of trials for both the RT subblock and the prediction block was identical. In this way it is possible to directly compare the RT on a given trial with the prediction made on the very same trial.

**Stimulus material** The sequence of stimuli in the grammatical subblocks was generated from the finite state grammar in figure 1. Sequences are produced by this grammar in the following manner:

1. Start in state #1 and randomly choose one of the arcs leaving that state while noting the letter corresponding to the followed arc.
2. In the next state repeat this process of choosing an arc and noting the corresponding letter
3. The process ends when state #7 is reached and the process starts over again to create strings of unbounded length.

**Display** As can be seen in figure 1 the alphabet of the grammar consists of four letters. The letters were translated into screen positions as shown in figure 2. In the grammatical RT subblock subjects were exposed to 100 trials; in each trial the  $\times$ -symbol appeared in one of the quadrants of the computer display and the subjects had to press the corresponding key on the numerical keypad on the keyboard. The keys 1,2,4 and 5 on the numerical keypad were used to ensure that the spatial configuration of the response keys matched the spatial configuration of the stimulus positions

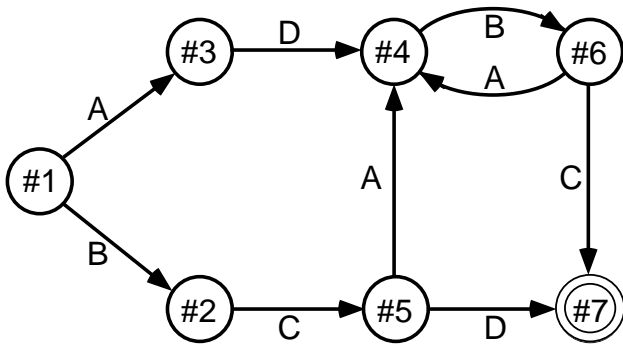


Figure 1: Finite state automaton used to generate strings for sequence learning experiments. A string is formed by starting in state #1 and then randomly choosing one of the arcs leaving that state meanwhile noting the letter corresponding to that arc. Continue stepping from state to state until the end state #7 is reached; from there the process starts over again from state # 1.

on the display. Subjects were instructed to hold their index finger over the middle of the four keys and press the appropriate key only with the index finger.

**Exit interviews** All subjects were asked a series of questions after the experiment was completed to assess whether subjects had acquired any explicit knowledge of the grammatical sequence.

### Results

The data of one of the subjects was not included in the analyses, because the subject had too many errors in three consecutive blocks due to misplacing the index finger over the numerical keypad. Comparison of the last two blocks revealed that the order of the subblocks, random before grammatical or vice versa, did not significantly influence reaction times.

**RT trials** Grammatical RTs decreased from 404.7 ms at the beginning of the experiment to 342.6 ms at the end; random RTs decreased from 414.2 ms to 370.3 ms. The mean RTs are displayed in Figure 3.

The first hypothesis predicts that RTs decrease more for the grammatical trials than for the random trials. In order to test this hypothesis, RTs were averaged over subjects and over two consecutive blocks. A repeated measures ANOVA with two within factors, block (10 levels)  $\times$  grammaticality (2 levels), indicates a significant interaction between grammaticality and blocks: as predicted, grammatical trial RTs decreased more over time than did random trial RTs,  $F(9, 198) = 3.87; p < 0.001$ . The analysis also yielded significant main effects for grammaticality and training: grammatical trial RTs were significantly smaller than the random trial RTs,  $F(1, 22) = 59.49; p < 0.001$ , and RTs became faster over blocks for both grammatical

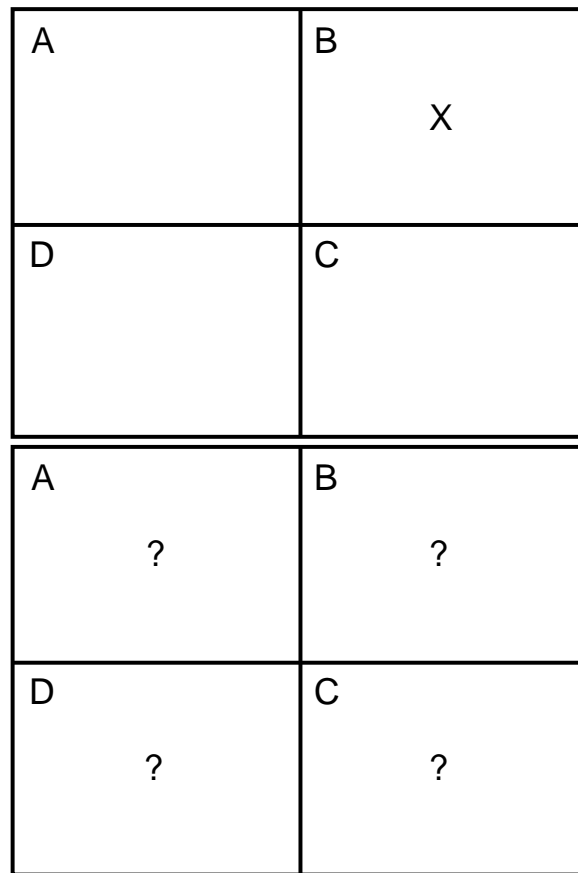


Figure 2: The top panel shows the computer display for the RT trials. Subjects have to press the key corresponding to the quadrant of the screen where the  $\times$  is shown. In the bottom panel the screen lay-out for a prediction trial: all quadrants have a question mark and subjects have to choose whatever letter they think will occur next. The letters in the top-left corner of the quadrants were not part of the actual display.

and random trials,  $F(3.78, 198) = 25.75; p < 0.001$  with Greenhouse-Geisser correction for non-homogeneous variances.

**Prediction trials** The percentage of correct predictions in grammatical subblocks increased from 33.6 % at the beginning to 52.2 % at the end of the experiment. The corresponding percentages for the random predictions are 30 and 34 % respectively. The proportions of correct responses on predictions for both random and grammatical subblocks are displayed in Figure 4.

The second hypothesis states that prediction performance should improve over time for the grammatical trials, but not for the random trials. In line with this prediction, a significant interaction between blocks (time) and

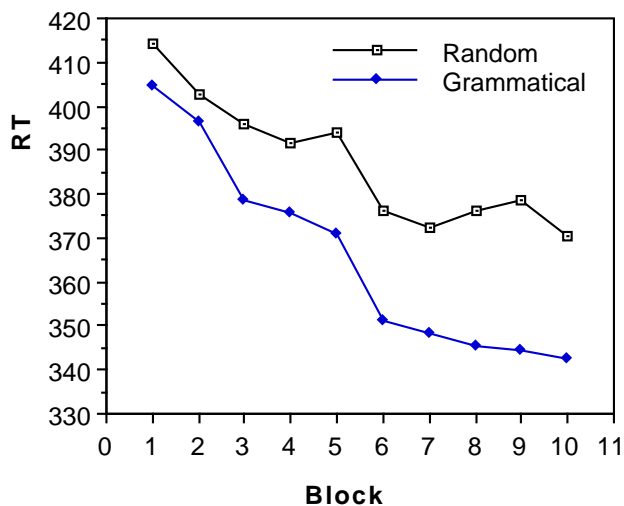


Figure 3: Mean reaction times for grammatical and random trials. Means are averaged over two consecutive blocks,  $N = 23$ .

grammaticality was found,  $F(7.9, 198) = 2.25$ ;  $p = 0.027$ , showing that the grammatical predictions did show more improvement over time than did the random predictions. More specifically, there was no improvement over time for the random trial predictions when analyzed separately,  $F(1, 228) = 0.845$ ;  $p = 0.359$ , as was to be expected.

**Prediction and RT trials: comparison** To compare performance on prediction and RT trials directly we added a block of trials in which the strings used for the RT trials and for the prediction trials were identical. Table 1 shows the mean RTs for correctly and incorrectly predicted items in this added block of trials. An anova with one within factor (correct vs. incorrect) confirms that correct predictions correspond to fast RTs,  $F(1, 22) = 6.44$ ;  $p = 0.019$ .

Table 1: Mean reaction times for correctly and incorrectly predicted trials.

Prediction	mean	sd
correct	360.96	49.64
incorrect	389.97	30.30

**Exit interviews** Subjects were asked whether they noticed anything particular in the sequence of stimuli. Although some subjects felt there was some ‘regularity’ in the sequence, none of the subjects could specify this, except for three subjects that said that the subsequence  $AB$  occurred rather frequently. This is the subsequence in the grammar which corresponds with the loop between the two top right nodes in Figure 1.

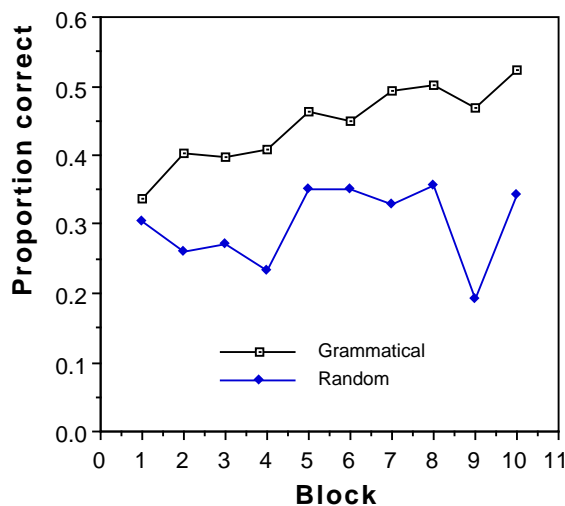


Figure 4: Proportion correct predictions of grammatical and random prediction trials,  $N = 23$ .

## Discussion

The results show that implicit learning occurs: subjects give faster responses on grammatical trials than on random trials and this effect becomes larger towards the end of the experiment. Secondly, subjects gradually get better at predicting following stimuli due to training as well. Thirdly, as expected, smaller RTs correspond with a better ability to predict the following stimulus.

## Models of sequence learning

Cleeremans and McClelland (1991) applied the SRN to implicit sequence learning. The SRN successfully describes subjects’ growing sensitivity to dependencies between successive stimuli. The success of the SRN model is due to its ability to capture the ‘statistical constraints’ inherent in the sequence of stimuli. The SRN model also correctly predicts, at least in a qualitative manner, the inverse relation between RTs and the proportion of correct predictions as we have shown above. A drawback of the SRN model is that it is not very well suited for describing individual differences. The SRN model construes implicit sequence learning in subjects as statistical learning. Subjects first grow sensitive to first order frequencies of symbols, then to second order frequencies, that is bigram frequencies, then third order frequencies et cetera. Individual differences in both the learning process and the resulting implicit knowledge base, that is knowledge of frequency constraints, are not brought out by the model. Below we will describe how hidden Markov models can be used to model individual behavior of subjects.

## The hidden Markov model

Hidden Markov models, henceforth HMMs, are also called stochastic finite automata since they are equivalent to finite

automata where the arcs between states have probabilities corresponding to them. The only restriction is that the probabilities on the arcs leaving a particular state should sum to one. This resemblance to finite automata is the reason for exploring the possibility of applying HMMs to implicit learning. Before presenting results of fitting HMMs to subjects' data we give a short introduction to HMMs.

Hidden Markov models have mainly been used in speech recognition applications such as Schmidbauer, Casacuberta, Castro, and Hegerl (1993), Chien and Wang (1997) although recently more psychologically oriented applications have come up as well such as in action learning (Yang, Xu, & Chen, 1997). The main reason that HMMs are used in speech recognition is that they are especially well suited for capturing temporal dependencies in a series of utterances which then helps in identifying phonemes. This feature can be used to model the temporal dependencies that are inherent in the series of stimuli that are typically used in implicit learning.

More formally a HMM consists of the following elements (notations adapted from Rabiner (1989)), also see figure 5 for clarification:

1. a set of states  $S_i, i = 1, \dots, N$
2. a set  $V$  of observation symbols  $V_k, k = 1, \dots, M$
3. a matrix  $A$  of transition probabilities  $a_{ij}$  for moving from state  $S_i$  to state  $S_j$
4. a matrix  $B$  of observation probabilities  $b_j(k)$  of observing symbol  $V_k$  while being in state  $S_j$
5. a vector  $\pi$  of initial state probabilities  $\pi_i$  corresponding to the probability of starting in state  $S_i$  at  $t = 1$

The equations describing the dynamics of the model are as follows:

$$S_{t+1} = A S_t + \zeta_{t+1}$$

$$O_{t+1} = B S_t + \xi_{t+1},$$

where  $S_t$  is the hidden process and  $O_t$  is the observed process;  $\zeta_{t+1}$  and  $\xi_{t+1}$  are zero mean martingale increment processes, cf. Elliott, Aggoun, and Moore (1995, p. 20) for further details. A hidden Markov process then is a Markov process with multiple indicators for each (hidden) state. By substituting  $S_t$  by its definition in terms of  $S_{t-1}$  in the defining equation for  $O_{t+1}$  it is easily seen that in fact  $O_{t+1}$  is dependent on all foregoing observations back to  $O_1$ . Hence, at any given point observations can depend on all foregoing observations. This is in contrast with a normal Markov model where the next observation only depends on the current observation.

### Characterizing sequence learning behavior

Fitting a hidden Markov model is in fact the inverse of producing a sequence of stimuli from a finite state automaton:

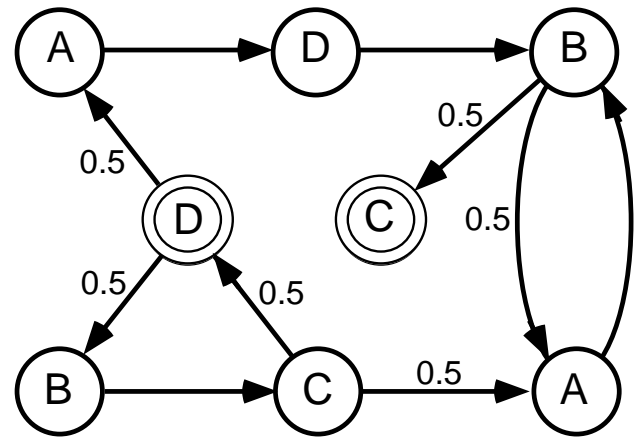


Figure 5: Representation of a hidden Markov model. This model produces exactly the same sequences as the grammar we used in the experiment with equal probabilities. Sequences are generated in the same manner as in FSAs: start in one of the states on the left with letter A or B, then follow the arcs leading from those states. A sequence ends when one of the accepting states is reached, that is the two states with the double circle around them. From there the process continues by going to one of the starting states again. For the accepting state with the letter D the arcs are drawn to the start states. For reasons of clarity the arcs from the accepting state with the C are left out. The arcs leading from one state to the next have probabilities corresponding to them which are given in the figure for some of the arcs.

finding the best automaton to describe a given sequence of observations. This procedure can be applied to any kind of sequence of categorical observations and hence also to a sequence of responses in a sequence learning experiment. In simulation studies we have shown that in fitting a HMM the right automaton can be induced from the data (Visser, Raijmakers, & Molenaar, accepted for publication). That is, having generated a sequence from the grammar used in the experiment we found the HMM in Figure 5 exploratively.

**Sequence learning data** In the prediction subblocks of the experiment subjects were presented with question marks on the screen at random points in the sequence of stimuli. In between the prediction trials normal RT trials were presented. For each subject this resulted in a sequence of responses consisting of the trials that were presented on the screen interspersed with their own predictions about the position of the next stimulus.

In order to characterize sequence learning we fitted HMMs on these sequences of responses. To bring out the learning we fitted separate HMMs on the initial and final segments of the sequence of responses. Both segments consisted of 500 trials. We expected to see a rise in number of hidden states of the model from beginning to end; that is,

we expected subjects to gradually build a more complex model of the grammar underlying the sequence of stimuli. A rise in number of states would reflect subjects' growing sensitivity to the structure of the sequence. For two subjects we indeed found such a rise in the number of states from two states at the start of learning to four states at the end of learning. Overall however, results were inconclusive. This is, we think, mainly due to the fact that only a small proportion of the series of responses that were analyzed were actually produced by the subject. Of the series of 500 trials that the HMMs were fitted on, only 125 were produced by the subjects, the others were generated by the finite state automaton and only *reproduced* by the subjects. As a consequence, of all the responses only a quarter could be useful in discriminating between beginning and end of the learning phase. Hence the low power of the test. In future research it would be useful to have longer sequences of freely generated responses to which HMMs can be fitted more reliably.

### Conclusion

In sequence learning both RTs and prediction have been used as a measure of performance. The results of this experiment show that when measured simultaneously it is possible to relate directly improvement in prediction performance and improvement in RT performance on grammatical trials. The direct comparison shows what is to be expected: fast RTs are indicative of the subjects' level of anticipation of the next trial and on the same count result in correct predictions. With this study it is also shown that prediction is possible even in a fairly complex rule system, that can not be verbalized by subjects.

The SRN model has proved to be a valuable model for describing the learning processes inherent in implicit sequence learning. However the model does not seem especially suitable to describe individual subjects' behavior. Therefore we introduced the hidden Markov model as a stochastic counterpart of the FSA to characterize individual learning behavior. Since hidden Markov models are an excellent means of describing temporal dependencies between responses they are in principle well suited for describing implicit learning behavior. Our results with fitting HMMs are promising in that we can reliably estimate them on the kind of sequences that are generally used in implicit sequence learning. It would be interesting to do experiments where subjects generate longer sequences of responses instead of the single predictions they made in the experiment described in this paper.

### References

- Chien, J. T., & Wang, H. C. (1997). Telephone speech recognition based on bayesian adaptation of hidden Markov models. *Speech Communication*, 22(4), 369–384.
- Cleeremans, A., & Jimenez, L. (1998). Implicit sequence learning: The truth is in the details. In M. Stadler & P. Freuch (Eds.), *Handbook of Implicit Learning* (pp. 323–364). Thousand Oaks (Ca): Sage Publications.
- Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *JEP: General*, 120, 235–253.
- Elliott, R. J., Aggoun, L., & Moore, J. B. (1995). *Hidden Markov models: Estimation and control*. New York: Springer Verlag.
- Jimenez, L., Mendez, C., & Cleeremans, A. (1996). Comparing direct and indirect measures of sequence learning. *JEP: Learning, Memory and Cognition*, 22–4, 948–969.
- Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, 19, 1–32.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE*, 77(2), 267–295.
- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, 6, 317–327.
- Schmidbauer, O., Casacuberta, F., Castro, M. J., & Hegerl, G. (1993). Articulatory representation and speech technology. *Language and Speech*, 36(2), 331–351.
- Seger, C. A. (1997). Two forms of sequential implicit learning. *Consciousness and Cognition: An International Journal*, 6(1), 108–131.
- Visser, I., Raijmakers, M. E., & Molenaar, P. C. (accepted for publication). Confidence intervals for hidden Markov model parameters. *British journal of mathematical and statistical psychology*.
- Yang, J., Xu, Y., & Chen, C. S. (1997). Human action learning via hidden Markov model. *IEEE Transactions on Systems, Man and Cybernetics*, 27(1), 34–44.

# A Constructivist Dual-Representation Model of Verb Inflection

Gert Westermann (gert@cogsci.ed.ac.uk)

Institute for Adaptive and Neural Computation, Division of Informatics  
University of Edinburgh, 2 Buccleuch Place  
Edinburgh EH8 9LW, Scotland UK

## Abstract

A constructivist neural network is presented that models impaired inflectional processing in German agrammatic aphasia. The model is based on a single mechanism and develops two types of representation through a constructivist learning process. The model accounts for data that has been taken as evidence for a dual mechanism theory of inflection, and it suggests an inflectional processing system that is based not on a distinction between regular and irregular cases, but between inflections that are easy and hard to learn. The model represents a successful single-mechanism neural network account of verb inflections.

## Introduction

The debate between rule-based and association-based theories of inflection has been continuing for many years and has moved from the initial focus on the English past tense to other languages such as the German participle (e.g. Clahsen, 1999; Marcus *et al.*, 1995). The reason for this shift is that in English, the issues of “regularity” and “high frequency” are confounded which makes it difficult to distinguish between the different theories. By contrast, in the German participle the regular case does not apply to the majority of all verbs, making it a so-called “minority default” (Marcus *et al.*, 1995).

A popular recent theory of how inflections are formed is the *Dual Mechanism Theory* (DMT) that postulates two qualitatively distinct mechanisms for the production of regular and irregular cases (e.g. Clahsen, 1999; Pinker, 1991, 1997; Marcus *et al.*, 1995). According to the DMT, regular inflections are produced by a mental symbolic rule, whereas irregulars are stored in an associative lexicon. Based on these mechanisms, the DMT claims to account for differences in the processing of regular and irregular inflections: whereas regular forms are applied productively to novel forms independently of their similarity to existing forms (e.g., *faxed*), irregular inflections show similarity effects both in existing “families” (*read* → *read*, *lead* → *led*, *breed* → *bred*) and in the extension to novel forms (*cleed* → *clad*).

However, while considerable empirical research has established processing differences between regular and irregular forms on many different levels from acquisition over psycholinguistic and ERP studies to impaired adult processing (see Clahsen, 1999, for an overview), little progress has been made in the specification of the DMT. Particularly problematic is the question in which way the

two mechanisms interact to produce the inflected form. Marcus *et al.* (1995) proposed the *Blocking Principle* which states that a lexical entry (indicating an irregular verb) blocks the application of the rule, but an implementation of this principle (Nakisa *et al.*, 1997) showed that in practice it involves parameters for which a useful setting cannot be found. Therefore, the DMT remains highly underspecified and thus hard to falsify. However, even in its underspecified form, the DMT is contradicted by some empirical data, e.g., frequency effects for regular English past tense (Stemberger and MacWhinney, 1986) and regular Dutch plural (Baayen *et al.*, 1997) forms, and similarity effects for regular German participles in agrammatic aphasia (Penke *et al.*, 1999).

In this paper I present a neural network model of inflectional processing in German agrammatic aphasia that accounts for dissociations between regular and irregular forms without postulating two qualitatively distinct mechanisms. Instead, the model develops two types of representations in a constructivist process, driven by the structure of the training data, and it displays emerging areas of functional specialization that correspond largely, but not completely, to the distinction between regular and irregular forms. The trained model is lesioned in different ways and it accounts for empirical data better than the DMT. Based on these results I propose a new theory of inflectional processing that is based on a distinction not between regular and irregular, but between “easy to learn” and “hard to learn” forms.

The rest of this paper is organized as follows: first, the structure of the German participle and the impairment profiles observed in agrammatic aphasia are reviewed. Then, the network model, the data, and the training regime are described, followed by a detailed analysis of the performance of the model in comparison with agrammatic aphasics. Finally, the resulting new theory of inflectional processing is presented and related to the DMT.

## The German Participle

German participles are comparable in usage to the English past tense in describing an event in the past. There are three groups of participles: *Weak* participles are formed by a (prosodically determined) prefix *ge-*, the verb stem, and the ending *-t*, e.g., *sagen* (say) → *gesagt* (said). *Strong* participles take the ending *-en*, e.g., *geben* (give) → *gegeben* (given) and they may also change the



verb stem, e.g., *gehen* (go) → *gegangen* (gone). A few strong verbs have idiosyncratic participle forms, e.g., *sein* (be) → *gewesen* (been). The third group are *mixed* verbs that take the weak ending *-t* but change their stems like strong verbs, e.g., *wissen* (know) → *gewusst* (known). It is generally claimed that the weak verbs form the regular class, while strong verbs are irregular, and the terms regular and irregular will here be used in this sense.

In contrast to English, German does not have a majority of regular tokens (each verb counted according to how often it occurs in a corpus), and the majority of types (each verb counted just once) is less pronounced than in English.

The CELEX database (Baayen *et al.*, 1993) lists 3015 German participles. After cleaning out some obvious errors and homophones and choosing the more frequent of different participle forms of one stem, 2992 participles remain. However, German verbs are often formed by modifying other existing verbs with a prefix or separable particle, e.g., the simplex verb *fahren* (drive) occurs in CELEX in 28 composite forms such as *hinausfahren*, *losfahren*, *fortfahren* etc. (drive out, drive off, continue). Since a prefix or particle do not alter the way in which the participle of a simplex verb is formed, all composite forms were combined into one simplex form.

For the simulation experiments described below, 20,000 verb tokens were randomly extracted from this corpus according to their frequency. To ensure that each verb occurred at least once, all verb types which had not been randomly selected were added onto the resulting corpus with a token frequency of one (this applied to 18 verbs).

The structure of the resulting training corpus is shown in table 1.

	type	token
Regular	518 (78.01%)	9306 (46.49%)
Irregular	134 (20.18%)	9717 (48.54%)
Mixed	12 (1.81%)	995 (4.97%)
Sum	664 (100.00%)	20018 (100.00%)

Table 1: The structure of the training corpus.

## Agrammatic Aphasia

Agrammatic (Broca’s) aphasia is a language disorder that is generally caused by a stroke predominantly affecting anterior parts of the left hemisphere. One of the characteristic symptoms of Broca’s aphasia is the tendency to omit or confuse inflections. Investigating the precise nature of these deficits can therefore lead to insights into the internal representation of inflectional morphology. Penke *et al.* (1999) analyzed data from eleven aphasic subjects who each produced 39 regular and 39 irregular participles in a sentence completion task with respect to regular and irregular errors, overregularizations and irregularizations, frequency effects, and effects of ablaut-patterns on error rates. They found irregular inflections to be selectively impaired in six of the subjects, and three showed no significant difference between regular

and irregular participles (the remaining two made more irregular errors but their total number of errors was too small to establish a significant difference between regulars and irregulars). Penke *et al.* (1999) concluded that irregular inflection can be selectively impaired in agrammatic aphasia.

## The Network Model

For the simulations described in this paper, a constructivist neural network (CNN) model was developed that builds the hidden layer of a radial basis function (RBF) network. Each hidden unit has a Gaussian activation function and thus acts as a *receptive field* for an area of the input space. The problem in building RBF networks is to decide on the number and positions of these receptive fields. The CNN algorithm solves this problem by constructing the hidden layer during learning, adding units when and where they are needed. The network starts with just two units in the hidden layer, each covering roughly half of the input space (see figure 1). The network tries to learn the task with this architecture (by adjusting the weights with quickprop), and when learning no longer improves the performance, a new unit is inserted. The place where the new unit is inserted is determined by the classification error resulting from treating inputs within one receptive field as similar: the receptive field that previously caused the highest error is shrunk and the new unit is inserted next to it. The idea here is that a unit which produces a high output error is inadequate, and therefore more structural resources are needed in that area. A similar network has already been successfully used to model the acquisition of the English past tense (Westermann, 1998).

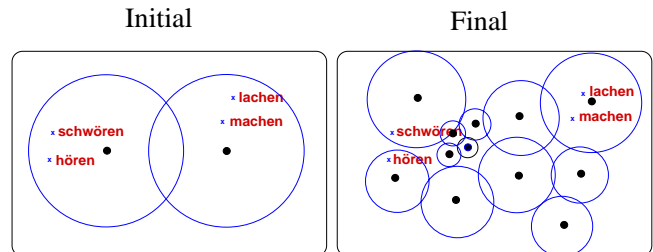


Figure 1: Receptive fields covering the input space at the beginning (left) and the end (right) of learning.

Figure 1 shows a hypothetical start and end state in a two-dimensional input space. While initially only two units cover the whole of the space, later hidden units have been inserted with different densities across the space to account for the specific learning task.

Figure 2 shows the network architecture. The input layer takes a phonological representation of the verb infinitive, and the output layer has one unit for each possible output class (see below). The hidden layer initially consists of only two units but is grown during learning. There are direct connections from the input to the output layer, and each hidden unit is fully connected to the output layer.

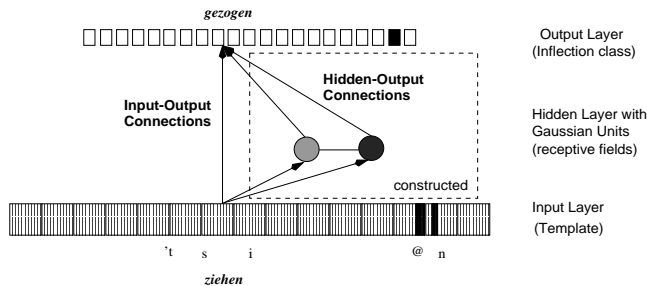


Figure 2: The initial architecture of the network.

## Data

The 664 German verbs were classified according to the way in which their participles are formed, resulting in a total of 22 classes, one of which was the “stem+*-t*” (regular) class, 6 were for mixed verbs, and 15 for irregular verbs.

The verbs were represented phonologically, and each phoneme was encoded by a 7-bit feature vector with features such as *fricative*, *plosive*, *voiced* etc. for consonants, and *front*, *high*, *open* etc. for vowels. Presence of a feature was encoded with 1 and absence with -1.

For the training of the network, the phonological representation of the infinitive of each verb was then inserted into a template consisting of three syllables: XCCCVCVCC-XCCCVCVCC-XCCCVCVCC; C stands for consonant, V for vowel, and X for whether the syllable is stressed or not. Since the endings of verbs are significant for the determination of the participle class, the verbs were right-aligned in this template so that the endings occurred in the same slots.

The resulting network had 150 input units (three syllables with seven phonemes each represented by seven features, plus one stress-bit per syllable), and 22 output units for the 22 inflection classes.

## Training

The task to be learned by the network was the mapping from the phonological representation of the verb infinitive to the class of its participle. Viewing the learning of the participle as a classification task avoids confounding it with phonological details such as different pronunciation of regular forms depending on the last stem phoneme (e.g., *holen* → *geholt* vs. *landen* → *gelandet*).

Five CNN models were trained on this corpus with different random initial weight settings. The networks were tested before the insertion of a new hidden unit. An output class was counted as correct when the corresponding unit, but no other unit, had an activation value over 0.7.

## Results

In order to model agrammatic aphasia, the CNN was lesioned in different ways. It was assumed that the removal of weights in the model corresponds to the destruction of neural tissue in the brain by a stroke.

## Localized Lesioning

The output in the CNN model is produced through two sets of connections: the direct connections between the input and the output layer that the network started out with, and the connections from the growing hidden to the output layer. A localized lesioning of these pathways in the CNN resulted in a double dissociation between regular and irregular verbs for four out of the five runs. The further analyses were conducted with these four networks.

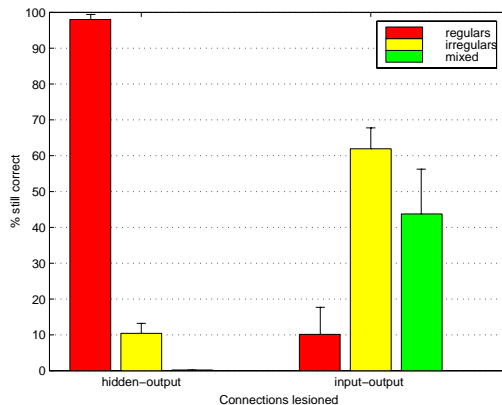


Figure 3: Double dissociation between regular and irregular (and mixed) verbs after lesioning the two pathways in the networks.

Figure 3 shows the results of lesioning the hidden-output (HO) and the direct input-output (IO) connections. Lesioning the HO connections resulted in a marked decrease of the performance of irregular and mixed verbs, with regular inflections remaining nearly fully intact. By contrast, lesioning the IO connections resulted in the opposite profile: performance of regulars was significantly more impaired than that of irregular and mixed verbs. It is important to note that this double dissociation emerged as a result of the structure of the training data together with the constructivist development of the model and was in no way prespecified.

Removing the HO connections in the network thus modeled the basic deficit in the inflection of agrammatic aphasics, namely, the breakdown of irregular and selective sparing of regular participles. Based on this result, the performance of the HO-lesioned CNN models was investigated with respect to the more detailed results reported by Penke *et al.* (1999).

Penke *et al.* (1999) found that all subjects who made more errors on irregulars than on regulars overgeneralized the regular ending *-t* to irregular verbs, but they only rarely irregularized regular verbs (i.e., their regular errors consisted mainly in using a wrong suffix or none at all). Testing the four corresponding CNN models for this behavior showed a good match of the aphasic profiles: the networks over-applied the regular class to 73.7% of all wrong irregulars (aphasics: 63.3%), but only 6.5% of all regular errors were irregularizations (aphasics: 14.3%). The other errors that can be made by the CNN models are no output, or ambiguous output when two (or more) output units are simultaneously activated.

Based on the assumption of two qualitatively distinct processing mechanisms for regular and irregular inflections, Penke *et al.* (1999) predicted and found a frequency effect in the aphasic production of irregulars, but not of regulars: there were significantly more errors for infrequent irregulars than for frequent ones, but no such effect occurred for regulars. When tested on the same verbs as the aphasic subjects, the CNN models equally showed a small frequency effect for irregulars but not for regulars: the error rate for low frequency irregulars (93.3%) was significantly higher than for high frequency irregulars (89.0%) (Wilcoxon,  $p = 0.068$ ), but error rates for regulars did not differ statistically (1.7% for low frequency and 2.4% for high frequency regulars,  $p = 0.273$ ).

Alternatively to a qualitative distinction, regulars and irregulars might represent two ends of a continuum: a regular verb can be said to be “very regular” if it is similar to other regulars and dissimilar to irregulars. It is “less regular” if it is dissimilar to other regulars but similar to irregulars. The reverse is true for irregulars (see also Daugherty and Seidenberg, 1992).

This assumption is attractive because it integrates mixed verbs which fall between regulars and irregulars in that they combine an irregular stem with the regular ending. Mixed verbs are generally ignored in the DMT because they are hard to consolidate with the proposed qualitative distinction between regulars and irregulars.

A regularity continuum would predict that “less regular” regulars, being more similar to irregulars, should be more error prone than “very regular” regulars in agrammatic aphasics. Penke *et al.* (1999) analyzed the distribution of verbs with respect to stem vowels and found that for the stem vowel <e>, irregulars outnumber regulars, making regulars with this stem vowel less regular. Therefore, regular verbs with <e> should have a higher error rate because they are similar to irregulars.

This prediction was confirmed in the analysis of the aphasic data: all regular suffixation errors occurred with <e>-stems. While Penke *et al.* (1999) interpreted their results within the framework of a qualitative distinction between regulars and irregulars (allowing grading effects for both mechanisms with the qualitatively distinct verb groups influencing each other), a more plausible interpretation is that of a regularity continuum where a single mechanism underlies the production of both forms.

Testing the CNN model, which is based on such a single mechanism, for this effect yielded the same pattern of results as in the aphasic subjects: when tested on the same verbs, 4 out of 5 of the regular errors were for the stem vowel <e>, indicating that these verbs are treated more like irregulars.

In summary, by lesioning the HO connections in the CNN model, detailed aspects of the performance of agrammatic aphasics on German participle inflections could be modeled. These results comprise both those that have been claimed to be evidence for the dual mechanism theory (double dissociations; frequency effects only for irregulars) and those that contradict the predictions of the dual mechanism theory (regularity continuum effect).

## Global Lesioning

As shown in the previous section, the lesioning of the HO pathway in the CNN model can account for a selective impairment in the inflection of irregular verbs and thus model the performance of agrammatic aphasic subjects. This selective and total lesioning of one pathway might suggest that the processing of regular and irregular verbs is subserved by locally different brain structures that can be selectively affected by a stroke. To establish whether the observed profile could be modeled without this assumption, the effects of globally lesioning the network to different degrees was investigated, without making a distinction between the IO and the HO connections. Over 200 trials, the network was lesioned in 5%-steps by randomly removing weights from both sets of connections.

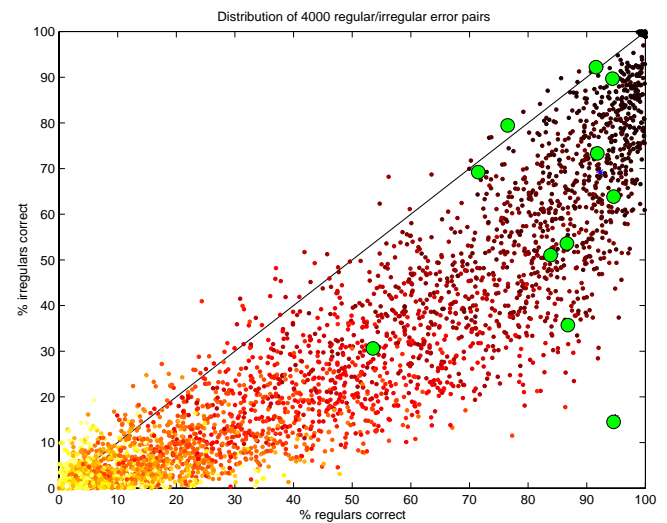


Figure 4: Performance on regulars vs. irregulars for 200 lesioning trials at 20 lesioning steps each (in 5%-steps). Greyscale indicates degree of lesioning (from dark to light). Data for the aphasic subjects are marked by circles.

The result of this global lesioning is shown in figure 4. The 4,000 lesioned networks showed some variety of regular vs. irregular errors, but, like with the aphasic subjects, there was never a selective sparing of irregulars with a breakdown on regular participles (top left of the plot). Instead, in most cases impairment of irregulars was stronger than of regulars (below the diagonal).

The data for the eleven aphasic subjects from (Penke *et al.*, 1999) are also displayed in figure 4. All aphasic data are within the range of performance predicted by the simulations, showing that although there is variability in the performance of agrammatic aphasics, differently lesioned CNNs can model the performance of each of them. The model is not over-general, however: like in aphasic subjects, a selective sparing of irregulars with a breakdown of regular inflections did not occur in any of the lesioning trials.

Why does global lesioning in the CNN lead to a profile in which irregular participles are more impaired than

regulars? An answer to this question can be found by analyzing the connections in the model. Many of the IO connections are inhibitory, suppressing the activation of the wrong inflection class by other IO connections. This profile is due to the distributed representation of the input: overlapping representations between classes make the inhibition of wrongly activated classes necessary, and with increased lesioning this inhibition is lost, resulting in the activation of wrong output classes for regular and irregular verbs equally. By contrast, the HO connections from one receptive field usually contain only one strongly excitatory weight to the correct output class. Therefore, the HO weights do not tend to activate a wrong output class. This different weight structure can be explained by the localist nature of the receptive fields: due to the constructivist growth process, receptive fields tend to cover only verbs from one class. Therefore, representations for different classes do not overlap and inhibition is not required. An analysis of the distribution of the receptive fields over the verbs showed that they had been preferentially allocated for the difficult-to-learn irregular verbs. Therefore, a partial lesioning of the HO connections affected predominantly irregulars. Taken together, irregulars were impaired by the removal of weights in both the IO and the HO connections, while regulars were affected only by lesioned IO connections. Together, a global lesioning therefore led to a more pronounced breakdown for irregulars than for regulars.

A global lesioning profile in which regular inflections are selectively impaired could only arise from a total lesioning of the IO connections together with no or weak lesioning of the HO connections. Based on the CNN model therefore the prediction is made that a selective impairment of regular inflections in aphasics would be evidence for a locally separate processing of regular and irregular inflections in the brain, whereas the selective impairment of irregulars cannot be taken as evidence for such a separation.

## A Dual-Representation Theory of Verb Inflection

The results described in this paper show that the CNN can account for detailed empirical results from agrammatic aphasic inflectional processing. At the same time, the CNN avoids the problems of the DMT, namely, underspecification and contradiction to some empirical data.

Whereas the DMT proposes two mechanisms operating on a single representation of a verb stem, the CNN develops so that a single mechanism operates on two representations of the verb. Initially, the direct phonological input is used in the IO pathway to produce the output class. For verbs for which the output cannot be learned based on this structural representation alone, the CNN develops through a constructivist process additional representations in the hidden layer. In contrast to the structure-based input representations, these new representations are identity-based and localist: the activation of a hidden unit receptive field only indicates the presence of a certain input, without information about

its structure. The CNN is therefore a single mechanism, but dual representation model. This dual representation view sheds a different light on the dissociations between regular and irregular forms. The DMT does not assume that any regular verbs are produced by the irregular mechanism, or vice versa. The common aphasic profile where both regular and irregular cases are partially impaired (albeit to different degrees) is therefore often attributed to performance errors or the unpredictability of aphasic impairment.

A more compelling explanation is offered by the CNN: here, the dissociations that become visible in the lesioning trials do not run clearly along the lines of regulars vs. irregulars. Instead, all verbs for which the inflection class cannot be learned in the direct IO pathway are shifted to the developing hidden layer and the HO pathway. This shift concerns regular, irregular, and mixed verbs, to different degrees. The dissociation between verbs is thus better described as *easy to learn* vs. *difficult to learn*, with the difficult forms relying on the hidden layer, whereas easy forms are produced in the IO pathway alone. This distinction can account better for the data such as mixed verbs, a regularity continuum, or the different aphasic profiles.

But what factors determine whether a form is easy or difficult to learn? The degree of difficulty is determined by several interacting distributional factors that can be derived from the principles of associative learning:

1. Frequency: a frequent transformation is easier to learn than an infrequent one. Therefore, inflection classes with a high summed token frequency will be easier to learn than those that only apply to rare verbs.
2. Class size: a transformation that applies to many different verbs is easier to learn than one that just applies to one verb. Therefore, inflection classes with many members (counted in types) are easier to learn than those confined to only a small group of verbs.
3. Similarity of class members to members of other classes: the inflection class of a verb is easier to learn if other similar verbs share the same class.
4. Ambiguity of inflectional morpheme: an inflection is easier to learn if it applies uniquely to members of its class, i.e., if it does not exist in other context as well. For example, the *-ed* suffix in English is highly indicative of the past tense: an analysis of the CELEX corpus showed that 99.6% of all word types in English that end in *-ed* are past tense forms. By contrast, the German irregular participle ending *-en* is much more ambiguous: it also occurs in verb infinitives (*gehen*, to go), noun plurals (*Wiesen*, meadows), and as part of noun singulars (*Drachen*, kite).

These factors influence each other, and further research will be needed to establish in detail how they interact. Nevertheless they show that the *regular—irregular* distinction is a good first approximation of the *easy—difficult* distinction: the regular inflection, although it does not apply to the most frequent individual

verbs, is the single most frequent inflection in both English and German: 57.2% of English past tense tokens and 46.89% of German participle tokens are regular. At the same time, these classes are also the biggest in type size (88.4% and 64.7%, respectively). However, the third point, similarity of class members to members of other classes, does not separate along the lines of regular and irregular verbs: many regular verbs are similar to irregulars which should make them harder to learn in this view. And in fact the regularity continuum that has been shown for aphasics indicates that regulars that are similar to irregulars are more prone to impairment than others, that is, they rely more on storage in the lexicon.

A similar analysis of factors influencing errors in past tense formation has been conducted with school children (Marchman, 1997), where their errors on an elicited past tense production task were determined by frequency, the number of similar sounding stems in the same and in different inflection classes, and the phonological characteristics of the stem and past tense forms.

Taken together, although the dissociations of verbs into easy and difficult corresponds largely to the regular-irregular dissociation, it nevertheless suggests that the regular case is a post-hoc extraction and idealization of the developed structure of the inflectional processing system.

## Discussion

The results presented in this paper suggest a novel account of inflection learning and processing: it is a single mechanism system in which dual representations emerge from a constructivist learning process together with the structure of the environment. The system separates verbs along the lines of easy vs. hard to learn and can thus better explain empirical results that have so far been taken as evidence for the Dual Mechanism Theory. The qualitative distinction between regular and irregular inflections that lies at the core of the DMT, is a projection of formal linguistic analysis onto the human data. Because according to formal linguistics, human language data does not correspond to the abstract “competence” but is instead corrupted as “performance”, any data that does not correspond to the predictions of the formal theory (i.e., regulars that behave like irregulars and vice versa) can therefore be attributed to performance. This method makes the DMT hard to falsify based on such data. By contrast, the CNN model is fully specified, and it shows how the actual human data can be modeled without recourse to a competence-performance distinction. Whereas the abstract category of “regularity” remains a good formal description of language structure, the fallacy is in drafting it into service as a *processing* category, as done in the DMT.

A way to test the validity of the CNN model empirically is to abandon the regular/irregular distinction in favour of an easy/hard distinction, by identifying “hard” regulars and “easy” irregulars. Such a distinction should then better predict impairment profiles in agrammatic aphasics and other aspects of dissociations in inflectional systems. More research along these lines will be needed

to empirically verify the dual-representation model of verb inflection.

While connectionist, single-mechanism models of inflections have been rejected by proponents of the DMT (e.g. Clahsen, 1999; Pinker, 1997; Marcus *et al.*, 1995), the CNN model presents evidence that such models can account for inflectional processing more successfully than theories that rely on qualitatively distinct processing mechanisms.

**Acknowledgements** The author is now at Sony Computer Science Lab, 6 rue Amyot, 75005 Paris, France. (gert@csl.sony.fr).

## References

- Baayen, H., Piepenbrock, R., and van Rijn, H. (1993). *The CELEX Lexical Database*. CD-ROM. Linguistic Data Consortium. University of Pennsylvania, PA.
- Baayen, R. H., Dijkstra, T., and Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, **37**(1), 94–117.
- Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences*, **22**(6), 991–1013.
- Daugherty, K. and Seidenberg, M. S. (1992). Rules or connections? The past tense revisited. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, pages 259–264, Hillsdale, NJ. Erlbaum.
- Marchman, V. A. (1997). Children’s productivity in the English past tense: The role of frequency, phonology, and neighborhood structure. *Cognitive Science*, **21**(3), 283–304.
- Marcus, G., Brinkmann, U., Clahsen, H., Wiese, R., and Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology*, **29**, 189–256.
- Nakisa, R. C., Plunkett, K., and Hahn, U. (1997). A cross-linguistic comparison of single and dual-route models of inflectional morphology. In P. Broeder and J. Murre, editors, *Cognitive Models of Language Acquisition*. MIT Press, Cambridge, MA.
- Penke, M., Janssen, U., and Krause, M. (1999). The representation of inflectional morphology: Evidence from Broca’s aphasia. *Brain and Language*, **68**, 225–232.
- Pinker, S. (1991). Rules of language. *Science*, **253**, 530–535.
- Pinker, S. (1997). Words and rules in the human brain. *Nature*, **387**, 547–548.
- Stemberger, J. P. and MacWhinney, B. (1986). Frequency and the lexical storage of regularly inflected forms. *Memory & Cognition*, **14**, 17–26.
- Westermann, G. (1998). Emergent modularity and U-shaped learning in a constructivist neural network learning the English past tense. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, pages 1130–1135, Hillsdale, NJ. Erlbaum.

# Contextually Representing Abstract Concepts with Abstract Structures

Katja Wiemer-Hastings (KWIEMER@LATTE.MEMPHIS.EDU)  
Department of Psychology, University of Memphis, CAMPUS BOX 526400  
Memphis, TN 38152 USA

Arthur C. Graesser (GRAESSER@MEMPHIS.EDU)  
Department of Psychology, University of Memphis, CAMPUS BOX 526400  
Memphis, TN 38152 USA

## Abstract

This paper proposes that abstract concepts are represented as contextually derived structures. According to this *abstract structure theory*, abstract concepts are related to mostly temporal and spatial structures that underlie and can be extracted from concrete situations. Linguistic context elements of abstract concepts, such as verbs and prepositions, express these structures, and can thus aid in the acquisition of such concepts. The paper presents results from a corpus study that supports the hypothesis, and discusses implications.

We propose that abstract concepts, such as *faith* or *notion*, are represented as abstract structures, which represent particular contexts in which they occur. Abstract concepts are not directly perceivable, but are often used in verbal descriptions of situations, or in utterances related to a situation. Such utterances have to unambiguously point out the entity that is referred to by the abstract noun. We argue that the relevant abstract structures can be inferred from their linguistic context, in particular, from verbs and prepositions used with the abstract nouns.

## Concept Constraints and Contextual Similarity

Similar concepts occur in similar linguistic contexts. Miller and Charles (1991) report that contexts of similar concepts are more often classified as belonging to the same concept. For example, the sentences *the patient rang for the \_\_\_\_\_* and *the \_\_\_\_\_ gave the patient an injection* both suggest that the word *doctor* or *nurse* would complete the sentence well. We found that a neural network could be trained to correctly select one out of seven abstract concepts based on linguistic context information in 72% of test cases (Wiemer-Hastings, 1998). An approach to learning verbs from context was described by Wiemer-Hastings, Graesser, and Wiemer-Hastings (1998; see also Hastings, 1994). This work shows a close link between context of use and concepts.

## Acquisition from Contexts

The relationship between concept and context similarity has implications for language acquisition: Unknown words can be learned from context. Berwick (1989) discussed a system that acquired word meaning based on contextual similarity of the word to the contextual representation of

familiar words. Sternberg and Powell (1983) have shown experimentally that human participants could infer unknown word meanings from the context of a short text passage.

So, linguistic context provides useful information for language acquisition. In the acquisition of concrete concepts, this information is a "bonus" added to more directly useful information that the learner has access to, that is, information about perceptual or functional characteristics, as well as uses of the objects in situations. It is not even clear to what extent the linguistic information actually *adds* to this perceptual information. It might instead just reflect the information contained in the perceptual context of use, thus being redundant with it.

Linguistic context (in particular, syntactic context) has been shown to facilitate the acquisition of verbs (e.g., Fisher, 1994). Levin (1993) has provided a classification of semantic verb classes based on syntactic verb frames. Thus, much critical lexical information can be extracted from the linguistic context. This is important when the available information is largely *confined* to linguistic information. This is the case for abstract noun concepts, which refer to complex situations and relationships within these situations. If linguistic context is informative for verbs, it should be helpful all the more to acquire abstract concepts, such as *ignorance* or *strategy*. Indeed, Quine (1960) has argued that abstract concepts must be acquired on the basis of linguistic information alone. In support of this hypothesis, we found that abstract concepts can be distinguished pretty reliably based on semantic and syntactic aspects of their context (Wiemer-Hastings, 1998). If the hypothesis holds, then it should also be possible to identify linguistic context elements that are related to abstract concept meanings, thus that they co-occur with similar frequency with similar abstract concepts, and do not co-occur with dissimilar abstract concepts.

## Context Dependence

Clearly, the characteristics of entities systematically constrain contexts in which they can occur. However, this statement implies that entities are something that is given a priori, and contexts are selected based on the entities. This may be true for concrete entities, such as furniture items. Concrete entities exist independent of aspects of particular

contexts. A chair is still a chair if it occurs in a new context. Concrete concepts have characteristics that put concrete constraints on how we interact with them. Their characteristics thus determine, to some extent, their use. In this sense, concrete entities are to some extent independent of contexts.

With respect to abstract concepts, it seems that the relationship between context and concept constraints is reversed: their use is not determined by their characteristics, but their characteristics are inferred from their use. It seems that the context is the a priori given in this case, whereas abstract concepts are used to describe and make sense of complex situations and processes. Abstract concepts do not exist independently. They can only "happen" in particular contexts. An *idea*, for example, is conceived mentally, and can be expressed in words. Its consequences can be observed in context. It has the pragmatic function of overcoming some obstacle. In a slightly different context, one may call the concept a *suggestion* or *recollection*, instead of an *idea*. Similarly, *truth* is a characteristic ascribed to a statement that describes a particular state of affairs correctly (see Barsalou, 1999). If the state of affairs is different from a statement, the concept truth does not apply anymore.

This difference in the relation between abstract and concrete concepts and their contexts also affects language acquisition. We first acquire words for concrete entities. Later, we learn that there are abstract concepts, but we need to infer their characteristics from the contexts in which they are used. That is, the context is processed before the abstract concept can be understood. For abstract concepts, we could accordingly postulate that they are understood to be similar to the extent that they are used in the same linguistic contexts.

### **Operationally Defined Context**

Context is a complex notion. In contrast to verbs, there are no particular syntactic frames associated with abstract concepts. However, if abstract concepts put constraints on admissible contexts of use, then some of their contextual elements should reflect important semantic aspects of the concepts. Are there any indications as to what context elements may play a role?

### **Explaining Contextual Effects with Scripts**

Context effects on concept processing (e.g., on the speed in word recognition) have been shown in many studies. A series of experiments showed that such effects are mostly due to global context rather than local context elements (Hess, Foss, & Carroll, 1995). Sharkey and Mitchell (1986) suggested that context effects on lexical processing are mediated by scripts (Schank & Abelson, 1977). Scripts are schemata of actions, such as seeing a doctor, or eating out in a restaurant. According to Sharkey and Mitchell, associated words activate target words not through associations (i.e., through strong connections in a semantic network), but by activating a script that in turn activates the target word.

This hypothesis has received empirical support. However, it cannot easily be applied to abstract concepts. Consider, for example, the difficulty in selecting a script for *idea*. One would likely assign the concrete concept *menu* to the *restaurant script*, but concepts like *idea* can occur in a wide, almost arbitrary, variety of concrete situations such as represented by scripts. Yet, there are particular aspects of contexts that must be true for an abstract concept to apply. In the case of *idea*, for example, the concept typically occurs in a context where there is a problem or obstacle of some kind; an agent who reflects or discusses possible ways to overcome the obstacle; and a thought or utterance (the *idea*) that leads to the problem-solving action. A temporal sequence with causally related elements emerges from this scenario. We collectively call such sequences and structures for other kinds of abstract concepts *abstract structures*.

### **Abstract Structures**

Abstract structures represent integrated processes, events, or particular relationships in situations. They are *abstract*, in that they apply to situations with different concrete aspects. They are *structures*, in that they organize sets of entities in a situation with respect to the causal, temporal, spatial and other relations that hold between them. The concept is similar to schemata or scripts. However, abstract structures are more abstract than scripts, so we opt to use a different term here, to avoid the association with concrete situations.

Abstract concepts have a temporal and spatial dimension. The temporal dimension is critical, because it represents the ontological class of a concept (i.e., whether the concept is a point-like event, a process, or a state), and the sequencing of events within a structure. The representation of many abstract concepts requires information about their time course, for example, *discussion* or *sequence*. Causal aspects usually depend on temporal information as well. For example, concepts like *effect*, *consequence*, *impact* etc. require some temporally preceding entity or event. In principle, this suggested representation format is compatible with a perceptual approach, which integrates perceptual aspects beyond vision. One such theory has recently been proposed by Barsalou (1999). He proposes a combination of situation percepts with introspective information to represent abstract concepts. Our approach does not challenge this view, but approaches the representation from a linguistic point of view, and focuses on dynamic aspects of the concepts within context.

This paper describes first results of an investigation of this *abstract-structure* hypothesis. If linguistic context serves as a basis for abstract concept acquisition, then it is necessary that it reflects critical relationships in the situation context, and thus directs the learner's attention to the relevant aspects in this situation to identify the referent of the abstract noun. In relation to this reasoning, in particular we test the following prediction: Temporal, spatial and other aspects of the abstract structure related to an abstract concept are expressed in, and can be inferred from, its linguistic context.

## Context Elements

According to the abstract structures hypothesis, there should be linguistic context features that express causal, temporal, and other information. With this in mind, we examined the linguistic contexts of abstract concepts selectively with respect to such elements. What context elements are likely to reflect spatial, temporal and other relationships between the agents and entities in a situation?

This paper discusses two elements of context: verbs, and prepositions. Selecting two groups of lexical items clearly does not follow the view that it is global context that is critical with respect to concept representation. However, it appears worthwhile to examine context elements that can be easily identified and test these first, instead of attempting to identify more complex structures in text. The selection of verbs and prepositions follows directly from our hypothesis that abstract concepts are represented as contextual structures. Both verbs and prepositions express the relationships pertinent to abstract concepts, according to our hypothesis.

Verbs describe the way in which agents interact with each other and with entities, and convey aspects relevant to abstract concepts, such as events and causality (Basili, Pazienza, & Velardi, 1996). They express causal (e.g., *cause, evoke, produce, lead to*, etc.), temporal (e.g., *follow, end, begin*, etc.) and spatial information (e.g., *leave, hide, bring, remove*). Verbs also express other important aspects of abstract concepts related to agent - object relations, such as evaluations (e.g., *like, want*, etc.), verbal expression (e.g., *announce, explain, suggest*, etc.), and others. The central role of verbs with respect to the processing and identification of agents and objects has been shown in a lot of research, even if not specifically for abstract concepts. Altmann and Kamide (1999), for example, show that verbs guide our attention to particular aspects of a situation, because they lead us to expect what particular kinds of entities will be made reference to subsequently. Whereas this finding generalizes to entities outside the visual domain is an open question, but it is a possibility.

Prepositions can explicitly be classified with respect to the same dimensions (see Table 1). Considering these two context elements in the linguistic context of abstract concepts, we examined the question if the verbs and prepositions that occur in the contexts of particular abstract concepts express semantic aspects of the abstract concepts. We predicted that if they do, then similar abstract concepts should co-occur with similar kinds of verbs and prepositions with similar frequency.

## Corpus Analysis of Abstract Concept Contexts

In order to test what kinds of verbs and prepositions occur with abstract concepts, one has to consider a representative number of context samples. For example, one would expect that very general predicates (such as *think about, talk about*) occur with all kinds of concepts, and to provide little basis for differentiation. In order to get at the systematic relationships between verb and preposition context and

abstract concepts, we must therefore look at a variety of contexts and record two aspects: a) patterns of co-occurrence between abstract concepts and verb / preposition classes, and b) the frequencies of the co-occurrences.

We conducted a corpus analysis to obtain both measures. Corpus analyses have been used frequently since large databases of naturally occurring text have become available electronically. Boguraev and Pustejovsky (1996) express the power of corpus analyses proposing that "Text corpora reflect language as it is used and evolves; by studying regularities of use and patterns of behavior of words, which only emerge from analysis of very large samples of text and / or speech, it is possible to induce (among other things) lexical properties (...)" (p. 5). The power of co-occurrence patterns in text for representing semantic aspects of concepts and texts has been demonstrated by the success of systems such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) and HAL (e.g., Burgess & Lund, 1997). However, since LSA uses co-occurrence information among all elements of text, it does not tell us much about *which* elements of context play a role in relation to individual kinds of concepts.

## The Corpus

A sample of thirty abstract nouns was selected randomly, but so that different ontological classes (e.g., state, process, event, emotion) were represented. The sample included the words *accident, agreement, approach, aspect, attempt, effect, decision, discovery, discussion, essence, fear, freedom, goal, idea, ignorance, impression, indifference, invention, miracle, notion, plan, pride, principle, recollection, result, silence, strategy, surprise, truth, and wisdom*. We collected our corpus from NexisLexis, an online database that contains full texts from newspapers, magazines and other sources, representing a wide range of topics. For each abstract noun, 250 sentences were collected that contained the particular abstract noun. The sampling was principally random. However, we made sure that no sentences were repeated, and that a variety of topics was represented. Altogether, we collected a corpus of 7500 sentences.

## Encoding Co-occurrence

For each noun, we looked at every sentence and recorded the verb and preposition that occurred in direct relation to the abstract noun. Verbs and prepositions were recorded with information about whether they preceded or followed the abstract noun. We additionally counted the frequency with which each verb and preposition occurred<sup>1</sup>.

This method yielded a large number of verbs and prepositions (about 1700). The raw data would have yielded long context vectors with very low average frequencies. Analyses based on such vectors would

---

<sup>1</sup> Verbs and prepositions were not recorded in combination. In many sentences, only one of the two occurred. Further, the combinations may lead to an enormously extensive data space that would be hard to reduce by classification.



presumably be distorted by noise. Therefore, we classified our recorded verbs and prepositions into semantic classes. Verbs were classified into the semantic classes constructed by Levin (1992). Her system contains 37 semantic verb classes that occurred in our corpus. They include verbs of *occurrence* (e.g., *happen*), *possession* (*give*, *obtain*), *communication* (*describe*, *announce*), and *psychological verbs* (*amaze*, *disturb*). We only considered verbs that could clearly be classified consistent with these classes.

Prepositions were classified into spatial, temporal, causal, modal, propositional, referential and possessive information, and further sub-classified within these groups (see Table 1).

Table 1  
Classification of prepositions

Preposition class	Subclass	Examples
Spatial location	distance	near, by, far from
	relation to 1 object	on, in, behind
	relation to 2 objects	between, amidst
Spatial motion-direction	related to origin	from, out of
	related to destination	into, towards
	related to path	through, across
Temporal	related to future	until, prior to
	related to past	after, since
	related to presence	during, while, at
Temporal	expressing time-range	after (time-range)
	expressing point-in-time	on, at
Causal	related to factor	due to, because of
	related to effect	in order to
	related to means	through, whereby
Modal / concomitative		with
Propositional Referential		about, on
	inclusive	with regard to
	exclusive / adversatives	
Possessive		except, contrary to
		of, from

### Abstract Concept Context Vectors

A noun-context element matrix was constructed that listed abstract nouns against context element classes (Table 2). The context elements contained the verb classes  $1$  to  $n$  after Levin (1992) that had non-zero occurrences in the corpus, followed by preposition classes  $1$  to  $n$ . The cells in the matrix contained the co-occurrence frequency in the corpus. Context elements were represented twice. The first time, the co-occurrence data only count times that the abstract noun *preceded* the particular context element class in context. The second time counted the times the context elements were followed by the abstract noun, respectively.

Table 2  
Co-occurrence matrix for context elements

	Verb class $1$	...	Preposition class $n$
Abstract noun 1	5	...	0
...	...	...	...
Abstract noun 30	0	...	58

### Evaluation

Thirty context-vectors were constructed based on the co-occurrence matrix, one for each abstract concept. Each vector represents how often a particular abstract noun occurs with different kinds of verbs and prepositions in context. These vectors were used to evaluate the hypothesis that linguistic context, in the form of verbs and prepositions, reflects semantic aspects of abstract concepts. If this hypothesis is correct, then the contexts of similar abstract concepts should be similar, resulting in a significant correlation of the cosines of the context vectors with human similarity judgments of the corresponding abstract concept pairs.

We tested what context information is related to abstract concepts in particular. Six different vectors were constructed to represent various aspects of context. We built vectors to represent only prepositions, only verbs, or both. For each of these, there were two versions: an extended, "ordered" version that contained word order information, and a short, "unordered" version that ignored word order information. For the ordered version, we counted co-occurrence separately for context elements preceding versus following the target noun. For the unordered vectors, co-occurrence counts within the verb and preposition classes were collapsed to disregard word order. This vector version thus represents merely how frequently which kinds of verbs and prepositions can in general co-occur with the abstract noun.

To test whether the verb and preposition context relates to abstract concepts, we compared the similarity of the context vectors to similarity judgments of the corresponding abstract concept pairs, provided by human raters. The 30 abstract concepts resulted in 435 vector / abstract concept pairs. Correlations were computed between two similarity measures: human similarity judgments of the concept pairs, averaged across 33 raters, and vector cosines for the context vector pairs. Both measures range from 0 (maximally dissimilar) to 1 (maximally similar).

### Results

Table 3 shows the correlation coefficients. The cosines of the full vectors, containing verb and preposition co-occurrences and word order information, were significantly correlated with the human ratings ( $r = 0.22$ ,  $p < 0.01$ ).

Table 3  
Correlation coefficients between the vectors and human similarity judgments

	Ordered Vectors	Unordered Vectors
Verbs and prepositions	0.22	0.17
Verbs only	0.22	0.18
Prepositions only	0.20	0.13

This correlation is modest, but highly significant. It indicates that more similar abstract concepts tend to have

similar co-occurrence patterns with verbs and prepositions. This coefficient is higher than the average human interrater correlation coefficient, computed for a random sample of 100 coefficients (mean  $r=0.18$ ,  $SD=0.16$ ).

To estimate the relative relevance of verbs and prepositions, we computed the vector cosines separate for verbs and prepositions. The cosines for both were significantly correlated with the human ratings ( $r_{prep} = 0.20$ ,  $p < 0.01$ ;  $r_{verb} = 0.22$ ,  $p < 0.01$ ). Thus, the co-occurrence patterns of both context elements, verbs and prepositions, are significantly related to abstract concept similarity.

Further analyses tested the relevance of word order information. The question here is whether the correlation is due to the information which verb and preposition classes can co-occur with an abstract concept in general, or whether the word order is critically important. Word order information may play an important role in abstract concept representation. For example, in the phrases "due to the discussion" versus "the discussion due to (...)" the prepositions express very different information about discussion. In the first example, the discussion causes some effect; in the second example, the discussion itself was caused by something.

We computed the cosines for the vectors that just represent co-occurrence with verb or preposition classes without separating co-occurrence counts according to word order. The cosines were correlated with the human ratings. The resulting correlations were significant ( $p < 0.01$ ), but the correlations were smaller than the ones obtained before. The correlation for the verb-only vectors was  $r=0.18$ ; the preposition-only vectors yielded a correlation of  $r=0.13$ . The combined verb and preposition vectors led to a correlation of  $r=0.17$ . Thus, word order does increase the correlation, especially in the case of prepositions.

We compared our results to correlations of the human similarity judgments with cosines from LSA for the same concepts. Since LSA takes into account all words in context, and for a lot more text, it is a good model to compare our vector matches to. In particular, if verbs and prepositions cover the important aspects of context related to abstract concepts, then the correlations of human ratings with LSA cosines should be comparable to the ones obtained in our study. If however the match with LSA cosines is substantially higher than our correlations, then verb and preposition context conveys only part of the relevant context information and other word classes should be included. We found that the correlation between human judgments and LSA cosines was significant ( $r_{LSA} = 0.23$ ,  $p < 0.01$ ), but not much higher than the correlations we obtained for our selective context vectors. This might mean that the correlation obtained from LSA is mostly due to the verbs and prepositions in the underlying text corpus. At least, the verb and preposition co-occurrence patterns can account for as much of the similarity ratings as LSA.

## Correlations of Verb and Preposition Context

In addition to these correlations with human ratings, we looked at a few correlations among the vector cosines. We found that the cosines of preposition vectors and verb vectors were significantly correlated ( $r=0.13$ ), but only if the vectors separated co-occurrence counts with respect to word order. In other words, similar abstract concepts tend to be preceded and followed by similar types of verbs and prepositions. This means that contexts with similar patterns of verb occurrence also resemble each other in the patterns of prepositions they contain.

## The Role of Ontological Information

It was mentioned before that verbs and prepositions convey ontological information (such as states, processes, events). To what extent can the human similarity judgments and context vectors be explained by two concepts being of the same as opposed to different ontological kinds? To test this, we created a Boolean variable that was "1" when both concepts in a pair were of the same ontological status, and "0" otherwise. We correlated this variable with the human similarity judgments, and with the context vector cosines.

The ontological status was significantly correlated to the human ratings,  $r=0.23$ ,  $p < 0.01$ . That is, the ontological status of two concepts may play a role in how people judge concept similarity. The ontological status was also related to the context vector cosines, but only to a selective group. First, it was correlated to the combined verb and preposition vectors. Interestingly, the correlation coefficient was exactly the same for the vectors containing information about word order and those not containing this information ( $r=0.13$ ,  $p < 0.01$ ). Furthermore, ontological status was correlated with the preposition-only vector cosines that did *not* contain word order information ( $r=0.10$ ,  $p < 0.05$ ).

This interesting result suggests that word order matters with respect to abstract concepts, but may be irrelevant, or even provide misleading information, with respect to the ontological status of the abstract concepts. The information represented by the preposition-only vectors without word order information simply reflects the frequency with which abstract concepts co-occur with the different kinds of preposition classes (Table 1). It makes sense that *statehood*, *eventhood*, etc. would be reflected in the kinds of verbs and especially prepositions that co-occur with the concepts. For example, event concepts may be surrounded by temporal prepositions such as *before* and *after*, whereas process nouns may be marked by prepositions such as *while* or *during*. Ontological status was also significantly correlated to the LSA cosines ( $r=0.12$ ) to a similar extent.

## Discussion & Implications

We have proposed that abstract concepts are represented by abstract structures that contain causal, temporal, spatial and other information pertinent to the abstract concept. Assuming that these contextual aspects are reflected by the verbs and prepositions that co-occur with particular abstract concepts, we have conducted a corpus study that examined

whether similar abstract concepts co-occur with similar patterns of verbs and prepositions. We found that the similarity of context vectors based on these word classes were significantly correlated with the similarity of the abstract concepts occurring in these contexts. That is, similar abstract concepts have similar co-occurrence patterns with verb and preposition classes. We found that the correlations of abstract concept similarity was not much higher with cosines of LSA vectors, indicating that verbs and prepositions may indeed be the most informative context elements with respect to abstract concepts. We did find a pretty substantial correlation between the verb and preposition vectors, however. This correlation suggests an alternative interpretation, namely, that different aspects of context are related to abstract concepts but that they are interrelated, thus that they do not add any further information to distinguish abstract concepts.

In future work we plan to examine to what extent a particular set of verb and prepositions can be used to identify the abstract structure corresponding to an abstract concept, and to kinds of abstract concepts (e.g., *states* versus *events*). Another interesting question is how many verb and preposition classes are most informative in relation to abstract concepts. Perhaps the correlations could be improved by choosing more classes with finer distinctions, or conversely, by reducing the class space even further.

Another interesting question is whether our abstract structure theory can explain context effects as reported by, for example, Schwanenflugel and Shoben (1983). Contexts preceding abstract concepts may instantiate the particular abstract structure underlying their representation and thus mediate priming effects. This could be tested by setting up contexts that differ in the amount of information they provide with respect to the relevant abstract structure.

### Acknowledgments

This work was supported in part by the National Science Foundation, Learning and Intelligent Systems Unit, under grant SBR-9720314. We further wish to thank three anonymous reviewers for their very helpful comments on an earlier draft of this paper, and for useful suggestions for follow-up work.

### References

- Altmann, G.T.M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73, 247-264.
- Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-609.
- Basili, R., Pazienza, M.-T., & Velardi, P. (1996). A context driven conceptual clustering method for verb classification. In B. Boguraev & J. Pustejovsky (Eds.), *Corpus processing for lexical acquisition* (pp. 117-142). Cambridge, MA: MIT Press.
- Berwick, R.C. (1989). Learning word meanings from examples. In D. Waltz (Ed.), *Semantic structures: Advances in natural language processing* (pp. 89-124). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Boguraev, B., & Pustejovsky, J. (1996). Issues in text-based lexicon acquisition. In B. Boguraev & J. Pustejovsky (Eds.), *Corpus processing for lexical acquisition* (pp. 3-17). Cambridge, MA: MIT Press.
- Burgess, C., & Lund, K. (1997). Representing abstract words and emotional connotation in a high-dimensional memory space. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 61-66). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Fisher, C. (1994). Structure and meaning in the verb lexicon: input for a syntax-aided verb learning procedure. *Language and Cognitive Processes*, 9, 473-517.
- Hastings, P. (1994). *Automatic acquisition of word meaning from context*. Doctoral dissertation, University of Michigan.
- Hess, D.J., Foss, D.J., & Carroll, P. (1995). Effects of global and local context on lexical processing during language comprehension. *Journal of Experimental Psychology: General*, 124, 62-82.
- Keil, F.C. (1979). *Semantic and conceptual development*. Cambridge, MA: Cambridge University Press.
- Landauer & Dumais (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
- Levin, B. (1993). *English verb classes and alternations*. Chicago: University of Chicago Press.
- Miller, G.A., & Charles, W.G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6, 1-28.
- Quine, W.V.O. (1960). *Word and object*. Cambridge, MA: MIT Press.
- Schank, R.C., & Abelson, R. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Schwanenflugel, P.J., & Shoben, E.J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9, 82-102.
- Sharkey, N.E., & Mitchell, D.C. (1985). Word recognition in a functional context: the use of scripts in reading. *Journal of Memory and Language*, 24, 253-270.
- Sternberg, R.J., & Powell, J.S. (1983). Comprehending verbal comprehension. *American Psychologist*, 38, 878-893.
- Wiemer-Hastings, K. (1998). Abstract noun classification: Using a neural network to match word context and word meaning. *Behavior Research Methods, Instruments, & Computers*, 30, 264-271.
- Wiemer-Hastings, P., Graesser, A.C., & Wiemer-Hastings, K. (1998). Inferring the meaning of verbs from context. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 1142-1147). Mahwah, NJ: Lawrence Erlbaum Associates.

# Adding syntactic information to LSA

Peter Wiemer-Hastings

Peter.Wiemer-Hastings@ed.ac.uk

School of Cognitive Science / ICCS, Division of Informatics  
University of Edinburgh, Edinburgh EH8 9LW Scotland

## Abstract

Much effort has been expended in the field of Natural Language Understanding in developing methods for deriving the syntactic structure of a text. It is still unclear, however, to what extent syntactic information actually matters for the representation of meaning. LSA (Latent Semantic Analysis) allows you to derive information about the meaning without paying attention even to the order of words within a sentence. This is consistent with the view that syntax plays a subordinate role for semantic processing of text. But LSA does not perform as well as humans do in discriminating meanings. Can syntax be the missing link that will help LSA? This paper seeks to address that question.

## Introduction

In the beginning, there was syntax. And it was good. But it did not give us what we really want to know about a text — what it means. Then there was latent semantic analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990, LSA), which provided a means of comparing the “semantic” similarity between a source and target text, and thereby giving some idea of meaning of the source. That was good too, almost as good as humans in a simple task, but not quite. Because LSA pays no attention to syntax at all — not even word order — one promising approach to improving LSA is by giving it some of the information that is provided by syntax. Knowledge about the syntactic structure of a sentence provides information about the relationships between the words: which words modify which other words, and the relationships between verbs and their arguments or thematic roles. The research presented here is an attempt to evaluate the benefits of providing LSA with thematic role information which comes from syntactic knowledge.

## Previous work

The primary goal of the AutoTutor project (Graesser, Franklin, Wiemer-Hastings, & the Tutoring Research Group, 1998; Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999) is to model human tutorial dialogue. It is based on studies of the discourse patterns of human tutors during tutoring sessions (Person, Graesser, Magliano, & Kreuz, 1994). These analyses have shown that human tutors do not have complete understanding of their students’ answers to questions, but they do get an approximation. For AutoTutor, LSA provides such

approximate understanding of student inputs by comparing them to expected answers, and using the LSA cosines as a metric of the extent to which the student entered what was expected.

We evaluated this approach by randomly selecting a set of 8 student answers to each of 24 questions in our domain of computer literacy (Wiemer-Hastings, Graesser, Harter, & the Tutoring Research Group, 1998). We asked human raters to evaluate these answers by providing an aggregate measure of the percentage of student answer propositions that “match” some expected answer proposition. Proposition was defined loosely as an atomic sentence. Match was left to the human raters to define. Then we performed the same analysis with LSA, modeling the match function by adjusting the cosine threshold. The best performance was realized with a 200-dimensional space with a cosine threshold of 0.5. This provided a correlation of  $r = 0.49$  with the average rating of the human judges. Because the distribution of ratings was skewed, we also calculated Cronbach’s alpha. The average alpha score between human raters was  $\alpha = 0.76$ . The alpha score between LSA and the average human rating was  $\alpha = 0.60$ . These results were very encouraging. LSA provided much of the discrimination shown by human raters, enough to use in the AutoTutor system. It could however, be improved.

The obvious information source that LSA ignores is syntax. It is a “bag-of-words” approach, simply adding together term vectors to make a vector for a text. This paper is an attempt to identify whether the addition of syntactic knowledge can strengthen LSA judgments.

## Related work

Partially as a result of the Behaviorist movement in psychology, linguistics and natural language processing focused for a long time primarily on the syntactic structure of sentences (Chomsky, 1981, for example). In the 70’s and 80’s, Schank sought to change this by claiming that semantics alone was sufficient (Schank & Riesbeck, 1981, for example). More recently, researchers from psychology have championed LSA as both a technique for determining the meaning of texts and as a model of human language.

Much of the recent interest (and controversy) regarding LSA can be traced to Landauer, Kintsch, and colleagues. They imported LSA from the realm of information retrieval and hailed it as part or parcel of a psy-

chological model of language understanding. Landauer and Dumais (1997) described LSA as a model of human language acquisition, using it to explain how the pace of lexical acquisition apparently outstrips the exposure to new words. Landauer has gone on to claim that LSA is a complete model of language understanding (Landauer, Laham, Rehder, & Schreiner, 1997). He explains away the existence of syntax by suggesting that it is only there to simplify the computational complexity of getting the words into an LSA-like representation in the first place.

Other psychologists have stressed the role which syntax can play in lexical acquisition. The syntactic bootstrapping (Gleitman & Gillette, 1994) theory shows how pre-verbal children can use their knowledge of syntax to help guide their acquisition of verbs.

Kintsch (1998) has appended LSA to his Construction/Integration model of text understanding as the semantic component. LSA provides a sort of spreading activation-like inclusion of related concepts when new information is integrated into a knowledge structure. This allows the system to perform a type of inference, making, for example, “driver” and “computer” available when “bus” is mentioned in a text.

In other related psychological approaches, MacDonald has proposed a used a variant of LSA to predict semantic priming (McDonald, 2000). And Ramsar and colleagues have used LSA to model analogical reasoning (Packiam-Allaway, Ramsar, & Corley, 1999).

The HAL system (Burgess & Lund, 1997) is similar to LSA in the sense that it is based on co-occurrences, but word order information enters the representation space through a weighting mechanism: A co-occurrence is weighted more heavily the fewer words intervened between the two words, within a window of usually ten words. So, two words that co-occur in immediate adjacency are weighted most strongly. This is not syntax, but it does grant some sensitivity to word order.

Burgess and Lund replicated earlier work by Finch and Chater (Finch & Chater, 1992) which showed that by applying a high-dimensional method to clustering the co-occurrences of words in a corpus, it is possible to infer lexical categories that correspond well with standard syntactic theories. Finch and Chater also showed that you could use these categories to infer basic grammatical rules (see also (Siskind, 1996; Christiansen & Chater, 1999) for other corpus-based approaches to acquiring such information). Thus, there seems to be sufficient information in a corpus of text to statistically infer something about the syntactic structure of that corpus.

This does not mean, however, that a technique like LSA already has the type of syntactic information that we are attempting to incorporate here. For any particular sentence, LSA creates a vector just based on the bag of words that are in that sentence. It has no information about the word order within that sentence or about the relationships between the words.

## Approach

Our initial success with LSA and the potential for improvement led us to examine how additional information

could be provided. One obvious possibility is to use more classical natural language understanding techniques as a pre-filter for LSA. The idea is to use parsing, anaphora resolution and other dialogue-processing techniques to prepare chunks of text for LSA to process semantically. Alternatively, this could be viewed as using LSA as the semantic component of a classical natural language understanding system.

We preprocessed the student sentences and the expected answer sentences in the following way: First, we performed a basic syntactic segmentation of the sentences. Although there are surface-level parsing methods generally available (Abney, 1996, for example), their grammars must be modified to conform to the application. If this approach is successful, we will move to automated methods. For this test, we simply separated the sentences into atomic clauses or propositions, and then segmented them by hand, breaking them down into strings which corresponded to:

- subject noun phrase
- verb, including adverbs and adverbial phrases
- object noun phrase (when applicable)

This provides two types of additional information:

1. the grouping of words which belong together into “components”
2. the pseudo-semantic role of the components as derived from syntactic argument structure

Second, we resolved anaphora in the sentences, replacing pronouns by their antecedents. Finally, when there was a conjunction, we distributed the arguments. For example, if there was a sentence like, “Subject verb object1 and object2”, it was broken into (“verb” “Subject” “object1”) and (“verb” “Subject” “object2”), using a verb-prefix notation.

We made no attempt to do any other processing based on discourse relations for two reasons. First, LSA normally ignores “stop words” like “if” and “because” anyway. Second, extracting any more complex discourse relations would require the use of semantic understanding which is the goal of this process. Table 1 gives some examples of sentences and their representations in this scheme.

There are three competing hypotheses of the effect on similarity judgments of using this additional information along with LSA:

1. Component grouping will increase discrimination because it adds information — the role of different components.
2. Component grouping will hurt discrimination because LSA works better on longer strings.
3. Component grouping will hurt grouping due to some complexity of combining individual component similarity scores.

Table 1: Example sentences and their representations

RAM stores the instructions to your programs.	(“stores” “RAM” “the instructions to your programs”)
If the new motherboard uses the same type of RAM, you can just take the SIMMs out of your old motherboard and install them in your new motherboard.	(“if uses” “the new motherboard” “the same type of RAM”) (“can just take out of your old motherboard” “you” “the SIMMs”) (“and install in your new motherboard.” “you” “the SIMMs”)

The following section describes our first attempt to test these hypotheses using a straightforward combination of the between-component cosines.

## Experiment 1

Given this type of representation, there remain a variety of ways to calculate the overall similarity between propositions based on the similarities of the components. In experiment one, we took the most straightforward approach, simply averaging the cosines of the respective components. In other words, we calculated the LSA cosine between the verb string from a student proposition and the verb string from an expected answer. We repeated this for the other sentence components. If there was an object string for one sentence and not for the other, a component score of zero was recorded. Then we averaged across the (normally two or three) components of the propositions.

Next, we aggregated the scores for each student answer proposition by taking the maximum average cosine across the different expected answer propositions. As in the previous experiment, the final score was the percentage of student answer propositions that achieved a score above the empirically-determined threshold. We tested thresholds between 0.05 and 0.95 in 0.05 increments. We measured the correlation between the LSA scores with the human ratings.

The best correlation was  $r = 0.18$  (not significant), with the threshold at 0.10.<sup>1</sup> This is far below the performance of the previous approach which used LSA to compare entire sentences. Thus, these findings do not support hypothesis 1.

The decrease in the overall performance could potentially be due to the difference between comparing sentences (as in the original experiment) and comparing propositions. But the aggregate score essentially factors that out to the extent that length of string does not affect LSA discrimination. String length does affect LSA discrimination however. Rehder et al (1998) used LSA to assess the domain knowledge of essay writers. To determine the effect of essay length on LSA discrimination, they truncated each essay after 10 words, 20 words, and so on. Below 60 words, they found fairly poor perfor-

mance. The performance steadily increased from there up to their 200 word maximum. Despite this finding, we have found performance approaching human abilities on our tutoring texts which have an average length of 16 words. Thus, we thought that any minor reduction in performance due to length would be offset by increased information provided by the pre-processing.

Analysis of cases of disagreement between LSA and the human raters showed that some items got very bad scores because one component consisted only of a “stop word” — a member of a list of 440 common words that includes prepositions, pronouns, and some very common adjectives, adverbs, verbs, and nouns. For example, one student proposition has a verb component group consisting of the string, “stores”, and the expected answer has the verb string, “has”. In this case (“RAM stores information being worked with”), the meanings of these two verbs are quite similar. But because “has” is on the stop word list, it has no representation in the LSA space, and the cosine comparison returns a value of 0.

On the other end of the spectrum, there was often an exact match between the subjects. For example, “RAM” and “CPU” are frequent subjects which, if they match at all, tend to match exactly, getting a 1.0 cosine. Because average “good” cosine matches are often in the 0.4 to 0.6 range, this tends to inflate the cosine average. This is especially the case for intransitive sentences where there are only two components. At the threshold that provided the best correlation with human raters, 0.10, the verb string only had to match at the 0.20 cosine level to put the entire proposition over threshold.

Another factor which seemed to affect the ratings was the fact that there are so different ways in which the same content can be expressed in natural language. For example, “RAM stores things being worked with” should have a fairly high semantic match for “The CPU uses RAM as a short-term memory storage” (whole string LSA cosine = 0.48). But because the components do not line up at all in this approach, the cosine average score is 0.03.

Based on these analyses, and under the hope that hypotheses 3 was the case instead of hypothesis 2, the approach was modified as described in the next section.

## Experiment 2

As previously mentioned, the shortness of the subject components seemed to have an inordinate effect on the overall scores. The average number of words in subject components was 1.6, and many subject strings include stop words like “the” which do not contribute to LSA

<sup>1</sup>Due to the tediousness of pre-processing the sentences by hand, these results were only calculated on the first third of the test set. Analyses of the correlations on the original task on this part of the test set showed that it had lower performance ( $r = 0.32, p = .01$ ), but not as low as the results of experiment 1. Immediate future work will be to process the rest of the test set.

cosines. Because of this, we tested in experiment two, an alternative scoring strategy. In this strategy, the score between two propositions was calculated as follows:

If there is a suitable match between the subjects, then return the average of the cosines of the other components.

Here, “suitable match” was defined as either a cosine of  $0.7^2$ , or a cosine of zero. In theory a zero cosine means a complete lack of semantic similarity. In practice, however, the cosine is only exactly 0 when one of the strings is empty modulo stop words. Thus, this allows the matching of vague subjects like “you”.

There are psychological theories of discourse which (vaguely) support this approach. One is the Given-New distinction of referents in discourse (Clark & Haviland, 1977; Brennan, 1995). The theory includes a discourse processing strategy in which the hearer searches the prior discourse context for an antecedent for Given information which is commonly the syntactic subject of a sentence. The rest of the sentence is New information which is attached to the antecedent. In our approach, we filter out expected answers which do not have matching Given information. Then we rate the similarity with the remaining items based on the similarity of the New information.

For this approach, the results were better than for experiment 1. The maximum correlation between the system and the human raters was  $r = 0.24$ , ( $p = 0.06$ ). This still does not approach the level of performance of the original system, however. This led us to attempt to address the other concerns raised above in experiment 3.

### Experiment 3

In experiment 3, we built on the Given-New approach presented above. This time, however, we joined the verb component of each proposition with its object component into one larger component. This corresponds to the VP in the basic  $S \rightarrow NP VP$  sentence, or to the predicate in the Subject/Predicate description of a sentence. Obviously this is a partial reversal from our previous approach of adding more information derived from syntax. The justification was to make the LSA comparisons less brittle with respect to distinctions between information in the verb and in the object.

The results for this approach were better than for experiment 2. The maximum correlation was  $r = 0.40$  ( $p < 0.01$ ), with a cosine threshold of 0.3. (The Cronbach's alpha score was  $\alpha = 0.49$ .)

Although this is an improvement, it is still not as good as the 0.49 correlation achieved by matching the entire sentence strings. Thus, these results do not support hypothesis 1. And taken together, their support for hypothesis 3 is ambiguous at best. This leaves us with the question: Why, when getting more information, does the discrimination still suffer?

### Discussion and Future work

In some ways our approach has been to find the best formula for combining the similarity ratings between the different components. The one which worked best, the one used in experiment 3, is non-linear. Perhaps a further search of combination methods can out-perform the basic LSA approach.

Taking the cue from other statistical NLP approaches and neural networks, perhaps we just have to find the right weight space which gives the best correspondence between the parameters (components) and the training data (human judgments). Ideally, if we were to attempt such an implementation, instead of aggregate human judgments over a set of items, we would have a rating for each pair of items. That would be much more demanding on the human raters, but would give more data to train the approach on.

Future work will focus on two fronts. First, we will acquire more data on which to evaluate this approach, both by adding more test items, and by getting additional human judgments as outlined above. Second, we will explore other methods of combining the added syntactic-derived information into LSA.

### Acknowledgments

This project was partially supported by grant number SBR 9720314 from the National Science Foundation's Learning and Intelligent Systems Unit. Many thanks to Mark Core for comments on this approach and to Katja Wiemer-Hastings for comments on the paper.

### References

- Abney, S. (1996). Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop*.
- Brennan, S. (1995). Centering attention in discourse. *Language and Cognitive Processes*, 10, 137–167.
- Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12, 177–210.
- Chomsky, N. (1981). Principles and parameters in syntactic theory. In Hornstein, N., & Lightfoot, D. (Eds.), *Explanation in Linguistics: The Logical Problem of Language Acquisition*. Longman, London.
- Christiansen, M., & Chater, N. (1999). Connectionist Natural Language Processing: the state of the art. *Cognitive Science*, 23(4), 417–437.
- Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In Freedle, R. (Ed.), *Discourse production and comprehension*, pp. 1–40. Earlbaum, Hillsdale NJ.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41, 391–407.

---

<sup>2</sup>0.5 was also tested, but it made a negligible difference

- Finch, S., & Chater, N. (1992). Bootstrapping syntactic categories using unsupervised learning. In *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*, pp. 820–825 Hillsdale, NJ. Lawrence Erlbaum Associates Inc.
- Gleitman, L., & Gillette, J. (1994). The role of syntax in verb learning. In Fletcher, P., & MacWhinney, B. (Eds.), *The Handbook of Child Language*. Blackwell, Oxford UK.
- Graesser, A. C., Franklin, S. P., Wiemer-Hastings, P., & the Tutoring Research Group (1998). Simulating smooth tutorial dialogue with pedagogical value. In *Proceedings of the 11th International Florida Artificial Intelligence Research Symposium Conference*, pp. 163–167. AAAI Press.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press, Cambridge, MA.
- Landauer, T. K., Laham, D., Rehder, R., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pp. 412–417 Mahwah, NJ. Erlbaum.
- Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.
- McDonald, S. (2000). *Environmental determinants of lexical processing effort*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.
- Packiam-Alloway, T., Ramscar, M., & Corley, M. (1999). Verbal versus embodied priming in schema mapping tasks. In *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society*. Laurence Earlbaum Assocs. Vancouver, Canada, August, 1999.
- Person, N. K., Graesser, A. C., Magliano, J. P., & Kreuz, R. J. (1994). Inferring what the student knows in one-to-one tutoring: The role of student questions and answers. *Learning and Individual Differences*, *6*, 205–229.
- Rehder, B., Schreiner, M., Laham, D., Wolfe, M., Landauer, T., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, *25*, 337–354.
- Schank, R., & Riesbeck, C. (Eds.). (1981). *Inside computer understanding*. Erlbaum, Hillsdale, NJ.
- Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.
- Wiemer-Hastings, P., Graesser, A., Harter, D., & the Tutoring Research Group (1998). The foundations and architecture of AutoTutor. In Goettl, B., Halff, H., Redfield, C., & Shute, V. (Eds.), *Intelligent Tutoring Systems, Proceedings of the 4th International Conference*, pp. 334–343 Berlin. Springer.
- Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In Lajoie, S., & Vivet, M. (Eds.), *Artificial Intelligence in Education*, pp. 535–542 Amsterdam. IOS Press.



# Categorization and the Ratio Rule

A. J. Wills (ajw43@cam.ac.uk)

Mark Suret (mbs22@cam.ac.uk)

I. P. L. McLaren (iplm2@cus.cam.ac.uk)

Department of Experimental Psychology, University of Cambridge,  
Downing St., Cambridge. CB2 3EB. U.K.

## Abstract

Many formal models of categorization adhere to two basic principles. First, the extent to which a stimulus is subjectively characteristic of a particular category can be represented by a single number. Second, the probability with which people choose a particular category label for a stimulus can be derived from these numbers via the Ratio Rule a.k.a. the Luce choice axiom (Luce, 1959). A categorization experiment employing artificial visual stimuli is presented and is shown to be problematic for these two principles. We demonstrate that, for the data presented here, the first principle can be retained if one replaces the Ratio Rule with a simple connectionist model.

## Introduction

Category learning is the task of acquiring the correct category label for each of a set of presented stimuli. The ability to categorize is central to cognition, and it has been the subject of a large number of studies. Over the last thirty years, these studies have typically involved abstract stimuli grouped into categories not necessarily definable in terms of a simple rule (e.g. Homa, Sterling, & Trepel, 1981; Medin & Schaffer, 1978; Posner & Keele, 1968). Psychologists have proposed a variety of formal models of our ability to learn and make decisions about such categories. The models differ in many respects - for example, the Generalized Context Model (Nosofsky, 1986) proposes the memorization of presented examples, whilst a number of other theories propose the formation of feature-category associations (e.g. Gluck & Bower, 1988; Kruschke, 1996; McClelland & Rumelhart, 1985). Despite such diversity, a great many theorists seem to agree on two fundamental principles. First, the extent to which a stimulus is subjectively characteristic of a particular category can be represented by a single number. We will refer to such numbers as *category magnitude terms*. Second, the probability with which a participant decides that a stimulus belongs to a particular category is determined by the Ratio Rule, a.k.a. Luce's Choice Axiom (Luce, 1959). In the current context, the Ratio Rule can be stated

$$P(i) = \frac{v_i}{\sum_{j=1}^n v_j}$$

where  $P(i)$  is the probability of choosing category  $i$  from  $n$  alternative categories and  $v_j$  is the category magnitude term for the  $j$ th alternative.

Theorists seldom justify their adoption these principles. Of greater concern is the fact that, as far as we are aware, there have been no direct tests of the Ratio Rule in the context of categorization. The evidence for the Ratio Rule, such as it is, comes from pair-comparison experiments and identification experiments (Bradley, 1954; Clarke, 1957; Hopkins, 1954). The evidence provided by such studies is equivocal at best, and some studies provide direct evidence against the Ratio Rule (e.g. Burke and Zinnes, 1965; Lamming, 1977).

Previously, we had made an unsuccessful attempt to disprove the Ratio Rule in the context of two-choice categorization decisions (Jones, Wills, & McLaren, 1998). The reason for this might have been that the Ratio Rule was basically correct for categorization decisions, but we suspected that it was because the predictions made by the Ratio Rule in a two-choice situation tend to be numerically close to the predictions of a number alternative accounts. Therefore in the experiment described here we tested a property more characteristic of the Ratio Rule - its predictions about probability ratios.

A long appreciated feature of the Ratio Rule is that it predicts that the ratio in which two alternatives are chosen is unaffected by the addition of a third alternative. For example, in a taste preference test between Coke and Pepsi, participants might choose Coke with a probability of 0.8. The Ratio Rule predicts that whilst the addition of lemonade might change the probability with which either Coke or Pepsi is chosen, it does not change the 4:1 ratio of probabilities.

The ratio we concentrate on in the current study is directly related to this property. It is the ratio between 1) the probability with which a particular response is made to a stimulus when three category labels are available and 2) the probability with which the same response is made to an equivalent stimulus when only two of the labels are available. For example, let's call the three labels A, B, and C, and say that A is the option which is disallowed in the two-choice example. Under the assumption that category magnitude terms for allowed alternatives are not affected by the number of alternatives available, it can be shown that the Ratio Rule predicts

$$\frac{P(B : B, C)}{P(B : A, B, C)} = \frac{v_A}{v_B + v_C} + 1 \quad \text{Measure 1}$$

This is not much use as it stands because we do not have any direct way of measuring the category magnitude terms, and different theories of categorization do not generally agree on how one might estimate the terms from observable data. The utility of Measure 1 lies in the fact that the Ratio Rule's predictions for the probability with which category A is chosen (when it is allowed) are similar in form. Specifically

$$P(A : A, B, C) = \frac{v_A}{v_A + v_B + v_C} \quad \text{Measure 2}$$

This correspondence means that, in situations where  $v_a$  is constant, any given change in  $(v_B + v_C)$  will produce the same direction of change in these two measures. One no longer needs to know what values the magnitude terms take. Instead one just needs to set up a situation where it is reasonable to assume  $v_A$  is constant across a set of stimuli. Then, to the extent different stimuli result in different values of  $P(A:A,B,C)$  and the 2 choice to 3 choice ratio (Measure 1), these differences must be in the same direction for both measures if the Ratio Rule is correct.

A number of similar correspondences can be set up, but we employ just one further here. Consider a second set of stimuli which are comparable to the first, except in their relative similarity to one of the three categories. In this situation it may be reasonable to assume that the magnitude terms for these two sets of stimuli differ only in respect to that category. Taking the category on which they differ as A, and the two magnitude terms as  $v_A$  and  $v_{A'}$ , the ratio of probabilities with which category B (or C) is chosen in response to these otherwise comparable stimuli is

$$\frac{P(B : A', B, C)}{P(B : A, B, C)} = \frac{v_A + v_B + v_C}{v_{A'} + v_B + v_C} \quad \text{Measure 3}$$

Note that in a situation where the two magnitude terms for category A can be assumed to be constant, and  $v_A > v_{A'}$  this third measure must exhibit the same direction of change as the other two. We investigated whether all three measures do indeed show the same direction of change in the context of a simple categorization task.

## Experiment

The experiment had two phases. In the training phase, all participants learned about the category membership of a set of novel, artificial stimuli. Each training stimulus belonged to one of three categories - A, B or C. In the transfer phase which followed participants were asked to determine the category membership of a set of test stimuli. Some participants were allowed to respond A, B or C, whilst for others the option A was disallowed. The stimuli presented in the transfer phase were designed to vary smoothly from being characteristic of category B and uncharacteristic of category C through to being characteristic of C and uncharacteristic of B. They were designed in this way so that (hopefully) the

three measures we were interested in comparing would be relatively smooth functions of the number of category B (or category C) elements. If reliable functions were found for our measures and these functions all exhibited the same direction of change then we would have evidence in support of the Ratio Rule. If the functions found exhibited different directions of change then this would be strong evidence against the Ratio Rule.

Participants with three response alternatives were presented with one of two sets of test stimuli. The members within a set were designed to be equally characteristic of category A. For one set they were somewhat characteristic of category A, whilst for the other they were uncharacteristic. All participants with two response alternatives received the test stimuli somewhat characteristic of A.

## Method

**Participants and Apparatus.** 36 Cambridge University students participated. They were tested individually in a quiet cubicle on a Acorn Risc PC microcomputer with a 14" color monitor. The computer's screen was at eye level, approximately 90 cm directly in front of where the participant sat. Responses were recorded via the "X", "B" and "M" keys of a standard PC keyboard. For this experiment the keys were re-labeled "A", "B" and "C" using bold red letters against a white background.



Figure 1: An example stimulus.

**Stimuli.** Each stimulus was a collection of twelve different small pictures (hereafter elements), arranged on an invisible four-by-three grid inside a 4.5 cm by 3.5 cm rectangle outline (see Figure 1 for an example). Every stimulus contained twelve elements drawn from a pool of 40 that we have used in a number of previous experiments (see Jones et al., 1998). No stimulus contained more than one copy of any particular element. At the beginning of the experiment, and separately for each subject, 12 elements from the pool were randomly designated as category A elements, a different 12 as category B elements, and a different 12 again as category C elements. The remaining four elements were designated as novel elements and were not employed in the training phase. Each training stimulus for each category was constructed by starting with all 12 elements characteristic of that category (e.g. category A elements for a category A training stimulus). Each element in the training stimulus then underwent a 10% chance of being replaced by a randomly chosen element from one of the other two sets (e.g. replaced by a B or C element in the case of a category A training stimulus). It was these modified stimuli that were

presented to subjects as training stimuli. This procedure produces training examples which are composed predominantly of elements characteristic of a particular category but also exhibit considerable variability in terms of the specific elements they contain. Ninety training examples were created for each subject, thirty from each of the three categories.

Participants received one of two sets of test stimuli - a *familiar-elements* set or a *novel-elements* set. Each stimulus in a familiar-elements set contained four A elements,  $x$  B elements and  $(8-x)$  C elements where  $x$  could be 0, 1, 2, 3, 4, 5, 6, 7 or 8. Ten examples of each of these nine types of test stimulus were created for each participant receiving a familiar-elements test set. The specific elements used to create each test stimulus were chosen randomly within the constraints provided by the number of A, B and C elements the stimulus was to contain. Ten examples of each of four dummy stimuli were also created, these stimuli being (8 A, 0 B, 4 C), (8 A, 4 B, 0 C), (0 A, 4 B, 8 C) and (0 A, 8 B, 4 C). The purpose of the dummy stimuli was to obscure from the participants that all test stimuli of interest (from the perspective of the experimenter) were constant in terms of the number of elements from category A they contained. Stimuli in the novel-element test sets were constructed in the same manner as familiar-element stimuli, except that the four novel elements (see above) were used instead of four randomly selected A elements.

The position of elements within a stimulus was randomly determined for each stimulus presented, with the constraint that exactly one element occurred at each location in the four-by-three grid. Where stimuli were accompanied by a category label, this was presented as a large sans-serif capital A, B or C in an outline rectangle (4.5 by 3.5 cm) immediately to the right of the stimulus itself.

**Procedure.** Participants were allocated to one of three groups such that an equal number (12) participated in each. The three groups, referred to hereafter as the *two-choice*, *three-choice* and *novel-elements* groups, differed in the number of response alternatives available in the test phase and the stimuli presented during the test phase.

The training phase was the same for all participants. After some general instructions the ninety training stimuli were presented sequentially and in a random order. Each training stimulus was presented for five seconds in the center of the monitor, accompanied by the appropriate category label. Two seconds of plain mid-gray mask in the stimulus and label rectangles preceded the next example. Participants were not required to respond during the training phase. They were simply asked to concentrate on the examples shown as they would later be asked to classify new, unlabelled examples. This training procedure had proved effective in a number of previous experiments (Jones et al., 1998; Wills & McLaren, 1997).

The training phase was followed by a test phase. There were 130 stimuli in the test phase (90 target stimuli and 40 dummy stimuli) which, again, were presented sequentially and in a random order. Test stimuli were not accompanied by a category label. Participants in the two-choice and three-choice conditions received a familiar-elements test set whilst participants in the novel-elements condition received a

novel-elements test set (see *Stimuli*). On the presentation of each test stimulus, participants were asked a question. Participants in the two-choice condition were asked "Is this a B or a C?". Participants in the three-choice and novel-elements conditions were asked "Is this an A, a B or a C?". In all conditions they responded by pressing the appropriate key on the computer keyboard. They then pressed the "Y" key, whereupon the next stimulus was immediately presented. There was no time limit for these decisions, and participants were put under no pressure to respond quickly.

The allocation of the category labels "A", "B", and "C" to the logical categories A, B and C was counter-balanced.

## Results

Figure 2a shows the probability with which participants responded with the category A label to test stimuli (Measure 2) as a function of the number of category B elements they contained (the conclusions of this study are unaffected if one plots against category C elements instead). The functions for the three-choice and novel elements conditions both appeared to show an inverted-U trend. The significant fit of a second-order polynomial to the nine mean data points confirmed this appearance for the three-choice condition,  $F(2, 6) = 5.6$ ,  $p < 0.05$ , but not for the novel-elements condition,  $F(2, 6) = 3.2$ ,  $p > 0.1$ . The quadratic co-efficient for the three-choice condition was significantly different from zero,  $b^2 = -0.006$ ,  $t(7) = 2.4$ ,  $p < 0.05$ .

The data points in Figure 2b are the average of the probability with which participants responded with their category B label to stimuli with  $x$  category B elements and the probability with which they responded with their category C label to test stimuli with  $x$  category C elements. In other words, it shows response probability as a function of the number of *category-appropriate* elements. Averaging these two probabilities is appropriate because, across subjects, there is no factor that determines which of the two categories providing variable numbers of elements to test stimuli should be described as category B and which as category C. A replication of this experiment with non-counterbalanced category labels failed to reveal any significant response bias. Number of category-appropriate elements in Figure 2 reduces from left to right in order to follow the convention that generalization functions (such as those shown in Figure 2b) are plotted as slopes with negative gradients.

For our current purposes it is not the data presented in Figure 2b which are of central interest, but the ratios calculated from the mean points it displays (Measures 1 and 3). These ratios are presented as a function of category-appropriate elements in Figure 2c. Inspection of this figure shows that the 2 choice to 3 choice ratios (Measure 1) appear to exhibit an increasing, accelerating trend whilst the 3 choice novel-elements to 3 choice ratios (Measure 3) exhibit a decreasing, accelerating trend. The significant fit of a second-order polynomial to the nine points of Measure 1,  $F(2, 6) = 803$ ,  $p < .0005$ , with a best-fit line for which all three co-efficients were significantly different from zero,  $b^2 = 0.049$ ,  $t(7) = 14$ ,  $p < .005$ ;  $b = -0.674$ ,  $t(7) = 24$ ,  $p < .0005$ ;  $a = 3.48$ ,

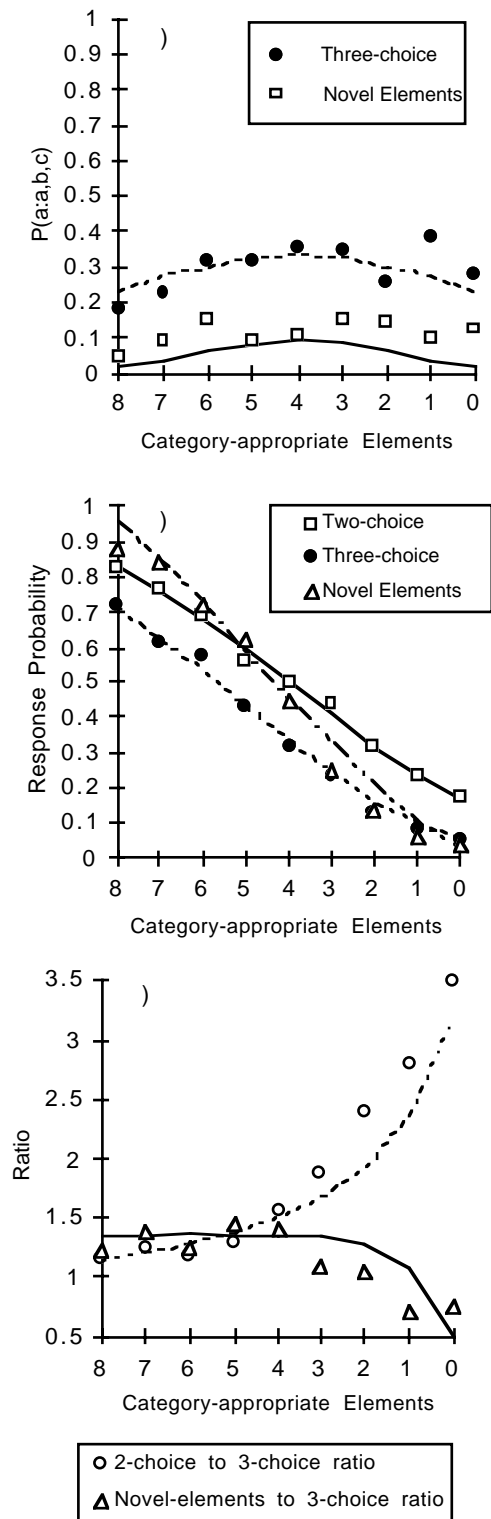


Figure 2: **a)** Probability of producing a category A response. **b)** Mean response probability (see text). **c)** Two ratios calculated from the data in Figure 2b. Plot symbols = empirical data. Lines = predictions of the winner-take-all model presented in the Modeling section.

$t(7) = 2.5, p < .05$ , supports this conclusion. The nine points of Measure 3 were also a significant fit to a second-order polynomial,  $F(2, 6) = 17, p < .005$ , and all three coefficients differed significantly from zero,  $b^2 = -0.021, t(7) = 3.0, p < .05$ ;  $b = 0.244, t(7) = 4.3, p < .005$ ;  $a = 0.632, t(7) = 3.8, p < 0.01$ .

## Discussion

Our results pose two central problems for the Ratio Rule as it is currently employed in many formal models of categorization.

First, the Ratio Rule predicts that the two ratios represented in Measures 1 and 3 should show the same direction of change over any interval of category-appropriate elements. However, the best-fitting quadratics for the corresponding functions show opposite directions of change (Figure 2c). Second, the Ratio Rule predicts that the probability of choosing category A in the three-choice condition (Measure 2) should show the same direction of change over any interval of category-appropriate elements as the other two measures. However, the best-fitting quadratics for measures 1 and 2 are of opposite shape (compare Figure 2a with Figure 2c). One might argue that these findings are of relatively little consequence because the discrepancies are in derived measures with no straightforward psychological interpretation, rather than in the response probabilities themselves. Such a position is disingenuous. The predictions under test arise naturally and unavoidably from a central (some would say defining) feature of the Ratio Rule - the fact that the ratio in which two alternatives are chosen is unaffected by the addition of a third alternative. These data provide evidence against that central tenet and hence bring the formulation into question.

If any one step in a chain of inferences is incorrect then the conclusions drawn from that process must be brought into question. Consequently, theoretical conclusions about the nature of categorization must be re-examined if our conclusion is found to be generally valid. Conversely, if the assumptions we have made in coming to our conclusions can be shown to be invalid then the Ratio Rule is not necessarily incorrect. Below we briefly consider some possible criticisms of our conclusion.

First, one could argue that we have disproved the Ratio Rule for means across participants, but this does not disprove the formulation for individual participants. This is a valid point, but as most formal theories of categorization have been applied to group means our conclusion still stands for these theories. Second, it is true that our stimuli are rather more complex than those typically used in category learning experiments. It may be the case that our results do not generalize to simpler stimuli, or that our stimuli are unusual in some other way. This seems to be an empirical matter, and one which is worthy of investigation. A third, substantial criticism is that we have assumed that category magnitude terms are, for our stimuli, univariate functions of the number of category-appropriate elements the stimulus contains (i.e. the magnitude term is determinable solely from this property). There are at least two distinct ways in which this assumption could be incorrect.

First, for specific models of categorization it may be possible to show that the category magnitude term for category A is not invariant under changes in the magnitude terms for categories B and C. For example, one might be able to demonstrate for the GCM model (Nosofsky, 1986) that the test stimuli were not at a fixed distance (in psychological similarity space) from category A examples. The difficulty here is that the procedure which Nosofsky uses to derive the psychological similarity space assumes that the Ratio Rule is correct. Some way around this circularity would have to be devised.

Second, one could quite reasonably argue that category magnitude terms are importantly affected by what response alternatives are available (as a number of theorists outside of the categorization literature have argued e.g. Restle, 1961; Tversky, 1972). If this were the case in our experiment then the derivation of Measure 1 would be invalid because it is directly based on this assumption.

Therefore one response to our results might be to retain the Ratio Rule but introduce a mechanism by which category magnitude terms can be affected by the alternatives available for decision. However, for most formal models of categorization this would require considerable revision of the basic principles upon which they were based. We wondered whether there was a direct replacement for the Ratio Rule that could accommodate our results without having to modify the rest of the theory.

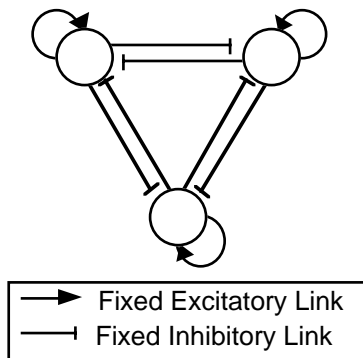


Figure 3: The winner-take-all model.

### Modeling

Previously we have proposed that response probabilities in categorization might be modeled by a simple winner-take-all connectionist system employing category magnitude terms as input activations (Wills & McLaren, 1997). Such a system is illustrated in Figure 3. In addition to the magnitude-term inputs, each unit has a fixed excitatory connection to itself and fixed inhibitory connections to the other units. These connections can cause the units to “compete” with one another until only one has a non-zero activation. In our system a decision is deemed to have been made when the highest activation exceeds its nearest competitor by some threshold value,  $S$ . This general architecture has been proposed previously by Grossberg (1976) amongst others, and has been employed in the modeling of a number of other

psychological phenomena (e.g. Houghton, 1990; Usher & McClelland, 1995).

For the purposes of this simulation we assume that category magnitude terms are defined by the function  $v = 0.047c + 0.012$ , where  $c$  is the number of category-appropriate elements the stimulus contains. The exact form of this equation is not critical. It was chosen because it describes the behavior of a simple localist delta-rule network with a learning rate of 0.0025. This learning rate was previously found to be successful in modeling the rate of learning in similar experiments (Wills & McLaren, 1997). The important thing to note is that we are preserving the assumption that category magnitude terms are independent of the response alternatives available.

The magnitude term input activations ( $r$ ) are assumed to be noisy and, for simplicity, this noise is assumed to be rectangular, have a mean of zero, and a range from  $-N$  to  $+N$ . Magnitude input activations are also constrained to lie between 0 and 1. The specific shape of the noise distribution is not critical and similar mean behavior could be produced with a Gaussian distribution. The output activations of the units are governed by the equations

$$o = \frac{o + En}{1 + En + D}, \text{ if } n > 0 \text{ and } o = \frac{o + En}{1 - En + D} \text{ otherwise,}$$

where  $n$  is the total input the unit, and  $E$  and  $D$  are excitation-rate and decay-rate constants respectively. These are standard activation equations with properties similar to those used by, for example, McClelland & Rumelhart (1985). Output activations in our model are constrained to be non-negative. Total input ( $n$ ) for a given unit is the sum of  $r$  and  $o$  for that unit, minus the sum of the outputs ( $o$ ) of the other units. For the current simulation  $E = 0.2$ ,  $D = 0.1$  and  $N = 1.1$ . The threshold parameter  $S$  was set to 0.18 for the two-choice condition, 0.65 for the three-choice condition and 0.72 for the novel-elements condition.

In the two-choice condition of our experiment, participants were not allowed to make category A responses. In our WTA model this was simulated by fixing the output activation of the category A unit at zero.

The results of our simulation are shown as lines in Figure 2. Note that the model respects all the major trends in the experiment and is numerically close to the observed data. A detailed discussion of the principles underlying the success of this model is not possible here, but it is important to note that the exact details of the implementation are not critical. Indeed, not even the expression in connectionist terms is essential. The model simply provides a mechanism by which a decision similar in principle to Thurstonian choice (Thurstone, 1927) can be made. We have demonstrated in other analyses that simply choosing the noisy alternative which is instantaneously the biggest does reasonably well in predicting the trends in Measures 1 and 2 (whether one employs Gaussian or rectangular noise).

However, only the connectionist system correctly predicts the trend in Measure 3. This is because it employs different decision thresholds in the three-choice and novel-elements conditions, which allows it to predict that Measure 3 falls below unity without having to make the counter-intuitive

assumption that the category A magnitude term for a stimulus containing no category A elements is greater than for a stimulus containing four category A elements (the only way a simple Thurstonian choice process could predict ratios smaller than one).

## Conclusion

The Ratio Rule as generally applied in formal models of categorization was shown to be incorrect for the experiment presented. Whilst further investigation is necessary, we suggest that our results may indicate a need to replace the Ratio Rule as currently employed with an alternative system (perhaps still based around the Ratio Rule). One possibility would be to substantially revise existing models so that they provided a mechanism by which category magnitude terms could be affected by the alternatives available for decision. We have shown that our results do not require that this modification be made. Rather, one simply needs to directly substitute the Ratio Rule with a decision mechanism based on the principles of Thurstonian choice. The noise employed in this mechanism may have one of a number of distributions. As a caveat, one distribution it is unlikely to have is a double exponential distribution because this would make it indistinguishable from the Ratio Rule (Yellott, 1977).

## Acknowledgments

The authors would like to thank Fergal Jones, Koen Lamberts, Donald Laming and Thomas Palmeri for their helpful comments. Thanks are also due to Stian Reimers and Neil Stewart who helped out with similar studies for which there was insufficient space in this article. This research was supported by a grant from the ESRC to I.P.L. McLaren.

## References

- Bradley, R. A. (1954). Incomplete block rank analysis: On the appropriateness of the model for a method of paired comparison. *Biometrics*, *10*, 375-390.
- Burke, C. J. & Zinnes, J. L. (1965). A paired comparison of pair comparisons. *Journal of Mathematical Psychology*, *2*, 53-76.
- Clarke, F. R. (1957). Constant-ratio rule for confusion matrices in speech communication. *Journal of the Acoustical Society of America*, *29*(6), 715-720.
- Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal Of Experimental Psychology: General*, *117*(3), 227-247.
- Grossberg, S. (1976). Adaptive pattern classification and universal recoding: Part I. Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121-134.
- Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 418-439.
- Hopkins, J. W. (1954). Incomplete block rank analysis: Some taste test results. *Biometrics*, *10*, 391-399.
- Houghton, G. (1990). The problem of serial order: A neural network model of sequence learning and recall. In R. Dale, C. Mellish, & M. Zock (Eds.), *Current Research in Natural Language Generation*. London: Academic Press.
- Jones, F. W., Wills, A. J., & McLaren, I. P. L. (1998). Perceptual categorization: Connectionist modelling and decision rules. *The Quarterly Journal of Experimental Psychology*, *51B*(3), 33-58.
- Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *22*(1), 3-26.
- Laming, D. (1977). Luce's choice axiom compared with choice-reaction data. *British Journal of Mathematical and Statistical Psychology*, *30*, 141-153.
- Luce, R. D. (1959). *Individual Choice Behavior*. New York: John Wiley & Sons.
- McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal Of Experimental Psychology: General*, *114*(2), 159-188.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207-238.
- Nosofsky, R. M. (1986). Attention, similarity and the identification-categorisation relationship. *Journal Of Experimental Psychology: General*, *115*(1), 39-57.
- Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353-363.
- Restle, F. (1961). *Psychology of judgement and choice*. New York: Wiley.
- Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 273-286.
- Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*(4), 281-299.
- Usher, M., & McClelland, J. L. (1995). *On the time course of perceptual choice: A model based on principles of neural computation*. (Technical Report PDP.CNS.95.5): Carnegie Mellon University.
- Wills, A. J., & McLaren, I. P. L. (1997). Generalization in human category learning: A connectionist explanation of differences in gradient after discriminative and non-discriminative training. *The Quarterly Journal of Experimental Psychology*, *50A*(3), 607-630.
- Yellott, J. I., Jr. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgement, and the double exponential distribution. *Journal of Mathematical Psychology*, *15*, 109-144.

From 1st September 2000, A.J.Wills should be contacted at: University of Exeter, School of Psychology, Washington Singer Laboratories, Perry Rd., Exeter. EX4 4QG. United Kingdom. <http://www.ex.ac.uk/Psychology/>

# Strategies and Tactics in Sentential Reasoning

**Yingrui Yang** ([yingruiy@princeton.edu](mailto:yingruiy@princeton.edu))  
Princeton University, Department of Psychology;  
Green Hall, Princeton, NJ 08544 USA

**Jean-Baptiste van der Henst** ([jvanderhenst@caramail.com](mailto:jvanderhenst@caramail.com))  
Princeton University, Department of Psychology;  
Green Hall, Princeton, NJ 08544 USA

**P. N. Johnson-Laird** ([phil@princeton.edu](mailto:phil@princeton.edu))  
Princeton University, Department of Psychology;  
Green Hall, Princeton, NJ 08544 USA

## Abstract

We propose a theory of the spontaneous reasoning strategies that individuals develop. These strategies depend on component tactics based on mental models. Reasoners vary their use of tactics in ways that are not deterministic. This variation leads different individuals to assemble different strategies, which include constructing incremental diagrams corresponding to mental models, and pursuing the consequences of a single model step by step. The number of models required by the premises predisposes reasoners towards certain strategies, e.g., multiple models tend to elicit incremental diagrams. Similarly, the connectives in premises also bias reasoners towards certain strategies, e.g., conditional premises tend to elicit reasoning step by step from a single model.

## Introduction

Psychologists have tended to neglect the strategies that individuals develop spontaneously to make complex inferences (cf. Schaeken, De Vooght, Vandierendonck, and d'Ydewalle, 2000). By a *strategy*, we mean a systematic sequence of elementary mental steps, i.e., tactics, that an individual follows in making an inference. Pioneering studies of strategies examined relational reasoning in which the task is, say, to infer who is tallest in a series of individuals. The results suggested that reasoners develop a variety of strategies (e.g. Wood, 1969; Quinton and Fellows, 1975). However, there has been a dearth of studies of strategies in sentential reasoning, which hinges on negation and connectives such as "if", "or", and "and". Some theorists have argued that sentential reasoning relies on a single deterministic strategy based on formal rules of inference (cf. Rips, 1994; Braine and O'Brien, 1998). We suspect that theorists have postulated a single deterministic strategy because their experiments have used too simple premises for strategies to differ, and because they have failed to gather evidence about reasoner's strategies. Indeed, we and our colleagues have proposed that naïve reasoners generally develop a variety of strategies (e.g. John-

son-Laird and Byrne, 1990; Byrne and Handley, 1997; Bucciarelli and Johnson-Laird, 1999).

## Experiment 1: A taxonomy of strategies

How can experimenters best observe the strategies that reasoners use in sentential reasoning? In our view, studies of strategies should examine inferential problems that are sufficiently time-consuming to force the participants to think, but not so difficult that they make many errors. We therefore used sentential problems based on three premises, but each set of premises was compatible with only two alternative possibilities. The task was to evaluate a given conclusion and to think aloud (cf. Ericsson and Simon, 1984). Here is a typical example of a problem:

Either there is a blue marble in the box or else there is a brown marble in the box, but not both. Either there is a brown marble in the box or else there is white marble in the box, but not both. There is a white marble in the box if and only if there is a red marble in the box. Does it follow that: If there is a blue marble in the box then there is a red marble in the box?

Henceforth, we use the abbreviations: "iff" for biconditionals of the form "if and only if", "ore" for exclusive disjunctions of the form "either \_ or else \_, but not both", and "or" for inclusive disjunctions of the form "\_ or \_, or both".

Our theory of strategies is based on mental models (Johnson-Laird and Byrne, 1991), and each mental model represents a possibility. All the problems in the experiment called for two mental models. The premises of the example above yield the following two models of the possible contents of the box, shown on separate lines:

blue	white	red
brown		

As the models show, the putative conclusion follows from the premises.

*Method.* Eight Princeton undergraduates, who had no training in logic, carried out twelve inferences, which each had a conclusion to be evaluated. The problems were based on three or four premises. Half of them had valid conclusions and half of them had invalid conclusions. The premises were mainly biconditionals and exclusive disjunctions, and the conclusions were conditionals except for two problems, which had exclusive disjunctions as conclusions. As in the example above, the contents of the problems concerned different colored marbles. The problems were presented in a different random order to each participant.

The participants were allowed to use pencil and paper. They were told to think aloud as they tackled each inference, and we video-recorded what they said, wrote, and drew. The camera was above them and focused on the paper on which they wrote, and they rapidly adapted to its presence.

*Results.* None of the participants made any errors in evaluating the given conclusions, though they were not always right for the right reasons. We transcribed the tapes verbatim apart from repetitions of words, filled pauses, and hesitations. These protocols also included a record of the step by step drawings of diagrams. We were able to make sense of almost all of what the participants said, drew, and wrote. Most participants used two or more distinct strategies, but two of them stuck to the same strategy throughout the experiment. What the protocols did not reveal were either the processes in developing a strategy, or the mechanisms underlying the tactical steps. We were able, however, to categorize the protocols from every participant for every problem into one of the strategies in the taxonomy in Table 1 below.

The taxonomy distinguishes five main strategies. It is based on all our experiments, but it may be necessary to add further strategies: no-one can ever know when the classification is complete. The five strategies were:

1. The incremental models strategy. Reasoners draw a diagram that integrates all the information from the premises. The diagram corresponds to a set of models (see the example above). Some participants drew the models in vertical columns down the page. Others arranged them horizontally. One participant merely drew circles around the propositions in the premises themselves to pick out one of the two models. Participants work through the premises in an order that allows them to increment their diagrams.

2. The step strategy. Reasoners pursue the step by step consequences of either a categorical proposition or a supposition. They accordingly infer a sequence of what logicians refer to as “literals”, where a *literal* is a proposition that does not contain any sentential connectives: it may be an atomic proposition, *A*, or its negation, *not A*. Consider the following problem, stated in an abbreviated form:

- Pink iff black.
- Black ore gray.
- Gray iff blue.

Does it follow that if not pink then blue?

One participant's complete verbatim protocol, illustrating the strategy, is:

Assuming we have no pink:  
 There is no pink. [He crosses out “pink” in premise.]  
 So there is no black. [Crosses out “black” in premises.]  
 There is gray. [Circles “gray” in premise.]  
 There is blue. Yes. [The conclusion follows.]

3. The compound strategy. Reasoners take two compound assertions, i.e., assertions containing a sentential connective, and draw a compound conclusion from them, e.g.:

Pink ore brown. [Reads premise]  
 Pink and white. [Points to diagram of premise:  
 pink → white]

If brown then not white.[A compound inference.  
 Writes: brown, ~~white~~]

White ore brown. [The required conclusion]

The strategy consists in a sequence of such compound inferences that yield an ultimate conclusion.

4. The chain strategy. Reasoners construct a chain of *conditionals* leading from one constituent of a compound conclusion to its other constituent. They make an immediate inference from any premise that is not a conditional to convert it into an appropriate conditional (see Richardson and Ormerod, 1997). Here is an example of a protocol:

[Crosses out terms in diagrams:

If not pink then not green. ~~pink~~ = ~~green~~  
 If not green then red. ~~green~~ or red  
 If red then white. red = white

Yes. [I.e. If not pink then white]

The valid use of the strategy to prove a biconditional or exclusive disjunction calls for two chains, but reasoners usually rely on just a single chain.

5. The concatenation strategy. Reasoners sometimes concatenate the premises to form a complex intermediate conclusion. They then draw an immediate inference from it to the required conclusion. For example, one participant concatenated the premises:

A and B.  
 B iff C.  
 C iff D.

to yield : A and (B iff C iff D). She then made an immediate inference to the required conclusion: A and D.

For the twelve problems in Experiment 1, we calculated the total number of times each strategy occurred in the protocols, and then expressed these numbers as percentages of the total number of occurrences of strategies. The results were as follows:

Incremental models strategy:	34% of overall use.
Supposition and step strategy:	21% of overall use.
Compound strategy:	19% of overall use.
Chain strategy:	25% of overall use.
Concatenation strategy:	0% of overall use.

The most salient feature of the protocols was that different participants used different strategies.



### A theory of reasoning strategies

A *deterministic* process is one in which each step depends solely on the current state of the process and whatever input it may have (Hopcroft and Ullman, 1979). Following Harman (1973), however, we assume that reasoning is not a deterministic process that unwinds like clockwork. Our first assumption is accordingly:

1. The principle of *nondeterminism*: thinking in general and sentential reasoning in particular is governed by constraints, but there is seldom just a single path it must follow. It varies in a way that can be captured only in a nondeterministic account.

Experiment 1 corroborated the principle of nondeterminism, and it did so at two levels. At a high level, the participants developed diverse strategies. At a low level, there was considerable variation within strategies.

Our second assumption is:

2. The principle of *strategic assembly*: naïve reasoners assemble reasoning strategies bottom up as they explore problems using their existing inferential tactics. Once they have developed a strategy, it can control their reasoning in a top-down way.

A corollary of the principle is that individuals are most unlikely to develop a reasoning strategy working “top down” from a high-level specification. Granted the principle, it also follows that the space of possible strategies is defined by the different ways in which inferential tactics can be sequenced in order to make inferences. Hence, an exhaustive enumeration of tactics provides the recursive basis for all possible strategies.

If the mechanism underlying reasoning depends on mental models, then each inferential tactic must be based on models. We therefore postulate a third assumption:

3. The principle of *model-based tactics*: inferential tactics are based on mental models.

The mechanisms for constructing models are, in turn, constrained by the nature of the human mind, which reflects innate constraints and individual experiences.

Our first test of the three principles was to show that mental models can underlie all the strategies and tactics in our taxonomy. The *incremental models* strategy is isomorphic to the cumulative construction of a single set of models based on the premises. The *step* strategy is based on a categorical premise or a supposition. Although the strategy is similar to the one strategy that Rips (1994) proposes, the model theory allows a greater freedom in the use of suppositions – a freedom that corresponds to their use by naïve reasoners. The main inferential step is to use a literal to update a set of models based on a premise in order to draw another literal as a conclusion. A premise, such as: Black ore gray, yields two models:

black  
gray

and the supposition, Not black, eliminates the first model and yields the literal conclusion: gray. The *compound*

Table 1: The model-based tactics underlying each of the five strategies: + indicates the use of a tactic, and (+) indicates its optional use.

Tactics	The five strategies				
	Increment models	Step	Compound	Chain	Concatenation
Make a supposition	(+)	+			
Concatenate premises		(+)	(+)		+
Construct models	+	+	+	+	+
Update models	+	+	+		
Immediate inference from models	(+)		(+)	+	+
Formulate intermediate conclusion from models		+	+		
Evaluate or formulate a conclusion from models	+	+	+	+	+

strategy relies on a series of compound inferences based on models. The *chain* strategy depends on the construction of a chain of conditionals. The chain has one explicit mental model and one implicit mental model. To prove a conditional of the form:

If A then D.

individuals can construct a chain leading from D to A, e.g.:

If D then not-C.  
If not-C then B.  
If B then A.

Such a strategy is invalid. So, why do reasoners construct this chain? The answer is that the conclusion holds in the mental models of the chain:

d     ¬c     b     a

Hence, mental models underlie the strategy. The *concatenation* strategy appears at first sight to rely on purely syntactic operations, and therefore to violate the principle of model-based tactics. In fact, the strategy depends critically on mental models. Given a pair of premises of the form:

A iff B.  
B ore C.

there are two ways in which to concatenate a conclusion:

1. (A iff B) ore C.
2. A iff (B ore C).

Which of these two conclusions follows from the premises? In fact, neither conclusion is valid. Yet, eight out of the eight participants in Experiment 3 who concatenated conclusions from the relevant premises generated conclusion 2. It is the one conclusion that has the same mental models as the premises. Ten participants in Experiment 2 used the tactic of concatenating a conclusion on one or more occasions. On 82% of occasions, the resulting conclusions were compatible with the mental models of the premises, and nine of the ten participants concatenated more conclu-

sions of this sort than not (Sign test,  $p < .02$ ). Concatenation is not blindly syntactic. It tends to be accepted only if it yields the same mental models as the premises. Table 1 presents the taxonomy of strategies and their underlying model-based tactics.

**Experiment 2: Development of strategies**

The theory predicts that the nature of the inferential problems given to reasoners should influence their development of strategies. According to the principle of strategic assembly, the characteristics of particular problems should trigger certain strategies “bottom up”. One instance of this prediction concerns the effects of number of models. Problems that include a categorical premise or a conjunction of them yield a single model. Hence, individuals can use a categorical premise as the starting point of their reasoning, and the step strategy is the easiest way to proceed because it places a minimal load on working memory. With multiple-model problems, the optimal way to keep track of the possibilities is to use the incremental models strategy. Multiple models, however, should also yield a greater number of errors. The aim of the present experiment was to test these predictions.

*Method.* Twenty Princeton undergraduates acted as their own controls and evaluated given conclusions to 36 problems presented in three blocks: twelve one-model inferences, twelve two-model inferences, and twelve three-model inferences. Typical problems were of the form:

<i>One-model</i>	<i>Two-model</i>	<i>Three-model</i>
A and B.	A iff B.	A iff B.
B ore C.	B ore C.	B iff C.
C iff D.	C iff D.	C or D.
A and not D?	A iff D?	A or D?

The participants evaluated the conclusions, and we used the same think-aloud and video-recording procedure as before.

Table 2: The percentages of the different strategies for the three sorts of problems in Experiment 2. The balances of the percentages (5%) were uncategorizable strategies.

	The strategies		
	Incremental models	Step	The other strategies
One-model premises	21	69	3
Two-model premises	26	56	15
Three-model premises	49	45	2

*Results.* As the model theory predicts, errors increased with the number of models: there were 8% of errors with one-model problems, 15% of errors with two-model problems, and 20% of errors with three model problems (Page’s  $L = 251.5$ ,  $p < 0.05$ , one-tailed). Table 2 presents the percentages of the different strategies for the different sorts of problem. The participants were sensitive to the properties

of the particular problems. As the theory predicts, they relied increasingly on the incremental models strategy as the problems required a greater number of models (Page’s  $L = 254.5$ ,  $p < .05$ , one-tailed). They tended to use the step strategy with one-model problems, but the use of the strategy declined with an increasing number of models. The results accordingly corroborated the principle of strategic assembly: reasoners develop strategies “bottom-up” depending on the sort of problem that they encounter.

**Experiment 3: Formulating conclusions**

This experiment was similar to Experiment 2, except that the participants had to draw their own conclusions.

*Method.* Twenty four Princeton undergraduates acted as their own controls and carried out four one-model inferences, four two-model inferences, and four three-model inferences, in counterbalanced orders. For each problem, they wrote down their answer to the question, “What, if anything, follows?” and we used the same procedure as before.

*Results.* The participants developed diverse strategies, and the realization of any particular strategy varied from trial to trial even for the same participant. As the model theory predicts, the percentages of invalid conclusions, modal conclusions about possibilities, and conclusions that failed to take into account all the premises, each increased significantly with the number of models. Table 3 presents the percentages of the different strategies in the experiment. As predicted, the use of the incremental models strategy increased with the number of mental models required by the premises. With one-model problems, the participants were likely to use the step strategy, but there was an increase in the use of the incremental models strategy with multiple-model inferences. This trend was reliable (Kendall’s coefficient of concordance,  $W = 0.228$ ,  $X^2 = 10.94$ ,  $p < .01$ , two-tailed).

Strategies should influence the form of the conclusions that reasoners draw. With incremental models, it is difficult to see what is common to a number of alternative possibilities, and so reasoners should tend to describe each possibility separately and to combine these descriptions in a disjunction. The other strategies, however, are unlikely to yield conclusions of this sort. These strategies focus on a single possibility, such as a supposition. We examined this prediction by dividing the participants in Experiment 3 into two post hoc groups. In the *model* group (9 participants), more than half of the participants’ identifiable strategies yielding conclusions were the incremental models strategy. In the *non-model* group (15 participants), more than half of the participants’ identifiable strategies yielding conclusions were some other sort. For the model group, 63% of the problems solved with the model strategy had a conclusion that was a disjunction of possibilities, but for the non-model group only 11% of the problems solved with a non-model strategy had such a conclusion (Mann-Whitney test,  $z = 2.87$ ,  $p < .005$  one-tailed). Different strategies do yield different sorts of conclusion.

Table 3: The percentages of the different strategies for the three sorts of problems in Experiment 3. The balances of the percentages (11% overall) are uncategorizable strategies.

	The strategies			
	Incremental models	Step	Compound	Chain
One-model premises	14	80	5	3
Two-model premises	33	22	20	9
Three-model premises	36	25	14	7
Overall	28	41	13	7

### Experiment 4: Strategies and premises

The principle of strategic assembly implies that the form of the premises should influence the development of strategies. A way to elicit the incremental models strategy should be use to disjunctive premises, which are naturally represented as sets of possibilities. A way to elicit the step and chain strategies is to use conditional premises, which have only a single explicit model required by these strategies. These effects should occur even when the premises are otherwise logically equivalent. Once individuals have developed a strategy, it should have a “top down” residual effect on their subsequent performance. It should be used for problems that would not normally trigger its use. The experiment tested these predictions.

*Method.* Twenty Princeton undergraduates acted as their own controls and drew their own conclusions to two sets of problems: four disjunctive problems and four logically equivalent conditional problems. Half the participants received the four disjunctive problems in a random order followed by the four conditional problems in a random order; and half the participants received the two blocks of problems in the opposite order.

*Results.* Table 4 presents the percentages of the different strategies for the two sorts of problems, and it gives the data separately for the two blocks of trials. As the theory predicts, the participants were more likely to use the incremental models strategy (56%) for the disjunctive problems than for the conditional problems (23%; Wilcoxon test  $T = 66, n = 11, p < .0005$ ). The table shows that the participants who first carried out the conditional problems rarely developed the incremental models strategy (10% of these problems), but their use of the strategy increased reliably for the disjunctive problems (55% of problems, Sign test,  $p < .02$ , two tailed). In contrast, those who first carried out the disjunctive problems often developed the incremental models strategy, and did not reliably reduce its use with the conditional problems. This difference between the two groups was reliable (Mann-Whitney  $U = 21, p < .05$ , two tailed). An obvious explanation for the differential transfer is that the incremental models strategy is simpler to use with any sort of sentential connective, whereas the step and chain strategies call for additional immediate inferences to convert

Table 4: The percentages of the different strategies for (a) the disjunctive problems and (b) the conditional problems in Experiment 4. The balances of the percentages are trials with erroneous responses or uncategorizable strategies.

(a) Disjunctive problems	The strategies	
	Incremental models	Step, Compound, and Chain
Presented first	58	35
Presented second	55	35
Overall	56	35

(b) Conditional problems	The strategies	
	Incremental models	Step, Compound, and Chain
Presented first	10	90
Presented second	35	60
Overall	23	75

disjunctive premises into conditionals.

The experiment corroborated the principle of strategic assembly. The nature of the sentential connectives biases reasoners to adopt particular strategies. The incremental models strategy, though it places a greater load on working memory, is more flexible than the other strategies, which are more finely tuned to conditional premises.

### General Discussion

Unlike some cognitive domains, such as arithmetic (Lemaire and Siegler, 1995), accounts of sentential reasoning have neglected strategies (for reviews, see Evans, Newstead, and Byrne, 1993; Garnham and Oakhill, 1994). Studies have failed to use appropriate methods to discover strategies; and in consequence theorists have often assumed that reasoners rely on a single deterministic strategy. We have tried to remedy the neglect and to advance a new theory of strategies in reasoning. Naïve reasoners use at least five distinct strategies. As the theory predicts, each strategy is built from tactical steps that rely on the manipulation of models (see Table 1). The *incremental models* strategy keeps track of all the mental models compatible with the premises. The *step* strategy pursues the step by step consequences of one model – either one derived from a categorical assertion in a premise or one created by a supposition. The *compound* strategy combines the models of compound premises to infer what is necessary or possible. The *chain* strategy pursues a model in a sequence of conditionals, which may be inferred from the premises, leading from one constituent of a conclusion to another. The *concatenation* strategy forms a conclusion by concatenating the premises, but normally only if the resulting conclusion has the same mental models as the premises. Because it relies on mental models, it gives rise spontaneously to illusory inferences (cf. Johnson-Laird and

Savary, 1999; Goldvarg and Johnson-Laird, 2000; Johnson-Laird et al., 2000; Yang and Johnson-Laird, 2000).

The model theory explains how people develop reasoning strategies. They are equipped with a set of inferential tactics. As they reason, the variation in their performance, leads them to assemble these tactics in novel ways so that they yield a reasoning strategy. As a result, they develop different strategies. All their strategies, however, depend on tactics based on mental models. The properties of inferential problems can accordingly influence the development of particular strategies. The problems in Experiments 2 and 3 called for one, two, or three models. As the theory predicts, the participants tended to use the conjunction in one-model problems as the starting point for the step strategy, which places a minimal load on working memory. As the number of models increased, they were more likely to use the incremental models strategy, which keeps track of the different possibilities compatible with the premises. Experiment 4 also bore out the theory's account of strategic assembly. Disjunctive premises, as predicted, tended to elicit the incremental models strategy, whereas conditional premises tended to elicit other strategies. The participants increased their use of incremental models on switching to disjunctive premises, but they did not decrease its use on switching to conditional problems. Although incremental models load working memory, the strategy is more flexible than those that are optimal for conditional premises.

What would have refuted our theory? At the lowest level, that of inferential mechanisms, the theory would have been refuted if there had not been an increase of difficulty with the number of models required by the problems. This phenomenon has been observed in previous studies (see Johnson-Laird and Byrne, 1991), but not before in inferences based on three sentential connectives. At the tactical level, the theory would have been refuted if reasoners used tactics incompatible with manipulations of models. Suppose, for example, that the concatenation had not been sensitive to the mental models of the premises, then a tactic would have been controlled purely by syntactic considerations, and it would have been contrary to the theory. At the strategic level, the theory would have been refuted if reasoners had uniformly developed a single deterministic strategy (cf. Rips, 1994). The moral of our results is clear. They support the three principles of the model theory of reasoning strategies.

### Acknowledgments

The authors thank Fabien Savary for carrying out Experiment 1. They are grateful for the helpful comments of Bruno Bara, Patricia Barres, Victoria Bell, Monica Bucciarelli, Ruth Byrne, Gino De Vooght, Zachary Estes, Vittorio Girotto, Yevgeniya Goldvarg, Paolo Legrenzi, Maria Sonino Legrenzi, Patrick Lemaire, Juan Madrugá, Hansjoerg Neth, Mary Newsome Walter Schaeken, Vladimir Sloutsky, Patrizia Tabossi, and André Vandierendonck.

### References

- Braine, M.D.S., and O'Brien, D.P., Eds. (1998) *Mental Logic*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Bucciarelli, M., and Johnson-Laird, P.N. (1999) Strategies in syllogistic reasoning. *Cognitive Science*, 23, 247-303.
- Byrne, R.M.J., and Handley, S.J. (1997) Reasoning strategies for suppositional deductions. *Cognition*, 62, 1-49.
- Ericsson, K.A., and Simon, H.A. (1984) *Protocol Analysis: Verbal Reports as Data*. Cambridge: MIT Press.
- Evans, J.St.B.T., Newstead, S.E., and Byrne, R.M.J. (1993) *Human Reasoning: The Psychology of Deduction*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Garnham, A., and Oakhill, J.V. (1994) *Thinking and Reasoning*. Oxford: Basil Blackwell.
- Goldvarg, Y., and Johnson-Laird, P.N. (2000) Illusions in modal reasoning. *Memory & Cognition*, 28, 282-294.
- Harman, G. (1973) *Thought*. Princeton, NJ: Princeton University Press.
- Hopcroft, J.E., and Ullman, J.D. (1979) *Formal Languages and Their Relation to Automata*. Reading, MA: Addison-Wesley.
- Johnson-Laird, P.N. and Byrne, R.M.J. (1990) Meta-logical problems: Knights, knaves, and Rips. *Cognition*, 36, 69-81.
- Johnson-Laird, P.N., and Byrne, R.M.J. (1991) *Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Johnson-Laird, P.N., and Savary, F. (1999) Illusory inferences: A novel class of erroneous deductions. *Cognition*, 71, 191-229.
- Johnson-Laird, P.N., Legrenzi, P., Girotto, V., and Legrenzi, M. (2000) Illusions in reasoning about consistency. *Science*, 288, 531-532.
- Lemaire, P., and Siegler, R.S. (1995) Four aspects of strategic change: Contributions to children's learning of multiplication. *Journal of Experimental Psychology, General*, 124, 83-97.
- Quinton, G. and Fellows, B.J. (1975). 'Perceptual' strategies in the solving of three-term series problems. *British Journal of Psychology*, 66, 69-78.
- Richardson, J., and Ormerod, T.C. (1997) Rephrasing between disjunctives and conditionals: Mental models and the effects of thematic content. *Quarterly Journal of Experimental Psychology*, 50A, 358-385.
- Rips, L.J. (1994) *The Psychology of Proof*. Cambridge, MA: MIT Press.
- Schaeken, W., De Vooght, G., Vandierendonck, A., and d'Ydewalle, G. (2000) *Deductive Reasoning and Strategies*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Wood, D.J., (1969). Approach to the study of human reasoning. *Nature*, 223, 102-103
- Yang, Y., and Johnson-Laird, P.N. (2000) Illusory inferences with quantified assertions: How to make the impossible seem possible, and vice versa. *Memory & Cognition*, in press.

# Problem Representation in Experts and Novices: Part 1. Differences in the Content of Representation

**Aaron S. Yarlas (yarlas.1@osu.edu)**

Center for Cognitive Science and School of Teaching & Learning, The Ohio State University  
21 Page Hall, 1810 College Road, Columbus, OH 43210 USA

**Vladimir M. Sloutsky (sloutsky.1@osu.edu)**

School of Teaching & Learning and Center for Cognitive Science, The Ohio State University  
21 Page Hall, 1810 College Road, Columbus, OH 43210 USA

## Abstract

Two experiments examined the content of novice and expert representations for both surface and deep structural elements of arithmetic equations. Experiment 1, which used a forced-choice categorization task in which surface features of equations (e.g., digits) competed with deep structural principles of mathematics (associativity and commutativity), found that experts were more likely to focus on principles in their judgments than were novices, who focused more often on surface elements. Experiment 2, using a similar task, introduced trials in which only principled elements varied. Novices were able to focus on principled elements in this case, but failed to transfer these representations when surface features were re-introduced. These findings indicate that novices had knowledge of the principles, but that they did not attend to them when competing surface features were present.

## Introduction

It has been well established that in various knowledge domains (e.g., physics, mathematics, or chess) experts approach problems in a manner different from that of novices (Chase & Simon, 1973; Chi, Feltovich, & Glaser, 1981; Larkin, 1983; Simon & Simon, 1978; Reed, Ackinclose, & Voss, 1990). In particular, while experts are more likely to focus on hidden relational properties of a problem, novices are more likely to focus on less important surface features of a problem. However, while there is some understanding of the content of mental representation (i.e., of which aspects of information are likely to be represented and which are likely to be left out), the process of construing the representation remains largely unknown. Do people attend to and encode those aspects that are left out, but then discard them, or do they fail to attend to and encode these "irrelevant" aspects?

The current paper (Part 1) focuses both on establishing differences in content of representation for experts and novices within a simple domain (arithmetic) and testing a number of viable explanations that could account for these differences. A subsequent paper (Part 2) focuses on examining differences in the process of construing representations for experts and novices.

There is a large body of literature indicating that in problem solving, reasoning, learning and transfer, and problem

categorization, novices tend to focus on surface features rather than on deep relational properties. These effects have been demonstrated in a variety of knowledge domains, including chess (Chase & Simon, 1973), mathematics (Blessing & Ross, 1996; Schoenfeld & Herrmann, 1982; Bassok, 1996, 1997; Novick, 1988; Reed, et al, 1990; Silver, 1981), physics (Chi, et al 1981; Simon & Simon, 1978; Larkin, 1983; Larkin, McDermott, Simon, & Simon, 1980), and computer programming (Adelson, 1984). Similar effects have been observed in a variety of knowledge-lean domains, such as deductive and inductive inference. When presented with deduction problems, untrained reasoners often tended to ignore the argument's logic (i.e., its deep structure) while relying on the argument's surface features, such as content and believability (Evans, Newstead, & Byrne, 1993; Johnson-Laird & Byrne, 1991). When presented with induction and analogy problems, novices and young children also often ignored deep relational structure while relying on the surface features (Gentner, 1989; Holyoak & Koh, 1987).

While there is little disagreement that novices focus on surface features, it remains unclear why novices tend to focus on surface features and not on deep relational properties. One possible explanation of novices' tendency to represent surface features is that novices merely have little knowledge of deep structural relations. However, while this possibility is capable of explaining expert-novice differences in extremely knowledge-demanding domains, such as medical diagnostics, chess, or advanced physics, it falls short of explaining these differences in fairly simple domains, such as elementary mathematics and physics. For example, researchers examining novices' representations in mathematics and physics often drew examples from students' textbooks, thus reasonably assuming that students should be familiar with the deep structure underlying these problems (Chi, et al, 1981; Larkin, 1983; Novick, 1988). The credibility of the lack of knowledge explanation is further undermined by findings that even those novices who receive instruction in a domain often continue to focus on surface features rather than the deep structure of a problem. These has been demonstrated across a variety of knowledge domains, including mathematics (Morris & Sloutsky, 1998) and physics (Kaiser, McCloskey, & Proffitt, 1986; McCloskey, 1983). Finally, the fact that findings on novices' representations in knowledge-lean domains are compatible with those in knowledge-

rich domains makes the low knowledge explanation even less plausible. At this point, however, lack of knowledge cannot be ruled out as an explanation for differences in problem construal by experts and novices. It is also possible that surface features are used more frequently by novices, and, as a result, they are more available than deep relational properties (cf. Anderson, 1990). Henceforth, we will refer to this possibility as the availability explanation.

Another possibility that appears more credible is that even when novices know about deep relations and are capable of extracting these relations, they still fail to represent these relations because surface features are more prominently present in the problem. In failing to represent relational features, they may either fail to encode relations, or these relational features may lose attentional competition to more salient surface features. However, this representational processing explanation can only be tested if the above-described explanations are eliminated as possibilities. In the current paper, then, the focus is on establishing difference in content of representations of experts and novices within the domain of arithmetic, and then testing the knowledge and availability explanations. If differences between experts and novices are found, and the data are inconsistent with the predictions of the alternative explanations, then the way is cleared to test the representational process explanation.

The goal of the current studies, then, is to establish why experts and novices differ in the content of their problem representations. To achieve this goal, we deemed it necessary to control for knowledge factors, while manipulating representational factors. In controlling for knowledge factors, we (a) used simplified tasks and (b) selected only those deep properties that were well familiar to a wide range of participants. In particular, we selected the commutative and associative properties of arithmetic, because these principles are learned in the elementary school and revisited in the beginning of the middle school (Everyday Mathematics: Teacher's Reference Manual, 1998), and therefore are likely to be familiar to the majority of middle school students and college undergraduates.

In this paper, we present two experiments. In Experiment 1, experts and novices in mathematics were asked to group arithmetic equations. These groupings could be based either on the commonality of surface elements (e.g., digits used, the number of constituent elements in the equations) or on the commonality of a deep mathematical relation (principles of commutativity or associativity). In Experiment 2, we introduced a two-phase grouping task. During the first phase, deep relations were "unmasked," such that surface elements were not varied among the compared equations. During the second phase, the deep relations were "masked" again by reintroducing competing surface elements.

## Experiment 1

The goal of this experiment was to validate the principles in question and to eliminate the possibility that expert-novice differences stem from differences in overall intelligence (or age) between novices and experts.

## Method

**Participants** Five samples were selected for the current experiment. The first group, which will be referred to as the "younger children", contained 20 first- and second-graders taken from an elementary school ( $M = 7.26$  years,  $SD = 0.59$ ; 8 girls and 12 boys). The second group, which will be referred to as the "older children", contained 16 sixth-graders taken from a middle school ( $M = 12.10$  years,  $SD = 0.38$ ; 5 girls and 11 boys). Both of these groups were selected from schools located in an upper middle-class suburb of Columbus, Ohio.

The third group of participants consisted of 25 undergraduates in an introductory psychology course at a large Midwestern university who participated for course credit. This group had an average age of 19.78 years ( $SD = 1.38$ ), with 11 women and 14 men.

These three groups of mathematics "novices" were contrasted with a group of mathematics "experts". This group consisted of 20 graduate students in a Mathematics department at the same university who participated for payment of ten dollars. This group had an average age of 28.88 years ( $SD = 6.05$ ), with 7 women and 13 men.

However, differences between "experts" and "novices" were not limited to expertise. Experts were also older and they might represent a self-selected group with respect to an overall ability. Therefore, we deemed it necessary to select a matching group that would be similar to experts in terms of age and overall ability, while differing in the level of expertise. This matching group consisted of 16 graduate students in a History department at the same university who participated for a payment of ten dollars. This group had an average age of 29.93 years ( $SD = 4.67$ ), with 8 women and 8 men.

**Materials** Five features of arithmetic equations were used in Experiment 1. Two of these features were considered "principled properties", in that they represented deep, relational principles of mathematical operations: the associativity and commutativity principles. The former states that for addition, subtraction, and multiplication, constituent parts can be decomposed and recombined in different ways (e.g.,  $a + b = [a - c + c] + b$ ). The latter states that the order of elements is irrelevant for addition and multiplication (e.g.,  $a + b = b + a$ ). The other three features were nonprincipled surface features that occur in arithmetic equations: (1) digits (e.g., 6, 3); (2) sign (e.g., -, +); and (3) the number of constituent terms in an equation. The numerical solutions of equations were controlled for by making these solutions either all equal or all equivalently different for each trial.

A forced-choice similarity paradigm was used in this experiment. Participants were presented with three cards at a time, a target card and two test cards, each which had printed on it an arithmetic equation. Participants were instructed to match the problem on the target card to one of the test problems with which they believed it was most similar. Each of the two test problems shared one feature with the target problem, and differed on the feature that the target shared with the other test problem, with all other fea-

tures held constant. All five features were pitted directly against each other, with the exception of the two principled features, yielding a total of nine feature comparisons. For example, on one of the trial in which commutativity competed with digit, the Target problem was  $6 + 3 + 4 = 3 + 4 + 6$ , the digit test problem was  $6 + 3 + 8 = 3 + 4 + 10$ , and the commutativity test was  $7 + 2 + 8 = 8 + 2 + 7$ .

There were four exemplar arithmetic equations representing each of the nine comparison sets, resulting in a total of 36 trials presented to participants. The numbers used in the arithmetic equations ranged from 1 to 15, and the operations used included addition, subtraction, and multiplication.

**Procedure** All participants were run individually by a male experimenter into a small, quiet room. Participants were instructed that they would be presented with math problems for which they were to group together problems that were similar. A warm-up trial was used to acquaint participants with the task. For the warm-up trial, the participant was presented with cards containing Gelman and Markman’s (1986) blackbird-flamingo-bat figures. The target card, which depicted a blackbird that looked similarly to the bat and dissimilarly from the flamingo, was placed equidistantly below the flamingo and bat cards, which were the test items. The experimenter pointed to each of the two test items, and asked the participant “which of these is more like this,” after which the experimenter pointed to the target item. After participants chose one of the test items, the experimenter asked the participant “why did you choose that one?” After the participant’s verbal explanation (either based on physical similarity or the commonality of species), the experimenter pointed out that the other test item could have also been chosen based on the other attribute, and made the point that similarity can simultaneously occur across multiple dimensions. All participants showed understanding of this concept and of the task.

Four trials for each of the nine features-principle comparisons resulted in a total of 36 trials, which took approximately 30 minutes. Trial order was determined using a block randomization procedure. The positioning of the test items in relation to the target (i.e., left or right) was counterbalanced across comparison type.

## Results and Discussion

The main goal of this experiment was to examine participants’ knowledge of principles in question. To achieve this goal, we considered as choices indicating knowledge only those for which the participants’ explanation of the choice was consistent with the principle. This was done because participants could select principled test stimuli for a reason that might have nothing to do with the principle in question. Only explanations *directly* referring to the principle in questions were considered choice-consistent. The proportion of consistent choices for each principle is the dependent variable used in the forthcoming analyses.

The degree to which participants in each sample made explanation-consistent principled choices was analyzed using a one-way ANOVA for each principle across samples. Table 1 presents overall percentages of explanation-consistent prin-

cipled choices aggregated across trials by principles and age groups. The ANOVA for explanation-consistent associativity choices yielded a significant difference among the samples in the proportion of choices made,  $F(4, 92) = 30.72$ ,  $MSE = .07$ ,  $p < .001$ . The percentage of explanation-consistent associativity choices increases monotonically across the five samples. The ANOVA for explanation-consistent commutativity choices also indicated that there was a significant difference among the samples in the proportion of choices made,  $F(4, 92) = 23.61$ ,  $MSE = .08$ ,  $p < .001$ . As shown in Table 1, the percentage of explanation-consistent commutativity choices also increases monotonically across the five samples.

Table 1: Means and standard deviations (in parentheses) for percentage of explanation-consistent principled choices in Experiment 1.

Sample	Principle	
	Associativity	Commutativity
Younger children	0.00 (0.00)	0.00 (0.00)
Older children	2.08 (6.04)	9.03 (17.32)
Undergraduates	10.66 (26.15)	20.44 (30.71)
History grads	26.39 (38.89)	36.81 (35.54)
Math grads	80.00 (36.34)	77.22 (35.95)

Bonferroni post-hoc tests (with  $\alpha = .05$ ) were used to compare the mean proportion of explanation-consistent principled choices for each sample. These tests yielded identical patterns for both the associativity and commutativity principles, indicating that there were not significant differences in the proportion of explanation-consistent principled choices by younger children, older children, and undergraduates, that History graduate students made significantly more explanation-consistent principled choices than younger children, and that Mathematics graduate students made significantly more explanation-consistent principled choices than each of the other four samples.

Results from Experiment 1 point to several important regularities. First, experts were found to consistently represent principles when categorizing arithmetic equations, whereas novices were more likely to focus on surface features rather than on principles; even when novices did focus on principles, they did so inconsistently. Second, very few younger children exhibited knowledge of principles in question. Third, expert-novice differences were not limited to age or general intelligence: history graduate students and math experts, equally aged groups with similar levels of overall intelligence, exhibited large differences in using deep principled features. Thus the experiment allows us to eliminate the possibility that general ability or development account for expert-novice differences.

However, Experiment 1 left an important question unanswered: it remains unknown why many novices failed to focus on deep principled features. One possibility is that novices merely lack knowledge of these principles. A sec-

ond possibility is that surface elements are more available than deep relational features due to a more frequent use of the former. The goal of Experiment 2 is examine the two possibilities.

## Experiment 2

To accomplish the main goal of this experiment (i.e., to distinguish among the above mentioned possibilities), it was necessary to observe whether novices represent principled features when these features do not compete with surface elements. In the current study, then, participants were given a number of trials in which the target problem shared a principled feature with one of the test problems, and shared no unique surface features with the other test problem. We refer to these trials as “unmasked” since principled features are no longer attentionally “masked” by surface elements.

In addition, in the current experiment the “unmasked” trials are followed by “masked” trials equivalent to the trials in Experiment 1, in which the surface elements are reintroduced to compete with principled features in participants’ similarity judgments. This will enable the examination of the degree to which representations of principled features will be maintained, or whether the surface features will draw attention away from principled features, such that there is no transfer of representation due to the positive learning set. If the former is true, then it is expected that participants’ explanation-consistent principled choices will be more frequent for the subsequent “masked” trials than they were in Experiment 1; if the latter is true, then there should be no difference between the frequency of these choices.

If novices are more likely to make explanation-consistent principled choices in “unmasked” trials, it indicates that they have knowledge of the principles in question, thus undermining the lack of knowledge explanation. If novices are more likely to represent principles in “unmasked” trials but there is no transfer to “masked” trials, this finding would undermine the availability explanation.

## Method

**Participants** Three samples were selected for Experiment 2, each representing a different age group. Two of the groups, the “younger children” and the “older children” used the same participants from Experiment 1; Experiment 2 was conducted approximately four months after Experiment 1 for both samples. The third group of participants consisted of 19 undergraduates in an introductory psychology course at a large mid-western university who participated for course credit. This group had an average age of 21.79 years ( $SD = 6.49$ ), with 12 women and 7 men.

**Materials and Procedure** The same principled features (i.e., associativity and commutativity) and surface features (i.e., digit, sign, and number of elements) used in Experiment 1 were used in Experiment 2. The same nine comparisons used in the previous experiment were again used here for the last 27 trials (three trials for each of the nine comparisons). In addition, in the current experiment, the first

eight trials consisted of ‘unmasked’ comparisons, thus leading to a total of 35 trials.

For the “unmasked” trials, each of the two principled features (i.e., commutativity and associativity) was compared four times against ‘control’ problems. For these trials, the two test problems were equivalently similar to the target on nonprincipled features, while one test problem shared a principled feature with the target problem. For example, for an unmasked-commutativity trial, the target equation was  $2 + 6 + 8 = 6 + 8 + 2$ , the commutativity Test equation was  $11 + 1 + 5 = 5 + 1 + 11$ , and the control Test equation was  $3 + 9 + 5 = 12 + 4 + 1$ .

## Results and Discussion

We first analyze performance in “unmasked” and “masked” trials across the three groups of novices. For purposes of clarity, we will refer to “masked” trials in Experiment 1 as *Masked 1*, whereas “masked” trials in Experiment 2 will be referred to as *Masked 2*. Again, when analyzing performance, we will focus on the proportion of choices made by participants for only the trials in which mathematical principles were present, and we will consider only those choices for which the participants’ explanation of the choice was consistent with the principle. We first present the analyses of Unmasked and Masked 2 trials, followed by comparisons across Masked 1, Unmasked, and Masked 2 conditions.

The degree to which participants in each sample made explanation-consistent principled choices in the Unmasked comparisons was analyzed using a one-way ANOVA for each principle across the three samples. The ANOVAs for explanation-consistent associativity and commutativity choices revealed significant differences among the samples in the proportion of choices made,  $F_s(2, 51) > 5.42$ ,  $ps < .01$ . As evidenced in Table 2, the percentage of explanation-consistent principled choices increased monotonically with age.

Table 2: Means and standard deviations (in parentheses) for percentage of explanation-consistent principled choices in Experiment 2.

Sample	Trial Type	
	Unmasked	Masked 2
Associativity		
Younger Children	0.00 (0.00)	1.17 (5.10)
Older Children	20.31 (29.18)	6.94 (15.11)
Undergraduates	27.63 (36.22)	18.13 (27.27)
Commutativity		
Younger Children	15.79 (30.29)	1.75 (7.65)
Older Children	67.19 (29.89)	15.28 (22.18)
Undergraduates	90.79 (20.77)	35.09 (36.71)

Bonferroni post-hoc tests (with  $\alpha = .05$ ) were used to compare the mean proportion of explanation-consistent principled choices for Unmasked trials for each sample. For associativity trials, this test indicates only one statistically



significant difference among samples, that undergraduates made more explanation-consistent associativity choices than younger children. However, for commutativity trials, all between-sample comparisons were statistically significant. Aggregated across both principles, less than 10% of the younger children's responses were principle-based, while almost 50% of older children's responses and over 60% of undergraduate students' responses were principle-based.

It should be noted that there were large differences in the proportion of participants focusing on commutativity and associativity, with the former being greater than the latter. However, even for associativity, where effects were smaller than for commutativity, around 50% of older children and undergraduates provided at least one explanation-consistent principled choice, thus exhibiting knowledge of the principle in question.

The degree to which participants in each sample made explanation-consistent principled choices in the Masked 2 comparisons was also analyzed using a one-way ANOVA for each principle across samples. The ANOVAs for explanation-consistent associativity and commutativity choices yielded a significant difference among the samples in the proportion of choices made,  $F_s(2, 51) > 4.1$ ,  $p_s < .05$ . Again, as evidenced in Table 2, the percentage of explanation-consistent principled choices increases monotonically with age.

Bonferroni post-hoc tests (with  $\alpha = .05$ ) were again used to compare the mean proportion of explanation-consistent principled choices for Masked 2 trials for each sample. For both principles, this test indicates only one statistically significant difference among samples that undergraduates made more explanation-consistent principled choices than did younger children. These data in conjunction with the results of the Unmasked condition suggest that even when participants knew the principle in question, they often focused on surface features.

Overall proportions of explanation-consistent principled choices in Masked 1, Unmasked, and Masked 2 trials aggregated across the principles and broken down by sample are presented in Figure 1. Participants' explanation-consistent principled choices on Unmasked trials generally increased in comparison to their choices on Masked 1 trials. Younger children gave more explanation-consistent commutativity choices for Unmasked trials than for Masked 1 trials ( $t = 2.27$ ,  $p < .05$ ), though there was not a significant difference in the amount of explanation-consistent associativity choices, which is due to a floor effect. Older children gave more explanation-consistent principled choices for Unmasked trials than for Masked 1 trials for both principles ( $t = 2.65$ ,  $p < .02$  for associativity, and  $t = 9.8$ ,  $p < .001$  for commutativity). Undergraduates gave more explanation-consistent commutativity choices for Unmasked trials than for Masked 1 trials ( $t = 8.59$ ,  $p < .001$ ), though there was a marginally significant difference in the amount of explanation-consistent associativity choices ( $t = 1.81$ ,  $p = .078$ ). These differences indicate that unmasking increased the proportion of principled choices in all samples.

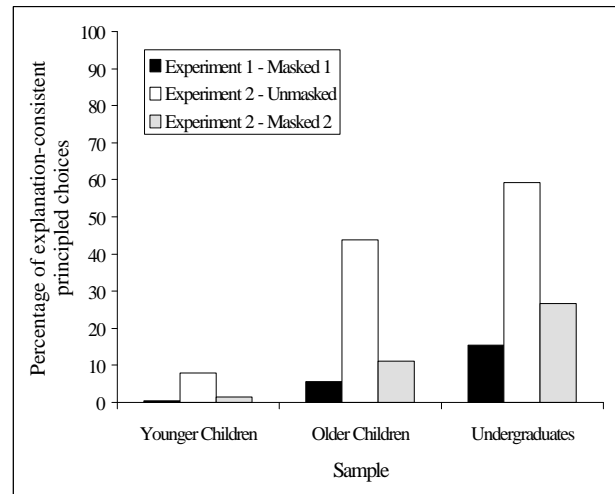


Figure 1. Percentage of explanation-consistent principled choices for each sample for unmasked” and “masked” trials in Experiment 2, and “masked” trials in Experiment 1.

Participants' explanation-consistent principled choices on Masked 2 trials generally decreased in comparison to their choices on Unmasked trials. Younger children gave more explanation-consistent commutativity choices for Unmasked trials than for Masked 2 trials ( $t = 2.37$ ,  $p < .05$ ), though there was not a significant difference in the amount of explanation-consistent associativity choices, which is due to a floor effect. Older children gave more explanation-consistent principled choices for Unmasked trials than for Masked 2 trials for both principles ( $t = 2.74$ ,  $p < .02$  for associativity, and  $t = 8.15$ ,  $p < .001$  for commutativity). Undergraduates gave more explanation-consistent commutativity choices for Unmasked trials than for Masked 2 trials ( $t = 7.23$ ,  $p < .001$ ), though there was a marginally significant difference in the amount of explanation-consistent associativity choices ( $t = 1.81$ ,  $p = .078$ ). These differences indicate that there was not pure transfer of representations from Unmasked to Masked 2 trials: once principled features had to compete again with surface features, the number of explanation-consistent principled choices decreased markedly.

An important question is whether the transfer led to a significant increase of explanation-consistent principled choices compared to when participants were never exposed to Unmasked trials. That is, whether being exposed to a positive learning set significantly increased subsequent attention to principles. To answer this question, we compared participants' explanation-consistent principled choices on the Masked 1 and Masked 2 trials. While the proportions of explanation-consistent principled choices are somewhat larger for each sample and each principle on Masked 2 trials than for Masked 1 trials (as evidenced in Figure 3), t-tests for each comparison revealed that none of these differences are statistically significant. Thus, the positive learning set of the Unmasked trials had a nonsignificant effect on the degree to which participants represented principled features of mathematics problems.

Overall, results of Experiment 2 indicate that 94% of the middle school participants and 100% of the undergraduate participants exhibited knowledge of principles in question

(i.e., provided an explanation-consistent principled choice on at least one trial), focusing on these principles in Unmasked trials. This finding severely undermines the lack of knowledge explanation. At the same time, the increase in younger children's principled choices due to "unmasking" was rather small, which points to a lack of knowledge. However, even in the two older groups, once nonprincipled features were reintroduced, representation of principled properties attenuated to levels similar to Experiment 1, a finding that undermines the availability explanation.

## Conclusion

The results of the two reported experiments establish a difference in the content of expert and novice representations for arithmetic problems. These results suggest that the observed differences do not stem from a lack of knowledge of deep principles by novices. The results further suggest that differences in content of problem representations in experts and novices may stem from different processing mechanisms underlying the construal of problem representations in experts and novices. The research presented in Part 2 will focus the examination of the processes of construal of problem representations by expert and novices.

## Acknowledgments

This research has been supported by a grant from James S. McDonnell Foundation to the second author.

## References

- Adelson, B. (1984). When novices surpass experts: The difficulty of a task may increase with expertise. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *10*, 483-495.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Bassok, M. (1996). Using content to interpret structure: Effects on analogical transfer. *Current Directions in Psychological Science*, *5*, 54-58.
- Bassok, M. (1997). Two types of reliance on correlations between content and structure in reasoning about word problems. In L. D. English (Ed.), *Mathematical reasoning: Analogies, metaphors, and images*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Blessing, S. B., & Ross, B. H. (1996). Content effects in problem categorization and problem solving. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *22*, 792-810.
- Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55-81.
- Chi, M. T. H., Feltovich, P. G., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121-152.
- Evans, J. St. B. T., Newstead, S. E.; Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, England: Lawrence Erlbaum Associates.
- Everyday Mathematics: *Teacher's Reference Manual* (1998). Chicago, IL: Everyday Learning.
- Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*, 183-209.
- Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. New York: Cambridge University Press.
- Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15*, 332-340.
- Johnson-Laird, P., & Byrne, R. (1991). *Deduction*. Hove, UK: Lawrence Erlbaum.
- Kaiser, M. K., McCloskey, M., & Proffitt, D. R. (1986). Development of intuitive theories of motion: Curvilinear motion in the absence of external forces. *Developmental Psychology*, *22*, 67-71.
- Larkin, J. (1983). The role of problem representation in physics. In D. Gentner & A. Stevens (Eds.), *Mental models*. Hillsdale, NJ: Erlbaum.
- Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Models of competence in solving physics problems. *Cognitive Science*, *4*, 317-345.
- McCloskey, M. (1983). Naïve theories of motion. In D. Gentner & A. Stevens (Eds.), *Mental models*. Hillsdale, NJ: Erlbaum.
- Morris, A. K., & Sloutsky, V. M. (1998). Understanding of logical necessity: Developmental antecedents and cognitive consequences. *Child Development*, *69*, 721-741.
- Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *14*, 510-520.
- Reed, S. K., Ackinclose, C. C., & Voss, A. A. (1990). Selecting analogous problems: Similarity versus inclusiveness. *Memory & Cognition*, *18*, 83-98.
- Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *8*, 484-494.
- Silver, E. A. (1981). Recall of mathematical problem information: Solving related problems. *Journal of Research in Mathematics Education*, *24*, 117-135.
- Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. In R. S. Siegler, (Ed), *Children's thinking: What develops?* Hillsdale, NJ: Lawrence Erlbaum Associates.

**This page left blank intentionally.**

**This page left blank intentionally.**

**This page left blank intentionally.**

# Accentuation of category differences: Revisiting a classic study

Janet K. Andrews (andrewsj@vassar.edu)  
Kenneth R. Livingston (livingst@vassar.edu)  
Vassar College Program in Cognitive Science  
124 Raymond Avenue, Poughkeepsie, NY 12604 USA

## Background

In 1963, Tajfel and Wilkes reported a study using as stimuli simple lines that were either unlabeled or were labeled with a letter A or B. Participants simply estimated how long the lines were, and the dramatic finding was that when the attachment of letter labels was systematically related to line length, such that A essentially meant "short" and B "long," participants significantly overestimated the difference between the longest A line and the shortest B line. This effect came to be known as the accentuation of intercategory difference, and Tajfel and Wilkes (1963) proposed that this basic cognitive/perceptual bias could explain the exaggeration of perceived differences between members of different social groups, thereby explaining an aspect of stereotyping. In the decades since the Tajfel and Wilkes (1963) study an extensive body of cognitive social psychology has been built on their results.

There is, however, reason to be doubtful of the strength of Tajfel and Wilkes' original findings. Recent research in the literature on basic perceptual and cognitive processes has also been exploring the operation of effects like those that Tajfel and Wilkes described as accentuation (e.g., Goldstone, 1994; Livingston, Andrews, and Harnad, 1998). One of the discoveries from this work is that it is extraordinarily difficult to demonstrate such effects when stimuli vary in only one dimension, a conclusion that seems to apply specifically to the case of lines varying only in length.

We think that Tajfel and Wilkes' results are consistent with an alternative explanation based on the demand characteristics of the task situation, rather than genuinely altered perceptual processing. In particular, when a category distinction is imposed on a continuous variation in the stimuli, it may have the effect of telling participants that they are supposed to treat within-category items as more similar or even as identical. While this is a real effect of category information on judgment, it is very different from the kind of basic, perceptual, and essentially involuntary process that these accentuation results have been taken to represent. We believe that it is very important to determine which interpretation of these line-length accentuation effects is correct. We therefore performed a series of studies intended to clarify the status of Tajfel and Wilkes' classic and much-cited results.

## Experiments and Results

Three replications of the Tajfel and Wilkes study were done, two using their original procedures with lines drawn on cards (Experiments 1 and 3), and one using computerized

presentation (Experiment 2). Experiment 1 produced results very similar to those of Tajfel and Wilkes, namely, a statistically significant overestimation of the difference in length between the longest A line and the shortest B line. However, surprisingly, Experiments 2 and 3 produced no such effects, but instead, accurate line length estimates across all stimuli. A re-examination of the data from Experiment 1 revealed that the group results were due to a very small subset of participants giving *identical* length estimates of the four lines in each labeled class. When their data were removed, the Tajfel and Wilkes' accentuation effect completely disappeared. Line length estimation was essentially accurate whether the lines were labeled or not.

## Conclusion

We believe that our results show that the original finding of intercategory difference accentuation reported by Tajfel and Wilkes is both extremely fragile and almost certainly the result of demand characteristics rather than altered perception of stimuli. This raises the question of how to interpret the accentuation effects found in the many studies using explicitly social stimuli that came after Tajfel and Wilkes. It is possible that these, too, reflect demand characteristics. However, they may be reflecting processes similar to those that produce real effects of categorization on the perception of nonsocial multidimensional stimuli. Work by Goldstone (1994), Livingston et al. (1998), and others has demonstrated both accentuation of intercategory differences and accentuation of intracategory similarity, but only with *multidimensional* stimuli and only when the categories require *learning* (as opposed to the simple attachment of labels to known categories, as in Tajfel and Wilkes). This suggests that it may be possible to explain social psychological accentuation effects in terms of more basic perceptual/cognitive processes after all, a possibility that further research can and should examine more fully.

## References

- Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123, 178-200.
- Livingston, K., Andrews, J., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 732-753.
- Tajfel, H., and Wilkes, A. (1963). Classification and quantitative judgement. *British Journal of Psychology*, 54, 101-114.

# Training and Transfer of Foreign Word Identification at Three Speeds

**Anita R. Bowles** (bowlesa@psych.colorado.edu)

Department of Psychology, Campus Box 345, University of Colorado, Boulder, CO 80309-0345

**Alice F. Healy** (ahealy@psych.colorado.edu)

Department of Psychology, Campus Box 345, University of Colorado, Boulder, CO 80309-0345

It is unclear whether slowing one's speech rate when talking to foreign language students improves their comprehension and ability to segment spoken input into separate words. Although some studies have found that slowed speech facilitates understanding (Flaherty, 1979), others showed either no effect (Blau, 1990) or an effect only for beginning foreign language students (Griffiths, 1992). The current series of experiments adds to this literature by examining the relationship between speech rate and beginning learners' foreign word identification in sentences.

## Method

### Subjects

In Experiment 1, 72 subjects participated; in both Experiments 2 and 3, 36 subjects participated. These subjects had no previous knowledge of Spanish; they were assigned to conditions in a pseudorandom order.

### Procedure

In each experiment, during training, subjects listened to 12 Spanish sentences containing words included in a beginning Spanish textbook pronounced by a native speaker. The sentences were presented one at a time through a computer. After each sentence was presented, the subjects attempted to type it and then received feedback in the form of the correctly spelled sentence. The sentences were presented in blocks of 12, with each subject trained for eight blocks.

In Experiment 1, subjects were assigned to one of four training lists. One third of the subjects given each list trained with sentences presented at a normal conversational rate (145 words per minute). Another third trained on sentences slowed to 104 wpm. The last third trained on speeded sentences presented at 203 wpm. SoundEdit software was used to expand or compress the normal speech used for the medium speed to create stimuli at the slow and fast speeds. Subjects were then tested on 48 Spanish sentences (the 12 sentences in each of the four training lists). One third of the sentences from each list were presented at the slow speed, one third at the medium speed, and the last third at the fast speed.

In Experiments 2 and 3, only two of the lists were used at training, but all four lists were used at test. Three slower speeds were used (70, 97, and 134 wpm). The stimuli in these experiments were produced at all three speeds by the native speaker. No computer expansion or compression was used. In Experiment 2, before the beginning of training, half of the subjects were given visual pretraining on the spelling

of the Spanish words in their training list, and half received no pretraining. During pretraining, subjects saw each word and copied it three times. In Experiment 3, half of the subjects were given visual pretraining, and half were given both visual pretraining and auditory pretraining on the sound of the words in isolation.

## Results

Subjects in all three experiments who trained at the fast speed had a significantly lower proportion of correct responses during training than did subjects who trained at either the slow or medium speeds. There was also a main effect of testing speed, with subjects scoring worse for sentences tested at the fast speed than at the slow or medium speeds. Importantly, in Experiments 1 and 2, subjects trained at the fast speed did significantly worse at test than did those trained at the slow or medium speeds, but there was no difference between the slow and medium speed groups. In contrast, in Experiment 3 (in which all subjects received some form of pretraining), there was no effect of training speed on accuracy at test.

## Conclusions and Implications

These experiments show that words in rapid speech are difficult for beginning foreign language learners to identify within sentences. However, with pretraining on individual words, the negative effects of training at a rapid rate are reduced. More importantly, the results suggest that slowed speech is not better for word identification than is normal conversational speech, even for novice foreign language students. These findings imply that slowing speech is not necessary for beginning foreign language instruction.

## Acknowledgments

This work was supported by Army Research Institute Contracts DASW01-96-K-0010 and DASW01-99-K-0002 and Army Research Office Grant DAAG55-98-1-0214.

## References

- Blau, E. K. (1990). The effect of syntax, speed, and pauses on listening comprehension. *TESOL Quarterly*, 24, 746-753.
- Flaherty, E. (1979). Rate-controlled speech in foreign language education. *Foreign Language Annals*, 12, 275-280.
- Griffiths, R. (1992). Speech rate and listening comprehension: Further evidence of the relationship. *TESOL Quarterly*, 26, 385-390.

# Temporal Tuning in the Acquisition of Cognitive Skill

Richard A. Carlson (racarlson@psu.edu)

Lisa M. Stevenson (lms152@psu.edu)

Department of Psychology, The Pennsylvania State University

University Park, PA 16802

Fluent cognitive performance depends on concurrent activation of appropriate representations of goals, procedures, and data. This observation is reflected in a number of current theories such as Anderson's ACT-R (Anderson & Lebiere, 1998). Together with the observation that part of the improvement with practice of many skills can be attributed to better coordination of actions with the environment (e.g., Neisser, 1992), this suggests what we call the *temporal tuning hypothesis*: As a consequence of practice, individuals will adjust the timing of their mental and information pickup activities so that the activation of information to be processed is optimally synchronized with ongoing mental activity. Previous research has demonstrated the learning of timing constraints in perceptual-motor (Dominey, 1998) and cognitive tasks (Carlson, Shin, & Wenger, 1994). We examined this hypothesis in three experiments in which participants performed computerized multiple-step arithmetic or spatial tasks.

We examined the possibility of temporal tuning by allowing subjects to control the pacing of their problem-solving performance by pressing keys to briefly display part of the information required for each step. Experiments 1 and 2 used a running arithmetic task in which subjects updated a total at each step. Subjects practiced solving 8-step problems for 10 blocks of 10 trials. In Experiment 1, operators for all steps were visible throughout each trial, and a new operand was displayed in response to a keypress at each step. In Experiment 2, both operator and operand appeared sequentially in response to the keypress. In both cases, the new information was displayed briefly, then masked. We manipulated constraints on timing by varying between subjects the delay between each keypress and the display of the available information. Delays ranged from 200 to 1100 ms. Experiment 3 used a spatial path-construction task with procedures and design similar to Experiment 2, to provide generality across task domains. In all experiments, we changed the timing constraints in final test blocks to verify that the practiced constraints had been learned.

If the structure of mental processes for performing each step allows temporal tuning, with practice subjects should learn to anticipate when they will be ready for the new information and request that information at a time that takes into account the delay. This would result in shorter keypress latencies with longer delays, measured from the onset of the information needed complete a step. Information requests might be initiated on a rhythmic basis, or on the basis of internal or external events that serve as process completion markers. If, on the other hand, participants must wait until a

step is completed to instantiate a goal for the following step, keypress latencies will not vary as a function of delay. This might be the case if, for example, problem-solving steps are realized by production rules with inaccessible internal structures.

In all three experiments, we found evidence of temporal tuning: With practice, subjects in conditions with longer delays learned to request information earlier than did those in conditions with shorter delays. When operators for all steps were continuously visible (Experiment 1), temporal tuning was more precise than when the operator for each step was displayed only on request (Experiments 2 and 3). For the arithmetic task, performance was also slower and less accurate when operators appeared step by step. This difference suggests constraints on the ability to anticipate and control the timing of mental activity. One possible constraint is that a goal based on the operator to be applied must be instantiated to initiate a procedure that provides a basis for anticipatory timing.

We consider alternative accounts of how temporal tuning might be accomplished. It appears that typical production-system models of cognitive skill would have to be extended to accommodate the phenomenon of temporal tuning. We consider the implications of this phenomenon for the role of on-line instantiation of goals in theories of cognitive skill.

## References

- Anderson, J. R., & Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Erlbaum.
- Carlson, R. A., Shin, J. C., & Wenger, J. L. (1994). Timing and the control of fluent cognitive sequences. Presented at the 35th annual meeting of the Psychonomic Society, St. Louis, November 12, 1994.
- Dominey, P. F. (1998). Influences of temporal organization on sequence learning and transfer: Comments on Stadler (1995) and Curran and Keele (1995). Journal of Experimental Psychology: Learning, Memory, and Cognition, 24, 234-248.
- Neisser, U. (1992). The development of consciousness and the acquisition of skill. In F. S. Kessel, P. M. Cole, & D. L. Johnson (Eds.), Self and consciousness: Multiple perspectives, (pp. 1-18). Hillsdale, NJ: Erlbaum.



# Hemispheric Effects in Fusiform Gyrus Across Face Encoding Tasks

Daniel J. Casasanto (dcasasan@mail.med.upenn.edu)

John A. Detre (detre@mail.med.upenn.edu)

Department of Neurology, University of Pennsylvania Medical Center  
3400 Spruce Street, Philadelphia, PA 19104 USA

## Introduction

Functional Magnetic Resonance Imaging (fMRI) was used to examine neuronal activation during two explicit face memory tasks. Numerous neuroimaging studies have shown bilateral activation of posterior temporal-occipital structures during processing of visually presented stimuli; preferential left-hemisphere activation has been associated with object processing, and preferential right-hemisphere activation has been associated with face processing (Sergent, 1995). Lesion studies suggest a special role for the right fusiform gyrus in the encoding of structural physiognomic information, an early stage of face processing (Sergent, 1993). It was hypothesized that intentional encoding of unfamiliar faces would be associated with preferential activation of right-hemisphere mesial temporal lobe structures, including the fusiform gyrus.

## Methods

### Task Design

For each of two face memory encoding tasks, healthy, right-handed volunteers viewed blocks of unfamiliar face photographs, alternating with blocks of a repeatedly presented pixelated control image (six 40s task/control blocks, 10 stimuli per block, 3.5s presentation, 0.5s ISI). Face stimuli were constructed from University of Pennsylvania ID card photographs. For the first task, full-head photographs were shown, including hair, neck, and upper shoulders. In some cases, clothing and jewelry were visible. For the second task, the same set of face photographs was used, but each photograph was cropped so as to include the brow, eyes, nose, and mouth, but exclude ears, hair, and any extraneous items. Two separate groups of six subjects were consecutively recruited for each of the two tasks. Subjects were instructed to remember the faces for a post-scan recognition test, and to attend the control images but not to memorize them. Scanning occurred during the encoding tasks but not during recognition testing.

## Image Acquisition and Processing

BOLD functional imaging data were collected at 1.5 Tesla in 20 contiguous 5mm axial slices, using a GE Signa EchoSpeed MRI scanner. Data were corrected for motion and static susceptibility-induced artifacts, and transformed into three-dimensional space. Using SPM 97 software, a statistical parametric map was constructed for each subject. Group activation maps were then constructed for each task using the SPMt Random Effects model. Activation exceeding a mapwise statistical threshold ( $\alpha=.05$ ) was quantified within the right and left fusiform gyri, and was compared using a hemispheric asymmetry ratio ( $AR=R-L/R+L$ ).

## Results

Suprathreshold activation was found bilaterally during both encoding tasks. Activation associated with encoding of full-head stimuli was slightly greater left than right ( $AR=-0.20$ ;  $Z=.91$ , *ns*). (See figure 1a.) Activation associated with encoding of cropped face stimuli was significantly greater right than left ( $AR=0.31$ ;  $Z=-2.45$ ,  $p<.05$ ). (See figure 1b.)

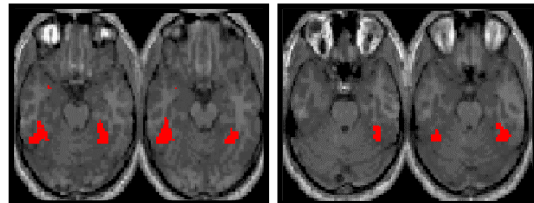


Figure 1a.

Figure 1b.

## Discussion

Cropped face encoding elicited the hypothesized preferential right-sided activation in the fusiform gyrus, while full-head encoding did not. One possible explanation for these findings is that the former task constrained subjects' encoding strategies to the visuospatial domain, while the latter task allowed verbal encoding of nameable objects, as well.

## References

- Sergent, J. (1993), *The functional organization of the human visual cortex*. Oxford: Pergamon.
- Sergent, J. (1995), *Brain Asymmetry*. Cambridge: MIT Press.

# Detecting Animals in Point-Light Displays

Leslie Cohen (lcohen@nimbus.ocis.temple.edu)

Temple University Department of Psychology, Weiss Hall, Philadelphia, PA 19122

Thomas F. Shipley (tshipley@astro.ocis.temple.edu)

Temple University Department of Psychology, Weiss Hall, Philadelphia, PA 19122

Eve Marshark (emarscha@unix.temple.edu)

Temple University Department of Psychology, Weiss Hall, Philadelphia, PA 19122

Kathy Taht & Denise Aster (ktaht@njaquarium.org)

New Jersey State Aquarium, 1 Riverside Drive, Camden, New Jersey 08103

## Introduction

In his work on the perception of biological motion, Johansson found that people can readily detect a human figure in point-light displays — displays where the motion of human walkers is represented by lights corresponding to major human joints (Johansson, 1973). Following up on his work, researchers have examined a number of effects related to the perception of biological motion. Observers have been shown to accurately detect gender and identify specific types of motion, such as dancing. It appears that people have a special ability to detect upright human figures in these, and similar, displays. When such figures are inverted accuracy of figure identification and detection sharply declines (Bertenthal and Pinto, 1994). Inversion effects such as those found in humans have been found by Pinto and Shiffrar (1999) in some non-human animal displays (horses and dogs), but not others (birds). It has been hypothesized that the ability to detect upright humans and the inability to detect some animals and inverted humans can be linked to the human motor system. Observers walk, and the information provided by their own walking may help organize the complex motion patterns that are present in point-light displays. Another possibility is that experience provides an organizational framework for point-light displays. If this were so, previous findings that observers did not accurately identify non-human animals and showed no inversion effect might both be traced to a lack of pertinent experience. Most of the subjects who participated in these studies had had extensive experience observing and interacting with moving people; few had a comparable history of interaction with non-human animals. The current study examines the potential role of experience in the identification and detection of animal figures in masked point-light displays.

## Methods

To test the effect of experience on the perception of point-light animals, the performance of professional seal trainers was compared with that of professional dog trainers and naive subjects on detection of point-light seals, dogs and humans. Subjects included professional seal trainers from the Camden Aquarium in New Jersey, professional dog trainers from the Philadelphia area, and Temple University undergraduates. On average, seal trainers had been employed by the aquarium or a similar agency for 3 years and dog trainers had spent 4 years training dogs at the time of this study. Dog trainers had no professional experience with seals; 5 of the 7 seal trainers had dogs as pets, and one also worked as a professional dog trainer.

Displays were generated from a video segment of a seal, dog or human walking. Seals, dogs, and humans were marked with spots at homologous joints and then videotaped as they

moved from one place to another on land. A 2-second point-light display was generated for each animal. Subjects were presented with a signal-detection task, in which they were to determine the presence and absence of point-light humans, seals and dogs when presented within a set of masking points. Each species was presented upright and upside-down. Two levels of masking were used. Signal-present displays had either one masking point for each point on the animal or two masking points for each point on the animal. Signal-absent displays were generated by combining 2 or 3 sets of masking points so that they had the same number of elements as the corresponding signal-present display. Masking points were generated by randomly perturbing the spatial location and phase of each element in the display. Subjects were shown a target display where the stimulus was shown repeatedly over a period of 20 seconds without any masking elements. They were then asked to decide whether that target was present in each of the following 40 trials. Each subject completed one block of trials for each of the 12 conditions.

## Results and Discussion

All groups detected humans more accurately than seals or dogs. There was no overall effect of expertise, seal trainers were no better than the other subjects at detecting seals, and vice versa. All subjects showed an inversion effect for humans, but there was no inversion effect for familiar animals. If anything, the opposite of the anticipated effect was found—a small inversion effect was present for the less familiar animal (e.g., seal trainers were better at detecting right-side-up dogs than up-side-down dogs). These findings suggest that experience does not play a role in the grouping of complex motion in point-light displays. These results support an account of perception of point light displays that is based on some unique, perhaps structural, aspect of humans. They may reflect the use of a motor code to represent motion. Such a code, which might normally allow us to copy the movements of others, might also unify the elements of a point-light display.

## References

- Bertenthal, B. I. & Pinto, J. (1994). Global processing of biological motions. *Psychological Science*, 5, 4, 221-225.
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14, 201-211.
- Pinto, J. & Shiffrar, M. (1999). Visual analysis of human and animal biological motion displays. *Abstracts of the Psychonomic Society*, 4, 1.
- Sumi, S. (1984). Upside down presentation of the Johansson moving light spot pattern. *Perception*, 13, 283-286.

# Familiarity and Categorical Inference

David Collister ([dc@psych.stanford.edu](mailto:dc@psych.stanford.edu))

Department of Psychology, Bldg. 420, Jordan Hall, Stanford University,  
Stanford, CA 94305-2130

Barbara Tversky ([bt@psych.stanford.edu](mailto:bt@psych.stanford.edu))

Department of Psychology, Bldg. 420, Jordan Hall, Stanford University,  
Stanford, CA 94305-2130

Many if not most categories have internal structure in that more “central” category members evoke optimal responses across a number of measures, including “goodness-of-example” ratings, priming, category verification times, production frequencies, and rates of learning. A number of studies (Rips, 1975; Osherson, Smith, Wilkie, López, & Shafir, 1990; Sloman, 1993) have presented evidence that people rely on internal structure when making inferences about members of a category. Atypical category members are judged more likely to have the properties of typical members, rather than vice versa. Similar members are judged more likely to share a property than are dissimilar ones.

Based on this evidence, models of categorical inference have been proposed that assume 1) category structure is due the number of properties shared by members, and 2) categorical inference operates across these properties (e.g., Osherson, et al., 1990; Sloman, 1993). However, not all measures of category structure appear to be about shared properties. For example, verification times and production frequencies are more closely related to the availability and familiarity of members rather than what properties they have in common. Further, measures based on these different types of category structure are not perfectly correlated. Some members are more typical than they are familiar. Other have the reverse relationship. Thus, familiarity may be another source of category structure for inference to operate over - one based more on the frequency of occurrence rather than the number of shared properties.

Four experiments were conducted that examined the role of familiarity in categorical inference. In all experiments, participants were shown one-premise syllogisms about various category items, and asked to evaluate the likelihood that the syllogisms were true. Items were selected from a number of natural and artifact categories such that some items varied in familiarity within different levels of typicality, and others had the reverse relation. In addition, syllogisms were about “blank” properties to minimize participants’ reliance on background knowledge and maximize their reliance on category structure (see Osherson, et al., 1990). In experiments 1 & 2, an asymmetric effect of familiarity was found that was opposite the usual effect of typicality: Participants were less likely to make inferences from familiar rather than unfamiliar items (experiment 1), and more likely to make inferences to familiar rather than unfamiliar items (experiment 2). In a third experiment, the

effect of familiarity was diminished when participants were asked to explain why they thought some syllogisms were better than others. Further, almost every reason given for preferring one syllogism over another was one based on some similarity between items, even when the similarity was acknowledged to be negligible. In the final experiment, the availability of items was increased through repeated exposure. Effects paralleled that of familiarity: Participants preferences for syllogisms increased and decreased with the availability of the conclusion and premise items respectively. The pattern of results across the experiments suggest that categorical inference may be affected differently by analytic versus nonanalytic task demands (e.g., Whittlesea & Price, 1999). When allowed to evaluate syllogisms without analytic demands (i.e., without having to give explicit justifications) people may be influenced (at least partly) by the availability of the items. For example, the fluency of processing that accompanies both more available and more familiar items may be a general phenomenon that accompanies a number of different cognitive processes. In this case, participants may misattribute the feelings of fluency as arising from some other process relevant to the problem at hand, e.g., an estimate of the prior likelihood that the items in question share a property, etc. However, when asked to justify their inferences people have to at least report if not rely on strategies that are more easily identified. In this case, people may discount ‘free-floating’ feelings of fluency and instead look for describable properties and relations between items that they can use to justify a response.

## References

- Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185-200.
- Rips, L. J. (1975). Inductive judgements about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*, 665-681.
- Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231-280.
- Whittlesea, B. W. A. & Price, J. R. (1999). Implicit/explicit memory versus analytic/nonanalytic processing: Rethinking the mere exposure effect. Manuscript in preparation, Simon Fraser University.

# How Words Get Special

**Eliana Colunga** (ecolunga@cs.indiana.edu)  
Department of Computer Science; Lindley Hall 215  
Bloomington, IN 47408 USA

**Linda B. Smith** (smith4@indiana.edu)  
Department of Psychology; 1101 East Tenth Street  
Bloomington, IN 47405 USA

Words seem to have a special status among perceptual signals. Having a label for an object changes the way it is categorized for both adults and children. For example, when asked to generalize an object name to new instances, children and adults generalize by shape. However, when asked to find an object that “goes with” another, they choose by overall similarity. A label also makes children’s choices shift from thematic to taxonomic and from surface to more conceptual similarities.

Recent studies by Woodward and Hoyne (1999) and Namy and Waxman (1998) suggest that the power of words is not there at the beginning of development but rather that it emerges. At 13 months of age, babies seem willing to pair objects with any kind of signal, such as gestures and non-linguistic sounds. However, by 20 months of age children are more constrained in what they will take as a label, only taking words as labels for objects. This paper is concerned with how this special status of words develops. We propose that words get their special status by virtue of being systematically used for labeling categories. We present a connectionist model of this process and test a prediction that derives from the model.

## The Model

We use a simple settling network to model an abstract version of Woodward and Hoyne’s results. The network has an Auditory Signal Layer and a Visual Signal Layer connected through a Hidden Layer.

The training set consists of 20 “words” and their corresponding “objects”. The words are presented on the Auditory Signal Layer and the objects on the Visual Signal Layer. We assume words are drawn from a constrained space of the possible values of the auditory dimension. The training set is constructed by randomly generating “words” and their corresponding “objects”; the pairings of words to objects are, thus, arbitrary. At the start of learning, words (that is, input from the constrained portion of the auditory space) have no special status over other inputs that may be paired with objects. During training, the word and its corresponding object (plus noise) are presented together and weights are updated using Contrastive Hebbian Learning. So, during training individual objects are systematically paired with words and unsystematically paired with other auditory or visual inputs.

After the network has reached 90% accuracy in the training set, the network is trained on novel word—object pairs and novel non-word—object pairs. Like the older children in

Woodward & Hoyne (1999), the network shows an advantage when learning novel word-object pairs, that is when pairing objects to patterns in the Auditory Signal layer which are within the constrained space of words.

In this model, all that matters for achieving “special status” is the systematic pairing of objects with points in a constrained region of auditory space. Thus, any signal that correlates systematically with any feature becomes subsequently easily associated with it. Such systematic correlations do exist in the input to children, beyond words as labels for objects. For example, animals make sounds, so animals (animate features) should become easily associated with (animal-like) sounds. In the following experiment we test this prediction.

## The Experiment

This study follows Woodward and Hoyne’s procedure, except that the objects used are all unusual animal toys. Thirty-six 13 month-olds and thirty-six 20 month-olds were shown an animal and the animal was labeled for them. In the Word condition the object was labeled with a novel word (i.e. “Look! Dax See? Dax”). In the Animal Sound condition, the object was labeled with a non-linguistic vocal sound (i.e. “Look! Yeep yeep yeep See? Yeep yeep”). In the Arbitrary Sound condition a non-linguistic, non-vocal sound (i.e. a clap) was used instead. Between training trials, the babies were shown and allowed to play with toy animals that later served as distracters during the test phase.

During the test phase, children were presented with the target object and a distracter on a tray. The child was then asked, “Can you get the <label>?”. The baby’s choice was coded as the object that he or she removed from the tray.

The results show that while 13 month-olds in all three labeling conditions learn the label-animal correspondences, 20 month-olds only learn the associations in the Word and Animal Sound conditions. This result suggests that it is the systematicity of prior learned pairings that determine which associations will be formed.

## References

- Namy, L. L., & Waxman, S. R. (1998). Words and gestures: Infants’ interpretations of different forms of symbolic reference. *Child Development, 69*, 295—308.
- Woodward, A. L., & Hoyne, K. L. (1999). Infants’ Learning about Words and Sounds in Relation to Objects. *Child Development, 70*, 65—77.

# Information Processing Speed During Functional Neuroimaging of Sentence Comprehension

**Ayanna Cooke (acooke@mail.med.upenn.edu)**

Department of Neurology; University of Pennsylvania Medical Center, 3400 Spruce St.  
Philadelphia, PA 19104

**Christian DeVita (devita@wernicke.med.upenn.edu)**

Department of Neurology; University of Pennsylvania Medical Center, 3400 Spruce St.  
Philadelphia, PA 19104

**David Alsop (alsop@oasis.rad.upenn.edu)**

Department of Radiology; University of Pennsylvania Medical Center, 3400 Spruce St.  
Philadelphia, PA 19104

**James Gee (gee@grip.cis.upenn.edu)**

Department of Radiology; University of Pennsylvania Medical Center, 3400 Spruce St.  
Philadelphia, PA 19104

**John Detre (detre@mail.med.upenn.edu)**

Department of Neurology; University of Pennsylvania Medical Center, 3400 Spruce St.  
Philadelphia, PA 19104

**Murray Grossman (mgrossma@mail.med.upenn.edu)**

Department of Neurology; University of Pennsylvania Medical Center, 3400 Spruce St.  
Philadelphia, PA 19104

This study continues previous work investigating the neural basis for grammatical and short-term memory components of sentence comprehension. We hypothesized that greater demands placed upon a third component, information processing speed in the form of a faster stimulus presentation rate, is shouldered by the caudate nucleus. To test this hypothesis, neural activity was measured with fMRI in 15 healthy, right-handed English-speakers as they determined the agent of the action in sentences with subject-relative (SR) or object-relative (OR) center-embedded clauses that included "short" (three-word) and "long" (seven-word) spans between the antecedent noun phrase (NP) and the "gap" where the NP is interpreted. These 13-word sentences were presented visually in a word-by-word fashion at 500 msec/word and 750 msec/word, the 500-msec rate requiring rapid information processing but the 750-msec rate requiring greater short-term memory.

Subjects were imaged on a 1.5T GE Echospeed scanner, and gradient echo echoplanar images were obtained to detect alterations in blood oxygenation (5mm slice thickness, effective TE 50msec, 64 x 40 matrix, voxel size of 3.75 x 3.75 x 5mm). The images were registered, aligned to Talairach space, smoothed with 8mm and 2.8sec Gaussian

kernels, and analyzed with SPM96 using appropriate protection for multiple comparisons.

Wernicke's area supported sentence comprehension of the four main sentence types (SR short, SR long, OR short, OR long), duplicating results from an earlier 7-subject study. In this study, left middle frontal cortex and left caudate were additionally recruited during OR sentences. Broca's area was additionally recruited during OR-long sentences, again confirming earlier results.

When words were presented at a rapid rate and short-term memory demand was minimized in the short antecedent-gap sentences, Wernicke's area and caudate were recruited. This was true for both SR and OR sentences. Long antecedent-gap sentences presented slowly to maximize short-term memory load revealed Wernicke's area but no caudate activity. This too was true for both SR and OR sentences. These observations support the hypothesis that the caudate contributes to sentence processing in the form of negotiating rapid information processing, not short-term memory demand, during sentence comprehension.

# Experimentally Uncovering Hidden Strata in English Phonology

Lisa Davidson (davidson@cogsci.jhu.edu)

Department of Cognitive Science, Johns Hopkins University  
3400 N. Charles St, Baltimore, MD 21218

## Introduction

The final state of the phonological grammar of a language *L* permits those structures that are legal in *L* and disallows those which are not. With respect to phonotactics, a possible assumption is that all consonant clusters prohibited by the phonology of *L* should be “equally illegal” for a speaker of *L*. However, several studies of the second language (L2) acquisition of consonant clusters have shown that not all clusters illegal in a speaker’s native language are equally difficult for learners acquiring a language with a different cluster inventory than their own (i.e. Broselow & Finer, 1991; Eckman & Iverson, 1993). Assuming that beginning L2 learners or naive speakers faced with foreign words use their native language grammars to produce these words, the graded performance on different consonant clusters sheds light on the nature of the final state of the native grammar. The present study uses an Optimality Theoretic (OT) approach (Prince & Smolensky, 1993) to examine the final state of a native speaker’s grammar and to account for this graded performance on clusters. New markedness constraints presented here that pertain to several different types of clusters are not only useful for characterizing speakers’ performance, but also for explaining consonant cluster typology more generally.

## Experiments 1 and 2

**Methods.** Previous research has explored minimal sonority distance (MSD) as a major factor affecting the differential acquisition of consonant clusters (Broselow and Finer, 1991; Eckman and Iverson, 1993; Hancin-Bhatt and Bhatt, 1998). Because English has an MSD of 1, it can be hypothesized that English speakers will have difficulty with clusters with an MSD of 0. Instead of using L2 learners, the present experiment examines the productions of native English speakers on Polish initial consonant clusters that have a sonority distance of 0 (SD 0) (/kt/,/kp/,/pt/,/č k/,/vz/), non-English clusters with an SD of 1 or more (SD 1) (/dv/,/vn/,/tʃ/,/zr/,/zm/) and English clusters with an SD of 1 or more (SD 1E) (/ʃr/,/ʃl/,/sm/,/sn/,/fr/). In experiment 1, subjects heard a pseudo-Polish word with one of the previous 15 clusters produced by a native speaker of Polish and then were told to read and memorize an English sentence containing a written version of the foreign word. The sentence then disappeared and the subjects repeated the sentence aloud. The subjects were explicitly instructed to pronounce the foreign words as they would if they were English words.

**Results.** A spectrogram of each target was examined to determine the response. An ANOVA showed that while subjects were significantly better on the English possible clusters than impossible clusters ( $p < .0001$ ), the distinction

between non-English clusters was not significant (SD 0: 33% correct, SD 1: 42% correct,  $p < .30$ ). It was concluded that SD cannot be the most important factor in determining the difficulty of clusters. A breakdown of subjects’ performance on individual clusters instead suggested the following groupings (from worst to best performance): A: /vn/,/vz/,/dv/ > B: /kt/,/kp/,/pt/,/č k/,/tʃ/ > C: /zm/,/zr/ > D: /ʃr/,/ʃl/,/sm/,/sn/,/fr/. Post-hoc comparisons indicated that each group was significantly different from every other.

Using a modification of the original procedure intended to minimize memory demands, experiment 2 successfully replicated experiment 1 (with the difference between groups B and C marginally significant,  $p < .12$ ).

## Discussion

An English phonological grammar that makes distinctions between different types of illegal clusters can be captured naturally within OT. In order to achieve graded performance, different markedness constraints must target each of the groups (A-D) of illegal clusters above. These rankings produce hidden strata that distinguish between increasing “foreignness” in nonnative words, much like the strata that correspond to decreasing nativization in loanwords (Itô & Mester, 1999). Native speakers are able to exploit the hidden rankings among these constraints by promoting faithfulness constraints to different points in the hierarchy when producing illegal clusters. Assuming that this movement is performed anew for every target, a speaker can exhibit performance that is not simply 0% or 100% on any given set of clusters. However, the fixed ranking of markedness constraints predicts that if a speaker can produce a more marked group of clusters (such as B) her performance on the less marked group (such as C) must be at least as good as or better. The relevant markedness constraints demonstrate that a universal, cross-linguistic account of consonant clusters must integrate sonority-like constraints with co-occurrence restrictions that employ non-sonority aspects such as place and cluster position.

## References

- Broselow, E. & Finer, D. (1991). Parameter setter in second language phonology and syntax. *Studies in Second Language Acquisition*, 7:1, 35-59.
- Eckman, F. & Iverson, G. (1993). Sonority and markedness among onset clusters in the interlanguage of ESL learners. *Second Language Research* 9, 234-252.
- Hancin-Bhatt, B. & Bhatt, R. (1998). Optimal L2 syllables. *Studies in Second Language Acquisition*, 19, 331-378.
- Itô, J. & Mester, A. (1999). The phonological lexicon. In T. Fujimura (Ed.), *Handbook of Japanese Linguistics*. Oxford: Blackwell.
- Prince, A. & Smolensky, P. (1993). *Optimality Theory: Constraint interaction in generative grammar*. To appear, MIT Press.

# Language after Hemispherectomy: Effects of Seizure Control

**Stella de Bode (sdebode@ucla.edu)**

UCLA, Department of Linguistics; 405 Hilgard Ave  
Los Angeles, CA 90095

**Susan Curtiss (scurtiss@ucla.edu)**

UCLA, Department of Linguistics; 405 Hilgard Ave  
Los Angeles, CA 90095

**Gary W. Mathern (gmathern@ucla.edu)**

Division of Neurosurgery  
UCLA Medical Center, Los Angeles, CA 90095

## Introduction and Rationale

There is no question that on-going seizure activity in children has a debilitating effect on all aspects of cognitive development, including language (Dulac et al., 1996; OLeary et al., 1983; Rossi et al., 1996). Comparing the effect of left- versus right-temporal lobe origin seizures, Cohen (1992) reported the expected correlation between side and auditory/verbal vs visual/spatial memory in children with complex partial epilepsy. Furthermore, specific aspects of linguistic performance were shown to be differentially affected in children with simple-partial left hemisphere epilepsy (Cohen & Le Normand, 1998). Linguistic comprehension tested in this study gradually improved to reach normal performance levels while production remained quite poor in comparison with controls. A general conclusion drawn from these and other studies is that cognitive development in children with epilepsy is severely compromised in comparison to neurologically intact children.

It is thus surprising to find that the effect of seizure control in post-surgical patients and its relevance to improved cognitive and linguistic functioning remains an area of great controversy. A few studies report that seizure-free patients perform no better cognitively than their counterparts who continue to have seizures after surgery (Grande et al., 1997; Seidel et al., 1997). However, post-surgical seizure control and linguistic outcome have not been specifically explored to our knowledge, though some reports indicate that this relation is far from linear (Vargha-Khadem & Mishkin, 1997). It seems that seizure control by itself does not guarantee improved linguistic functioning in hemispherectomy patients. However, patients who become seizure-free after surgery performed early in life are asserted to demonstrate a better linguistic prognosis.

This is a preliminary report of a pilot study examining the validity of the latter claim and adding to the quantitative investigation of the effects of seizure control on language outcome by evaluating spoken language outcome as a function of seizure control,  
age of seizure onset, and  
seizure duration

in a large population of pediatric hemispherectomies.

## Methods

Subjects consisted of 42 patients who underwent hemispherectomy for intractable seizures at UCLA Medical Center. Age at onset of seizures: 0;0 – 11;0 years, age at surgery: 0;3 – 17;3 years, seizure duration: 0;3 – 14;1 years, post-surgical evaluation: no less than 5 years. Postoperative spoken language outcome was rated on the basis of free language samples from 0 – no language to 6 – fluent mature speaker.

## Results and Discussion

Based on the analysis of the entire population the following results were obtained: 1) age at seizure onset positively correlated with language outcome ( $p > 0.013$ ); 2) postsurgical seizure control positively or negatively correlated with language outcome in a statistically significant way ( $p > 0.0082$ ). Seizure duration did not reach statistical significance.

Although it is clear from this preliminary analysis that clinical variables related to seizure activity are relevant in predicting post-surgical cognitive outcome in hemispherectomy patients our major concern and intuition was that for many children in our sample the respective prediction would not prove accurate. Indeed, we have found that some children go on to develop language despite on-going seizures while other patients *with* seizure control remain without any language. We thus hypothesized that 1) post-surgical seizure control is just one measure of the integrity of the remaining hemisphere and cannot be approached without accounting for specific etiologies. As a result of this prediction variables of language outcome and seizure control are currently analyzed separately for the three major etiologies (cortical dysplasia, Rasmussen's encephalitis and infarct). 2) Different language outcomes in the patients without seizure control can be explained by differentiating between those seizures that result from a structural lesion in the remaining, presumably "healthy" hemisphere and seizures that result from functional damage sustained as a result of pre-surgical seizures/abnormal functioning from the removed hemisphere. To date our hypotheses have been confirmed.

# Compositional Functions in Nominal Combination

Zachary Estes (zcestes@princeton.edu)  
 Princeton University  
 Department of Psychology; Green Hall  
 Princeton, NJ 08544-1010 USA

Compositional functions are the cognitive processes by which two independent units of meaning are understood as a single compound meaning. For instance, some nominal combinations are interpreted attributively (e.g., ‘sponge memory’ as “a good, *absorptive* memory”), while others are interpreted relationally (e.g., ‘rodeo magazine’ as “a magazine *about* rodeos”). One question of recent interest is whether attributive combination is cognitively distinct from relational combination, or whether attribution is simply a *resembles* relation (e.g., ‘sponge memory’ as “a memory that *resembles* a sponge in some way”). If the two combination-types respond differently to the same manipulation, then one can infer that they are in fact distinct processes.

Attributive and relational combinations were preceded by prime combinations that either did or did not share the same attribution/relation in a sense-nonsense judgment task. In Experiment 1, the prime also shared either the modifier or the head concept with the target. In Experiment 2 there was no lexical overlap between prime and target. Experiment 3 tested whether priming of attributive combination was purely associative. All experiments also included an uninformative baseline prime. See Table 1.

Table 1: Sample stimuli, Experiments 1, 2 and 3.

Prime-type		
Experiment 1		
M-Consistent:	sponge towel	rodeo documentary
M-Inconsistent:	sponge nurse	rodeo clown
H-Consistent:	warehouse memory	motorcycle magazine
H-Inconsistent:	childhood memory	library magazine
Experiment 2		
Consistent:	warehouse mind	motorcycle documentary
M-Control:	warehouse guard	motorcycle gang
H-Control:	gutter mind	epic documentary
Experiment 3		
Consistent:	warehouse brain	
Reversed:	brain warehouse	
Inconsistent:	seesaw relationship	
Target:	sponge memory (attributive)	rodeo magazine (relational)

When the prime combination used the same attribution/relation and one of the same constituents as the target combination, then comprehension of that target was facilitated (see Figure 1). When there was no lexical overlap between prime and target, only attributive combination was

facilitated (see Figure 2). However, this facilitation for attributive combination was due to associative priming and did not generalize to other attributions (see Figure 3). Thus, although attributive combination may be more susceptible to associative priming than relational combination, the two compositional functions behaved similarly.

Figure 1: Priming of response times, Experiment 1.

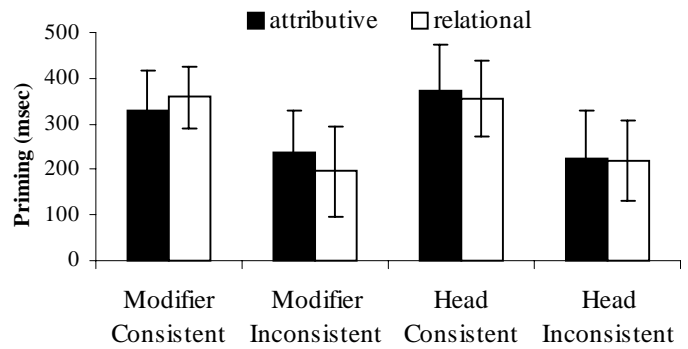


Figure 2: Priming of response times, Experiment 2.

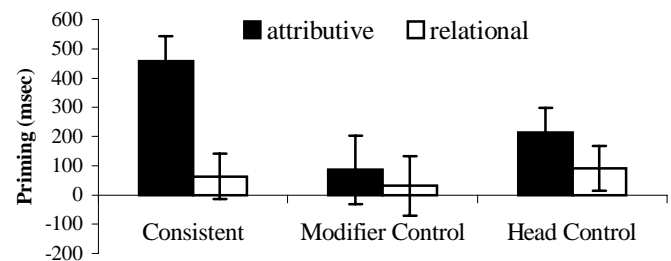
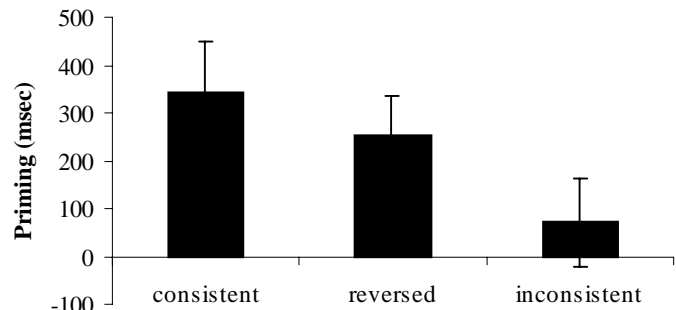


Figure 3: Priming of response times, Experiment 3.



## Acknowledgments

This research was supported by the National Science Foundation. I thank Sam Glucksberg, Yevgeniya Goldvarg, James Hampton, Matt McGlone and Mary Newsome for helpful comments on this research.



# Eye Movements in Human Face Learning and Recognition

**Richard J. Falk (richard@eyelab.msu.edu)**  
**Andrew Hollingworth (andrew@eyelab.msu.edu)**  
**John M. Henderson (john@eyelab.msu.edu)**

Department of Psychology, and Cognitive Science Program; Michigan State University  
129 Psychology Research Building, East Lansing, MI 48824

**Sridhar Mahadevan (mahadeva@cse.msu.edu)**

Department of Computer Science and Engineering, and Cognitive Science Program; Michigan State University  
2325A Engineering Building, East Lansing, MI 48824

**Fred C. Dyer (fcdyer@msu.edu)**

Department of Zoology, and Cognitive Science Program; Michigan State University  
403 Natural Science Building, East Lansing, MI 48824

Any theory of face recognition must specify what is encoded in order for a face to be recognized at a later time. Theories of face recognition tend to highlight the importance of either individual feature encoding or holistic processing (see Valentine, 1988, for review). However, very little information is available about where exactly people look when processing a face. The current study examined the nature of eye movements in the learning and recognition of human faces. The goals of the study were twofold: (a) to determine where people look when learning and recognizing faces, and (b) to determine if fixation patterns change as a function of face inversion.

Sixteen participants studied 20 color photographs of faces for 10 seconds each in preparation for a recognition memory test. In the test phase all 20 previously viewed faces (familiar) and 20 novel faces (unfamiliar) were presented in pseudo-random order until the participant responded (mean response time = 2397 ms). Half of the familiar faces and half of the novel faces were presented in the upright orientation. The remaining faces were presented in the inverted orientation. Eye movements were recorded during both the study and test phases using a dual-Purkinje image eyetracker.

Mean percent correct was lower for inverted (66%) than for upright faces (79%),  $p < .05$ , suggesting that the participants were engaged in a representative face processing task.

During the study phase, 56% of total viewing time was spent fixating on the eyes, 18% on the nose, 12% on the mouth, and the remaining 13% on the rest of the face (ears, chin, cheeks, and forehead). Thus, fixating on the eyes is an important part of the face encoding process. Because the amount of total viewing time differed from the study to the test phase, viewing time on specific regions was compared as proportions of total viewing time. Overall, as Figure 1 shows, the proportion of total fixation time spent on facial features for the study and the test sessions was very similar. However, there were reliable differences in the proportion of total viewing time for the mouth and the ear features,  $p < .05$ . These findings suggest that similar features are chosen for analysis during both face learning and recognition. Figure 1 also shows that the familiarity and inversion manipulations produced very little change in the proportion

of time devoted to these selected features with only the mouth showing a reliable difference,  $p < .05$ .

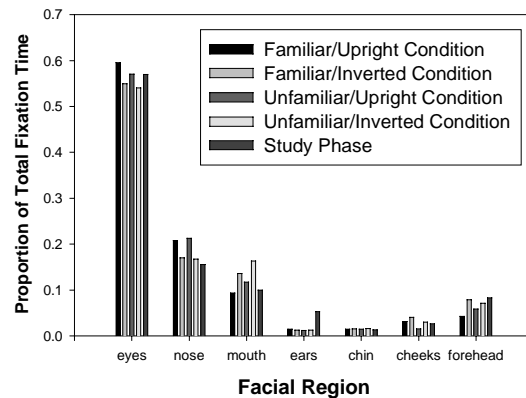


Figure 1: Proportion of total time in facial region for study condition and four test conditions.

If holistic processing is more likely with upright faces than inverted faces, we might expect less sampling of individual facial features in the former condition compared to the latter. Instead, the proportion of fixations on which the eyes moved from one facial region to a new region was not reliably different as a function of orientation (.84 upright, .81 inverted,  $F < 1$ ). Therefore, the decrement in recognition performance due to inversion does not appear to be a consequence of the differential sampling of facial features.

Overall, the results indicate: (a) that similar facial features are selected for analysis during face learning and recognition, and (b) that there is very little difference between the fixation patterns for upright and inverted faces.

## Acknowledgements

This research is supported in part by a Knowledge and Distributed Intelligence (KDI) grant from the National Science Foundation (ECS-9873531).

## References

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, 79, 471-491.

# Comprehension of Active and Passive Sentences in Portuguese and English: The Prototypicality Effect

Rosângela Gabriel (rosangela.gabriel@psy.ox.ac.uk)

Department of Experimental Psychology    Pós-Graduação em Letras  
University of Oxford    PUCRS  
South Parks Rd    Porto Alegre - RS – Brasil  
Oxford OX1 3UD    Cep 90619-900

Kim Plunkett (kim.plunkett@psy.ox.ac.uk)

Department of Experimental Psychology  
University of Oxford  
South Parks Rd  
Oxford OX1 3UD

## Abstract

There have been several investigations into the acquisition of passive constructions, most based on empirical data from English children. These have thrown up a variety of theories regarding the nature of the strategies underlying passive acquisition. However, is it reasonable to assume that in different languages children use the same strategies to learn the passives? Will passives in all languages demand the same cognitive skills from the language learner? Or, in each language will children show a different pattern of development?

Slobin (1981) states that in each type of language, children initially isolate and generalise basic sentence forms. According to Slobin, prototypical events and canonical sentence forms constitute a nucleus for the growth of language. For the purpose of the research reported here, it is important to specify the meaning of *basic sentence form*, *prototypical event* and *canonical sentence*. The former combines structural and typological characteristics, and will show some variation depending on the number of constituents requested by the verb as well as the frequency of a given structure in a language. The second is defined in conceptual terms, following Hopper & Thompson's (1980) Transitivity Hypothesis. A more prototypical transitive event will present two or more participants, an action, an actor high in potency, and an affected non-actor. Lastly, a canonical passive sentence resembles Givón's (1990) 'promotional passive' or Maratsos *et al.*'s (1985) 'typical passive', embodying three important features. First, a non-agent will be the pragmatic topic of the sentence, placed in the syntactic subject position. Second, the semantic agent will optionally appear in a special oblique case. Finally, an actional verb will be coded in a more stative form (be/get + past participle). We hypothesised that order of acquisition of sentences structures will follow Slobin's (1981) prediction: first children will acquire the more prototypical and basic sentence forms and only later will children be able to generalise to less prototypical sentence forms.

In order to investigate this hypothesis, two studies were designed, testing comprehension in four different types of active and passive sentences: more prototypical transitive scenes (irreversible, reversible) and less prototypical transitive scenes (dative and locative). Subjects from Study 1 are all monolingual English speakers residing in UK whereas subjects from Study 2 are all monolingual Portuguese speakers living in the south of Brazil. In both studies participants were children (aged 3 to 10) and adults. The results show a prototypicality effect on the acquisition of passives in both languages. More prototypical passive sentences as well as more prototypical actives were understood at younger ages than less prototypical sentences. This cross-linguistic similarity might indicate that the process of pattern formation is an important cognitive strategy for the language learner.

## References

- Givón, Talmy. *Syntax: a functional-typological introduction* - vol. II. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1990. p.563-644.
- Gordon, P. & Chafetz, J. (1990). Verb-based versus class-based accounts of actionality effects in children's comprehension of passives. *Cognition*, 36, 227-254.
- Hopper, P. J. & Thompson, S. A. Transitivity in grammar and discourse. *Language*, v. 56, n. 2, 1980, p. 251-299.
- Maratsos, M. P., Fox, D. E. C., Becker, J. A. & Chalkley, M. A. (1985). Semantic restrictions on children's passives. *Cognition*, 19, 167-191.
- Slobin, D. I. The origins of grammatical encoding of events. In: Deutsch, Werner. *The child's construction of language*. Academic Press, 1981, p. 187-199.

# Why Are Some Problems Easy? New Insights into the Tower of Hanoi

Glenn Gunzelmann (glenn@andrew.cmu.edu)  
Department of Psychology, Carnegie Mellon University  
Pittsburgh, PA 15213

Stephen Blessing (blessing@carnegielearning.com)  
Carnegie Learning, 372 N. Craig St., Suite 101  
Pittsburgh, PA 15213

## Introduction

Researchers have found large differences in difficulty and varying amounts of transfer among isomorphs of the Tower of Hanoi (Kotovsky, Hayes, & Simon, 1985; Hayes & Simon, 1977). Because the tasks have the same formal structure, these differences must result from the surface representations. To explain these findings, Kotovsky, et. al. pointed toward the ability to relate the rules to real-world knowledge and representation influence such as the externalization of rules (rules embedded in the external problem representation; also see Zhang, 1997).

Despite this research, many questions remain about the processes underlying problem solving and transfer of learning. This experiment uses standardized presentations of the isomorphs and presents more problems per participant than in past experiments. These manipulations should enhance transfer and help clarify findings that involve differential difficulty.

## Method

Participants were presented with 12 problems for each of three isomorphs of the Tower of Hanoi (the Standard Tower of Hanoi, Monster Move, and Paint Stripping; order of isomorphs was varied across participants) and two filler tasks. For each task, participants were presented with a description, a set of rules, and an explanation of the interface before beginning. They were instructed to solve each problem by reaching the goal presented on the screen. After solving all of the problems, participants were asked questions to determine how noticeable the relationships among the isomorphs were.

## Results and Discussion

The verbal reports were used to help determine what information may have transferred from the source isomorph to the target. While some participants claimed to notice a similarity, only 2 (of 37) were able to accurately describe it. Despite this lack of awareness, transfer of learning was clearly shown. Time to solve decreased across isomorph position,  $p < .01$  (Figure 1). Also, any of the isomorphs was sufficient to produce transfer. In addition, the degree of transfer was much greater than has been found previously, owing to the standardized interface as well as increased practice. Performance on the Tower of Hanoi was not facilitated by previous exposure to another isomorph (likely due to a floor effect; see Figure 1). These findings, combined with the lack of awareness about the similarities, suggest that more general procedural

knowledge (execution of general strategies) is largely responsible for the transfer. They also suggest that the Tower of Hanoi was relatively easy for participants to solve.

The comparison of isomorphs showed that the Monster Move isomorph was most difficult, followed by the Paint Stripping isomorph, with the standard Tower of Hanoi being the easiest,  $p < .001$ . Representational influences seem to drive this effect, with the rules for the Tower of Hanoi being largely inferable from the presentation. In contrast, all of the Monster Move rules need to be learned explicitly, while at least some of the Paint Stripping rules are not intuitive based solely on the presentation. These results suggest that the incorporation of problem constraints (rules) into the problem representation can reduce problem difficulty by reducing cognitive load. These results can be generalized beyond the simple problems used here, and suggest simple ways of achieving improved performance in virtually any task domain.

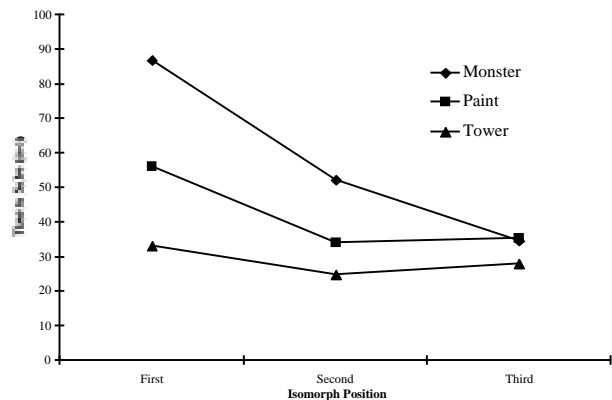


Figure 1. Average time (sec.) to solve problems for each of the isomorphs for each isomorph position.

## References

- Hayes, J. R., & Simon, H. A. (1977). Psychological differences among problem isomorphs. In N. J. Castellan, D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory: II* (pp. 21-41). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? Evidence from the Tower of Hanoi. *Cognitive Psychology*, *17*, 248-294.
- Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science*, *21*, 179-217.

# Stages of Phonological Processing in Spoken Production

Matthew Goldrick (goldrick@jhu.edu)  
Brenda Rapp (brenda@mail.cog.jhu.edu)  
Paul Smolensky (paul@mail.cog.jhu.edu)

Department of Cognitive Science, Johns Hopkins University, Baltimore MD 21218 USA

Processing theories of spoken word production have generally drawn a distinction between two general stages of phonological processing. The first stage (“lexical processing”) involves retrieval of stored phonological information specific to a particular lexical item. The second stage (“post-lexical processing”) transforms this information into a form which can be used to access subsequent processes for speech execution.

Investigations into impairments of these stages of processing (e.g. Kohn & Smith, 1994) have often proceeded by making representational and processing assumptions about each stage and using these assumptions to predict patterns of impaired performance following damage to each stage. This approach has not been unproblematic: there is widespread disagreement as to the nature of processing at each stage, preventing consistent interpretation of empirical results.

Here we report on results obtained using a different approach. We propose that picture naming tasks require the use of lexical phonological processing, while repetition tasks may bypass this stage by making use of non-lexical acoustic-to-phonological conversion procedures. This proposal predicts that damage to lexical processing will affect picture naming, as the task must access the lexical process, but it will not affect repetition, as this task can bypass the impaired process. In contrast, damage to post-lexical processing will impair both tasks equally, as both tasks must make use of this process. This method will allow us to more closely examine the characteristics of lexical and post-lexical phonological processing.

## Determination of Deficit Locus

BON is a 62 year old right handed woman who suffered a left hemisphere stroke affecting the superior posterior frontal and lateral parietal regions. CSS is a 62 year old right handed man who suffered a stroke affecting the parietal regions of the left hemisphere and the right basal ganglia. Their comprehension and articulation were normal.

Each subject was administered a common set of stimulus items in repetition and naming. Consistent with a post-lexical deficit, BON exhibited impaired performance in both naming (79% accuracy) and repetition (84%). The difference in performance was not significant. Consistent with a deficit to lexical phonological processing, CSS exhibited close to normal performance in repetition (96% accuracy), while being impaired in picture naming (87%). The accuracy difference between these tasks is significant.

All subsequent analyses examined performance on a larger set of items: picture naming performance for CSS (n=1680); and performance on all spoken output tasks for BON (n=851).

## Phonological Analyses

BON was significantly worse on low frequency (9.8% error) versus high frequency phonemes (5.6%); on dorsal segments (20.7%) versus coronal segments (8.5%); and on coda segments (8.7%) versus onset segments (2.5%). (The coda versus onset effect could not be attributed to segment frequency or place of articulation effects.) CSS exhibited none of these effects. This contrasting pattern of performance supports the approach and suggests that the post-lexical process is especially sensitive to the characteristics of phonemes. Subsequent analyses of her performance suggest the post-lexical process respects grammatical constraints and language-particular frequency.

## Lexical Analyses

Compared to BON, significantly fewer of CSS’s errors result in nonwords (57% of his errors are nonwords, compared to 70% for BON). This suggests the lexical process (unlike the post-lexical process) is biased to produce lexical items. Analysis of CSS’s whole word substitution errors suggests a definition of phonological neighborhood which, unlike other measures of density (e.g. Luce & Pisoni, 1998), does not require that neighbors share the target’s phonemes in the same position. Analyses including this density measure suggest effects of frequency, length and neighborhood density on his performance.

## Conclusion

Using task differences to determine the loci of phonological deficits is a fruitful approach. The results suggest that lexical processing involves lexically-guided retrieval of stored segmental information; this information is then used by the post-lexical process to construct a fully specified phonological representation.

## References

- Kohn, S. E. & Smith, K. L. (1994). Distinctions between two phonological output deficits. *Applied Psycholinguistics*, *15*, 75-95.
- Luce, P. A. & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, *19*, 1-36.

# Of Words, Birds, Worms, and Weeds: Infant Word Learning and Lexical Neighborhoods.

**George J. Hollich (ghollich@yahoo.com)**

Department of Psychology  
Ames Hall; Johns Hopkins University  
Baltimore, MD 21218 USA

**Peter W. Jusczyk (jusczyk@jhu.edu)**

Department of Psychology  
Ames Hall; Johns Hopkins University  
Baltimore, MD 21218 USA

**Paul A. Luce (paul@deuro.fss.buffalo.edu)**

Department of Psychology  
University at Buffalo  
Buffalo, NY 14260 USA

How specific are infants' representations of words? Do words that sound similar to each other present any special difficulties, or benefits, in early lexical acquisition? That is, experience in encoding certain kinds of phonotactic sequences and metrical patterns could facilitate the acquisition of new word-to-world mappings (Jusczyk, 1997). Alternatively, competition from existing lexical items that share similar phonotactic and phonetic properties could also inhibit children's ability to encode a new item (see Luce & Pisoni, 1998; Marslen-Wilson, 1989; McClelland & Elman 1986; Norris, 1994). Thus, for example, children who know the word, "hat," could conceivably learn the word, "had," more quickly than a phonetically unrelated word because their experience with the "ha-" sound structure makes forming an acoustic package easier. On the other hand, competition from the "hat" representation, could make "had" very difficult to learn and inherently confusable with "hat."

Two studies are reported that examine infants' abilities both to detect the similarity among such "lexical neighbors," words that differ by a single phoneme, and to learn a referent for a novel neighbor after an exposure to a high number of these similar sounding words. In all studies, the lexical neighbors were constructed of CVC non-words that differed in the initial consonant, the vowel, or the final consonant of a prototype. All lists were controlled for word phonotactics, frequency, and their relation to English lexical neighborhoods.

In study 1, 15-month-old infants exhibited a novelty preference for a neighborhood prototype, after being familiarized in the head turn preference procedure with twelve lists of twelve neighbors. The mean looking time in seconds, with the standard error in parentheses, to the novel and prototypical words was 7.95 (0.52) and 6.70 (0.68), respectively. This suggests that, even by 15 months, infants are capable of detecting the neighborhood similarity among words.

In study 2, 17-month-olds were tested on their ability to learn the referent of two novel prototypes after being exposed to their respective lexical neighbors. In one condition, the high-density condition, six lists of twelve neighbors were

used. The low-density condition utilized six lists of three neighbors plus nine filler items. Results obtained with the intermodal preferential looking procedure indicated that word learning was significantly better in the low density condition, both in overall looking times and in infant reaction times to the targeted word. The mean difference in looking times between the target and non-target in the high- and low-density conditions was  $-0.14$  (0.14), and  $0.59$  (0.21) seconds, respectively.

Taken together, these results fit well with current models of spoken language recognition, many of which suggest a competitive effect for words arising from dense lexical neighborhoods. However, preliminary results from a control study seem to indicate that some exposure to a neighborhood may be better than no exposure at all. Thus, 17-month-old infants that were tested on their word learning ability after being exposed to twelve lists of only filler items performed worse than those from the low-density condition reported above did. This suggests that some exposure to lexical neighborhoods might facilitate and strengthen infants' ability to form a representation of the new word, while too much exposure might fatigue the system and/or introduce strong competitive effects.

## References

- Jusczyk, P. (1997). *The discovery of spoken language*. Cambridge, MA: The MIT Press.
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing, 19*, 1-36.
- Marslen-Wilson, W. D. (1989). Access and integration: Projecting sound onto meaning. In W. D. Marslen-Wilson (Ed.), *Lexical access and representation*. Cambridge, MA: Bradford.
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1-86.
- Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 24*, 1495-1520.

# Modelling Language Acquisition at Multiple Temporal Scales

Steve R. Howell (showell@hypatia.psychology.mcmaster.ca)

Department of Psychology, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada

Suzanna Becker (becker@mcmaster.ca)

Department of Psychology, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada

The problem of incorporating time in a neural network is an important one. Networks with feedback, such as Simple Recurrent Networks (SRNs) (Elman, 1990) have been argued to represent time more realistically through its effects on the processing of input, compared to standard feedforward networks. In effect, SRNs' context units act as a memory, which incorporates a "smeared-out" representation of the network's internal states over time.

A problem does exist with this representation, however, especially for complex domains like that of language. The nature of argument agreement, embeddings and similar phenomena means that the SRN must be able to represent important past states (such as head noun for verb agreement) in spite of the declining effects of past context. While most word inputs will be related most strongly to words co-occurring close by in the input stream, verb agreement, for example, is largely determined by its corresponding noun, even in long, multiply-embedded sentences. SRN models should therefore be able to preserve representations of vital early structure for later use, in spite of the generally appropriate decline of short-term context. This issue has been addressed in an architectural fashion by others (Weckerly and Elman, 1992), but can perhaps be addressed more generally by allowing for more than one duration of context in an SRN's operation.

It is possible to apply the concept of hysteresis to the SRN's context units. That is, the update from the hidden units to the context units may be other than the usual 1-to-1 copying; the context units may also incorporate self-recurrent connections of varying strengths. In particular, we have been experimenting with SRNs using the hysteresis function suggested by Wermter, Arevian, and Panchev (1999) on the self-recurrent connections:

$$\text{Context}_i(t+1) = (1-\text{Hy}) * \text{Hidden}_i(t) + \text{Hy} * \text{Context}_i(t)$$

We have conducted initial experiments using a test corpus derived from the original simplified test corpus used by Elman (1990). Our version differs from the original in that it includes not only consonant to vowel relations, but also word-to-word relations. That is, some of the consonant-vowel combinations (words) can only occur immediately following others.

Thus in addition to the network needing to learn, for example, that u's only come after G's or U's (Guuu), it must also learn that Guuu only comes after Da. It is in this capacity that the hysteresis parameter should most come into play, for it specifies, in effect, the duration of retention

of the states of the context units. For short term letter to letter relations, small to zero hysteresis values should be adequate, as demonstrated originally by Elman's success. In that experiment, the network error declined consistently within a word, but jumped at word boundaries, representing the fact that word distribution was random in that corpus. In our experiments, manipulating the hysteresis parameters was expected to bias the network in favour of either short or long term relationships. Also, simulated annealing of the learning rate, another technique not typically used with SRNs, is used in both control and experimental networks. In pilot work this feature smoothed oscillations in the gradient descent of error.

The initial results of a number of simulation runs from different random initial conditions indicate that small hysteresis values (of other than 0) are indeed an advantage in learning this prediction task, with error per epoch declining noticeably, though not exceptionally, faster with  $0.2 > \text{Hys} > 0.1$ . Presumably this modest net gain is actually composed of both a larger gain for word-to-word relationships and a small decline for letter-to-letter prediction. Explorations of the exact nature of this advantage are underway, as is investigation of the best range of hysteresis parameters for various language tasks.

With the ability to change the hysteresis of context layers, it becomes useful to incorporate multiple hidden layers into an SRN (Wermter, 1999), with layers having a different 'span' of context via different hysteresis settings. We also describe a model with multiple hidden layers that is being applied to more complex language corpora, and is designed to be able to learn at multiple time scales simultaneously, by capturing longer-range temporal structure in progressively higher layers.

## References

- Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.
- Weckerly, J., & Elman, J.L. (1992). A PDP approach to processing center-embedded sentences. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Wermter, S., Arevian, G. & Panchev, C. (1999). Recurrent Neural Network Learning for Text Routing, *Proceedings of the Ninth International Conference on Artificial Neural Networks*, 2, 470-475.

# Perceptual and Experience-Dependent Influences on Location Memory Processes

Alycia M. Hund

(alycia-hund@uiowa.edu)

Department of Psychology, University of Iowa  
E11 Seashore Hall, Iowa City, IA 52242

John P. Spencer

(john-spencer@uiowa.edu)

Department of Psychology, University of Iowa  
E11 Seashore Hall, Iowa City, IA 52242

## Abstract

This experiment investigated how people use perceptual (i.e., axes of symmetry) and experience-dependent information (i.e., remembered possible target locations) to maintain location information in memory over short-term delays. Participants pointed to target locations on a tabletop following variable delays. Analyses of directional error indicated that location memory is repelled from axes of symmetry and attracted toward remembered possible target locations.

## Introduction

Previous experimental findings suggest that two factors produce systematic biases in location memory. First, remembered locations are repelled from axes of symmetry (McNamara, Hardy, & Hirtle, 1989), suggesting that people use symmetry axes to organize space into categories. Second, remembered locations are attracted toward an average or prototypical location within each category (e.g., Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher, Newcombe, & Sandberg, 1994), suggesting that people use a longer-term memory of the *possible* target locations to facilitate memory. Generally, these two factors have been studied using tasks that require participants to remember many locations within task spaces with several symmetry axes. Thus, the factors that underlie memory biases often are confounded, making it difficult to determine whether biases are caused by two competing processes or one process.

Recently, Spencer and Hund (2000) used a simplified task to investigate how location memory biases change over short-term delays. Participants pointed to three remembered locations within a large, homogeneous task space following variable delays. Results indicated that responses were repelled from a midline axis of symmetry. Furthermore, these repulsion effects increased in magnitude over delay.

Here we extend these results to examine whether both factors—repulsion from symmetry axes and attraction toward remembered possible locations—influence how people maintain location information in memory. More specifically, we separated the possible target locations from the midline symmetry axis to determine whether location memory biases result from one or two memory processes.

## Method

Sixty right-handed adults participated. On each trial, 1 of 3 possible target locations appeared on a large tabletop. Following a variable delay, participants pointed to the remembered location. They received accuracy and timing feedback after each trial.

Targets were presented in different layouts relative to the midline axis of the task space (e.g.,  $-60^\circ$ ,  $-40^\circ$ ,  $-20^\circ$  v.  $20^\circ$ ,  $40^\circ$ ,  $60^\circ$ ) such that the mean of the possible target locations was not at midline. In addition, three bias conditions (no bias, bias left, bias right) were included to examine experience-dependent memory effects. Trials were divided evenly among the 3 possible targets in the no bias condition. In the bias conditions, 2/3 of all trials were to a biased target (left or right) and 1/3 of the trials were equally divided between the two remaining targets.

## Results

As reported in Spencer and Hund (2000), directional responses to all targets were biased away from midline. These repulsion effects increased systematically over delays. In addition to midline repulsion effects, participants' responses were biased away from the  $45^\circ$  diagonal symmetry axes when the targets were centered near these axes. Finally, biasing either the left or right target shifted participants' responses leftward or rightward, suggesting that they used a longer-term memory of possible target locations to remember the target location on each trial.

## Discussion

Data demonstrate that location memory biases result from two memory processes. Over short-term delays, memory is repelled from perceived axes of symmetry and attracted toward remembered possible target locations. In addition, data indicate that adults can select particular reference axes to facilitate memory. Future studies are needed to clarify the factors the influence reference axis selection and how experience-dependent effects are built-up over learning.

## References

- Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial locations. *Psychological Review*, *98*, 352-376.
- Huttenlocher, J., Newcombe, N., & Sandberg, E. H. (1994). The coding of spatial location in young children. *Cognitive Psychology*, *27*, 115-147.
- McNamara, T. P., Hardy, J. K., & Hirtle, S.C. (1989). Subjective hierarchies in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 211-227.
- Spencer, J. P., & Hund, A. M. (2000). Location memory biases induced by experience-dependent and visually based reference frames. *Manuscript in preparation*.

# A Study of Age-of-Acquisition Ratings in Adults

Gowri K. Iyer ([giyer@crl.ucsd.edu](mailto:giyer@crl.ucsd.edu))

Joint Doctoral Program 'Language & Communicative Disorders', SDSU/UCSD  
9500 Gilman Drive, Mail Code 0526, La Jolla, CA 92093-0526

Cristina M. Saccuman ([saccuman@crl.ucsd.edu](mailto:saccuman@crl.ucsd.edu))

Joint Doctoral Program 'Language & Communicative Disorders', SDSU/UCSD  
9500 Gilman Drive, Mail Code 0526, La Jolla, CA 92093-0526

Elizabeth A. Bates ([bates@crl.ucsd.edu](mailto:bates@crl.ucsd.edu))

Department of Cognitive Science, UCSD, 9500 Gilman Drive, Mail Code 0526, La Jolla, CA 92093-0526

Beverly B. Wulfeck ([wulfeck@crl.ucsd.edu](mailto:wulfeck@crl.ucsd.edu))

Department of Communicative Disorders, SDSU, San Diego, CA 92182-1518

Certain word attributes have been demonstrated to be important determinants of speed of processing in lexical tasks (such as picture naming, recognition tasks). Traditional accounts of lexical tasks using words and pictures have held that the most important among these word attributes is word frequency. However, there are a number of studies that indicate that, in some lexical tasks, apparent frequency effects may be wholly or partly accounted for by word age-of-acquisition (AoA), or word-learning age (Carroll & White, 1973a; Morrison et al., 1995).

In the literature, the methods used to obtain AoA data can be broadly grouped into two. The first method is objective and relies on the data collected directly from vocabulary tests and parental reports of children's abilities (Walley & Metsala, 1992). The second method is subjective and involves researchers obtaining age-of-acquisition ratings from adults. This second method allows for easier data collection and has been used in several studies (Carroll & White, 1973a, 1973b; Morrison et al., 1997; Snodgrass et al., 1996). Such studies have suggested that adult ratings of word acquisition age are a reliable tool to measure real word learning age and are also a better predictor (as compared to frequency and familiarity ratings) of subjects' performance on certain lexical tasks such as picture naming and recognition. Until recently, most studies have collected these adult AoA ratings using off-line techniques and using a relatively small number of stimuli (words and/or pictures).

The present study is an on-line experiment where we examined the AoA phenomenon in 50 normal, monolingual adults using a larger set of stimuli (520 words and/or pictures). The basic task, adapted from Carroll and White (1973b), involves subjects rating each item presented on a computer screen, on a 9-point age scale (2, 3, 4, 5, 6, 7-8, 9-10, 11-12, 13+ years) marked on the keyboard. The subjects' rating

responses and time taken to make these decisions are recorded. Results are discussed with reference to previous AoA studies and developmental norms. These results confirm that AoA ratings are good predictors of real word-learning age, and may be better predictors of naming latencies when compared to existing frequency norms and familiarity ratings. These results also raise some interesting theoretical issues regarding what these AoA adult rating measures tap into and its relevance to lexical access. Researchers have not been able to truly understand why the adult ratings are an important variable. However, many have tried to explain the relative advantage of AoA ratings over other word attributes such as frequency and familiarity.

## References

- Carroll, J. B., & White, M. N. (1973a). Word frequency and age of acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology*, 25, 85-95.
- Carroll, J. B., and White, M. N. (1973b). Age of acquisition norms for 220 pictureable nouns. *Journal of Verbal Learning and Verbal Behavior*, 12, 563-576.
- Morrison, C.M., Chappell, T.D., & Ellis, A.W. (1997) Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology*, 1997, 50A (3), 528-559.
- Morrison, C. M., & Ellis, A. W. (1995). The roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 21, 116-174.
- Snodgrass, J.G., & Yuditsky, T. (1996). Naming times for the Snodgrass and Vanderwart pictures. *Behavior Research Methods, Instruments, and Computers*, 28 (4), 516-536.
- Walley, A. C., & Metsala, J. L. (1992). Young children's age-of-acquisition estimates for spoken words. *Memory and Cognition*, 20, 171-182.



## The development of word recognition: The use of the possible word constraint by 12-month-olds

Elizabeth K. Johnson ([zab@jhu.edu](mailto:zab@jhu.edu))

Johns Hopkins University, Department of Psychology

Peter W. Jusczyk, ([jusczyk@jhu.edu](mailto:jusczyk@jhu.edu))

Johns Hopkins University, Department of Psychology

Anne Cutler ([anne@mpi.nl](mailto:anne@mpi.nl))

Max-Planck-Institute for Psycholinguistics

Dennis Norris ([dennis.norris@mrc-cbu.cam.ac.uk](mailto:dennis.norris@mrc-cbu.cam.ac.uk))

MRC Applied Psychology Unit, Cambridge, United Kingdom

Fluent speech contains no reliable pauses between words. By 7.5 months infants can segment words from fluent speech. At this age infants rely most heavily on prosodic cues such as word initial stress and distributional cues such as the transitional probabilities between syllables (Saffran et al, 1996). During the second year of life, infants' word recognition abilities undergo considerable improvement. Finally, by the age of 24 months, infants appear to be approaching adult-like word recognition skills (see Jusczyk, 1999 for review).

Many adult word recognition models emphasize the importance of existing items in the lexicon for recovering words from fluent speech (see Brent, 1999 for review). A question that arises about these segmentation abilities has to do with how infants integrate information from different types of word boundary cues. For instance, Norris et al (1997) suggested that a Possible-Word Constraint (PWC) could ease word recognition by limiting the number of lexical candidates activated by a given input. This constraint requires that, whenever possible, the input should be parsed into a string of feasible words. Any segmentation resulting in impossible words (i.e. a single consonant) is impossible. Norris et al (1997) used a word spotting task to demonstrate that adults find words such as "apple" more easily in a possible condition (ie vuffapple) rather than in an impossible condition (fapple).

In the present study, we investigated whether or not 12-month-olds could use the PWC to aid them in word recognition. In Expt. 1, we exposed 32 infants to lists of 2 words: rush and lop or rack and win. After 30 seconds of familiarization to both words, the head-turn preference procedure was used to determine whether these words were easier to recognize in a possible as opposed to impossible condition. During the test phase, infants were presented with test lists containing the target words embedded within possible (i.e. nprush) or impossible (i.e. prush) conditions. Infants tested with targets in possible conditions listened significantly longer to the lists containing the targets words ( $p < .01$ ), whereas infants tested with lists containing targets in impossible conditions did not. This result suggests that

12-month-olds, like adults, may use existing knowledge about possible words to constrain their hypotheses concerning words in the input. Infants familiarized with the word "rush" did not recognize the word "rush" when it was buried within a container like "prush." This result fits with the PWC because positing that "rush" is contained within "prush" would leave a residue which cannot form a word on its own: the consonant "p."

Experiment 2 tests whether infants use PWC when processing fluent speech. Infants are familiarized with pairs of words (rest and low or rise and lay). However, in the test phase they hear passages rather than word lists. Half of the infants in each of these two conditions are being tested with passages containing the target words in a possible condition (i.e. "delay"), while the other half are being tested on passages containing the words in impossible conditions (i.e. "play"). Preliminary results suggest that infants will recognize the words when they are presented in a "possible" condition.

In addition to Expt. 2, we have been carrying out 2 additional studies similar to Expts. 1 and 2. However we have moved the target to the beginning of the filler rather than the end. For example, the target word dull has been buried in a possible container such as "dullkef" or an impossible container such as "dullk." In combination, these 4 studies provide some interesting evidence concerning the role of context on infants' ability to find words in fluent speech.

### References

- Brent, M.R. (1999). Speech segmentation & word discovery: a computational perspective. *TICS* 3, 294-301.
- Jusczyk, P.W. (1999). How infants begin to extract words from speech. *TICS* 3, 323-327.
- Norris, D., McQueen, J.M., Cutler, A. & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cog Psychology* 34, 191-243.
- Saffran, J.R., Aslin, R.N. and Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926-1928.

# Familiarity for Nouns and Verbs: Not the Same as, and Better than, Frequency

Natalie Kacinik (kacinn01@student.ucr.edu)  
Connie Shears (shearc01@student.ucr.edu)  
Christine Chiarello (christine.chiarello@ucr.edu)

University of California, Riverside  
Life Sciences Psychology Building  
Riverside, CA 92521, U.S.A

Much of the study of language processing has centered on the single word recognition paradigm. Many cognitive theories regarding semantic memory have emerged from this research, and several lexical dimensions (i.e., imageability, frequency) have been found to influence the processing and recognition of nouns (Balota, Ferraro, and Connor, 1991). Indeed, efforts are typically made to balance word stimuli on factors such as length and frequency. However, the importance of different orthographic and semantic dimensions in determining the speed and accuracy with which words are responded to has not been extensively investigated. Moreover, most of this research has either not considered different word types (nouns vs. verbs), or has focused on concrete, imageable nouns, largely because of the lack of word norming corpora available for other word types.

Recently, new measures have been developed (Chiarello, Shears, & Lund, 1999) computing typicality of grammatical class (noun vs verb) and examining grammatical class differences in imageability and frequency, using established corpora such as Francis and Kucera (FK, 1982), as well as using the more contemporary Usenet corpus. While these semantic dimensions and word class comparisons have provided valuable tools for word recognition researchers, most studies have failed to consider word familiarity as an important determinant of speed and accuracy of responding (but see Gernsbacher, 1984, and Balota, Cortese, & Pilotti, 1999).

We report a series of regression analyses using data obtained from 2 lexical decision experiments and other corpora. We investigated the influence of variables identified in Chiarello et al. (1999) [i.e., imageability, length, noun-verb distributional distance (NVDD), FK and Usenet frequency, and recently collected familiarity ratings] on the speed and accuracy of lexical decision responses to nouns and verbs. Familiarity, measured on a 7 pt. scale, was defined as 'common in everyday experience'.

Overall, familiarity was found to be highly correlated with RT ( $r = -.70$ ,  $p < .001$ ), thereby accounting for nearly half of the variance. Although significantly correlated with imageability, NVDD, FK and Usenet frequency ( $r = .22$ ,  $.23$ ,  $.39$ , and  $.40$ , all  $ps < .005$ ), regression analyses indicated that much of the RT variance accounted for by familiarity was unique. The importance of these variables in predicting RT also varied by word class (nouns vs verbs). Specifically, familiarity, then frequency, and then imageability were found to be the most important predictors

of noun RT, whereas familiarity, then imageability, then frequency, and finally NVDD were found to be the most important predictors of verb RT.

In conclusion, our results support and extend Gernsbacher's (1984) earlier demonstration of familiarity as a powerful contributor to word recognition, possibly because it is a contemporary metric of actual encounters, related to the variety of contexts a word has been experienced in, and the ease with which individuals can recall those contexts (Audet & Burgess, 1999). Our findings indicate the need for researchers to consider the importance of processing differences based on *the familiarity of stimuli to the subject population* (i.e., controlling only for frequency and imageability may not be enough). Finally, we also demonstrate the need for researchers to carefully consider the issue of word class, as different dimensions appear to be more or less important for the processing of nouns and verbs.

## References

- Audet, C., & Burgess, C. (1999). Using a high-dimensional memory model to evaluate the properties of abstract and concrete words. In M. Hahn, & S.C. Stoness (Eds.), *Proceedings of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.
- Balota, D.A., Cortese, M.J., & Pilotti, M. (1999). Item-level analyses of lexical decision performance: Results from a mega-study. In *Abstracts of the 40th Annual Meeting of the Psychonomics Society*. Los Angeles, CA: Psychonomic Society.
- Balota, D., Ferraro, R., & Connor, L. (1991). On the early influence of meaning in word recognition: A review of the literature. In P.J. Schwanenflugel (Ed.), *The psychology of word meanings*. Hillsdale, NJ: Erlbaum.
- Chiarello, C., Shears, C., & Lund, K. (1999). Imageability and distributional typicality measures of nouns and verbs in contemporary English. *Behavior Research Methods, Instruments, & Computers*, 31, 603-637.
- Francis, W., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.
- Gernsbacher, M.A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 113, 256-281.

# Random Indexing of Text Samples for Latent Semantic Analysis

Pentti Kanerva Jan Kristoferson Anders Holst

kanerva@sics.se, janke@sics.se, aho@sics.se

RWCP Theoretical Foundation SICS Laboratory

Swedish Institute of Computer Science, Box 1263, SE-16429 Kista, Sweden

Latent Semantic Analysis is a method of computing high-dimensional semantic vectors, or context vectors, for words from their co-occurrence statistics. An experiment by Landauer & Dumais (1997) covers a vocabulary of 60,000 words (unique letter strings delimited by word-space characters) in 30,000 contexts (text samples or “documents” of about 150 words each). The data are first collected into a  $60,000 \times 30,000$  words-by-contexts co-occurrence matrix, with each row representing a word and each column representing a text sample so that each entry gives the frequency of a given word in a given text sample. The frequencies are normalized, and the normalized matrix is transformed with Singular-Value Decomposition (SVD) reducing its original 30,000 document dimensions into a much smaller number of latent dimensions, 300 proving to be optimal. Thus words are represented by 300-dimensional semantic vectors.

The point in all of this is that the vectors capture meaning. Landauer and Dumais demonstrate it with a synonym test called TOEFL (for “Test Of English as a Foreign Language”). For each test word, four alternatives are given, and the “contestant” is asked to find the one that’s the most synonymous. Choosing at random would yield 25% correct. However, when the semantic vector for the test word is compared to the semantic vectors for the four alternatives, it correlates most highly with the correct alternative in 64% of the cases. However, when the same test is based on the 30,000-dimensional vectors before SVD, the result is not nearly as good: only 36% correct. The authors conclude that the reorganization of information by SVD somehow corresponds to human psychology.

We have studied high-dimensional random distributed representations, as models of brainlike representation of information (Kanerva, 1994; Kanerva & Sjödin, 1999). In this poster we report on the use of such a representation to reduce the dimensionality of the original words-by-contexts matrix. The method can be explained by looking at the  $60,000 \times 30,000$  matrix of frequencies above. Assume that each text sample is represented by a 30,000-bit vector with a single 1 marking the place of the sample in a list of all samples, and call it the sample’s *index vector* (i.e., the  $n$ th bit of the index vector for the  $n$ th text sample is 1—the representation is unitary or local). Then the words-by-contexts matrix of frequencies can be gotten by the following procedure: every time that the word  $w$  occurs in the  $n$ th text sample, the  $n$ th index vector is added to the row for the word  $w$ .

We use the same procedure for accumulating a words-by-contexts matrix, except that the index vectors are not unitary. A text-sample’s index vector is “small” by comparison—we have used 1,800-dimensional index

vectors—and it has several randomly placed  $-1$ s and  $1$ s, with the rest 0s (e.g., four each of  $-1$  and  $1$ , or eight non-0s in 1,800, instead of one non-0 in 30,000 as above). Thus, we would accumulate the same data into a  $60,000 \times 1,800$  words-by-contexts matrix instead of  $60,000 \times 30,000$ .

Our method has been verified with different data, a ten-million-word “TASA” corpus consisting of a 79,000-word vocabulary (when words are truncated after the 8th character) in 37,600 text samples. The data were accumulated into a  $79,000 \times 1,800$  words-by-contexts matrix, which was normalized by thresholding into a matrix of  $-1$ s, 0s, and  $1$ s. The unnormalized 1,800-dimensional context vectors gave 35–44% correct in the TOEFL test and the normalized ones gave 48–51% correct, which correspond to Landauer & Dumais’ 36% for their normalized 30,000-dimensional vectors before SVD, for a different corpus (see above). Our words-by-contexts matrix can be transformed further, for example with SVD as in LSA, except that the matrix is much smaller.

Mathematically, the 30,000- or 37,600-dimensional index vectors are orthogonal, whereas the 1,800-dimensional ones are only nearly orthogonal. They seem to work just as well, in addition to which they are more “brainlike” and less affected by the number of text samples (1,800-dimensional index vectors can cover a wide-ranging number of text samples). We have used such vectors also to index words in narrow context windows, getting 62–70% correct, and conclude that random indexing deserves to be studied and understood more fully.

**Acknowledgments.** This research is supported by Japan’s Ministry of International Trade and Industry (MITI) under the Real World Computing Partnership (RWCP) program. The TASA corpus and 80 TOEFL test items were made available to us by courtesy of Professor Thomas Landauer, University of Colorado.

## References

- Kanerva, P. (1994). The Spatter Code for encoding concepts at many levels. In M. Marinaro and P. G. Morasso (eds.), *ICANN '94, Proc. Int'l Conference on Artificial Neural Networks* (Sorrento, Italy), vol. 1, pp. 226–229. London: Springer-Verlag.
- Kanerva, P., and Sjödin, G. (1999). Stochastic Pattern Computing. *Proc. 2000 Real World Computing Symposium* (Report TR-99-002, pp. 271–276). Tsukubacity, Japan: Real World Computing Partnership.
- Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato’s problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* 104(2):211–240.

# Auditory and Visual Continuity Perception: A Unifying Theory

Leah M. Knightly (leah@psych.ucla.edu)  
UCLA Department of Psychology; Box 951563  
Los Angeles, CA 90095-1563 USA

## Introduction

Researchers in the field of auditory and visual perception have been intrigued by our ability to unify partially occluded objects (Shipley & Kellman, 1992; Kellman & Shipley, 1991) and partially masked sounds (Dannenbring, 1976; Ciocca & Bregman, 1987). In vision, an object may be partially occluded by another object yet we may perceive the object as continuing behind the occluder. In audition, sounds may be partially masked by another sound, yet we may hear the sound as continuing through the mask. Though these phenomena are considered to be analogous (Bregman, 1990), separate theories exist to predict the conditions under which continuity perception occurs in vision (Relatability Theory - Kellman & Shipley, 1991) and audition (Frequency Proximity and Trajectory principle - Ciocca & Bregman, 1987). The purpose of this paper is to propose that the conditions under which continuity perception occurs for edges and tones may be predicted by one theory. This theory, introduced here and inspired by Relatability Theory, is called "Continuity Theory Audio-Visual (AV)."

## Continuity Theory (AV)

Continuity Theory (AV) predicts that a partially occluded stimulus will be perceived as continuing behind an obstruction if the linear extensions of the stimulus on either side of the obstruction meet within the bounds of the obstruction. For the visual domain, this means that an edge partially covered by an occluder will be perceived as continuing behind the occluder if the linear extensions of the edges meet within the area occupied by the occluder. For the auditory domain, this means that a tone partially masked by a noise burst will be perceived as continuing through the noise burst if the linear extensions of the pre and post-noise frequencies meet within the duration of the mask.

Evidence that Continuity Theory (AV) can predict the conditions under which edges and tones are perceived as continuous is provided through a critical analysis of the results obtained in two studies – Shipley & Kellman (1992) on unit formation in vision and Ciocca & Bregman (1987) on perception of tones through noise. In Shipley & Kellman (1992) participants perceived partially occluded figures as unified if the linear extensions of their edges met within the bounds of relatability (see Kellman & Shipley, 1991 for details). In Ciocca & Bregman (1987) listeners perceived sounds as continuing through a burst of noise depending on

the frequency and trajectory (i.e. linear extension) of the pre and post-noise tones. Close examination of the results from these two studies reveals that one theory is sufficient to describe the conditions under which continuity perception occurred. This theory is Continuity Theory (AV). The advantage of the theory is it can account for the results obtained in vision and audition. In addition, the theory includes size/duration of the occluder/mask as a factor in unit formation, a variable not incorporated in other theories of continuity (Kellman & Shipley, 1992; Ciocca & Bregman, 1987) yet considered to be important in continuity perception (Vicario, 1982).

## Future Directions

In summary, Continuity Theory (AV) provides a simple and general cross-modal rule that predicts continuity perception for those conditions tested in Shipley & Kellman (1992) and Ciocca & Bregman (1987). Future work should involve testing Continuity Theory (AV) for those conditions not examined in Shipley & Kellman (1992) (i.e. when linear extensions meet at an angle  $< 90^\circ$ ) and in Ciocca and Bregman (1987) (i.e. when linear extensions meet at an angle  $> 90^\circ$ ).

## References

- Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: The MIT Press.
- Ciocca, V. & Bregman, A. S. (1987). Perceived continuity of gliding and steady-state tones through interrupting noise. *Perception and Psychophysics*, 42 (5), 476-484.
- Dannenbring, G. L. Perceived auditory continuity with alternately rising and falling frequency transitions. *Canadian Journal of Psychology*, 1976 Jun, v30 (n2):99-114.
- Kellman, P. J. & Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cognitive Psychology*, 23, 141-221.
- Shipley, T. F. & Kellman, P. J. (1992). Perception of partly occluded objects and illusory figures: Evidence for an identity hypothesis. *Journal of Experimental Psychology: Human Perception and Performance*, 18 (1), 106-120.
- Vicario, Giovanni, B. (1982). Some observations in the Auditory Field. In J. Beck, *Organization and Representation in Perception* (pp. 269-283). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

# The Role of Working Memory in Homograph Recognition

Yuki Kobayashi (yukikoba@srt.l.u-tokyo.ac.jp)

Department of Psychology, University of Tokyo; 7-3-1 Hongo Bunkyo-ku  
Tokyo, 1130033 Japan

Recognition of homographs is usually assumed to consist of two process stages: automatic access to mental lexicon and inhibition of meanings irrelevant to context (Miyake et al., 1994). Working memory is related to the inhibition: Kobayashi and Takano (1999) showed that readers with larger working memory capacity can inhibit irrelevant meanings faster than those with smaller capacity. The homographs used in the study had only two major meanings. The number of meanings can effect to recognition of homographs. I predicted that readers with larger working memory capacity cannot inhibit when the number of meanings increases more than two meanings.

## Method

### Subjects

The subjects were 29 undergraduates from the University of Tokyo. All were native speakers of Japanese.

### Materials

**Lexical Decision Task** Fifteen homographs were selected as first primes. Five homographs each had four major meanings, and ten each had two major meanings. They were all written in *kana* (i.e., Japanese phonogram). Targets were these homographs written in *kanji* (i.e., Chinese ideogram). Second prime was a pair of *kanji* related to the target in meanings (consistent condition), an asterisk (neutral condition), and a pair of *kanji* related to another target in meanings (inconsistent condition).

**Japanese Reading Span Test** We used Osaka and Osaka's (1994) Japanese version of the test.

### Design

The independent variables were consistency (consistent vs. neutral vs. inconsistent) and number of irrelevant meanings (three vs. one). I examined reading span scores as a pseudo-independent variable, too. The dependent measure was RT for targets.

### Procedure

**Japanese reading span test** Osaka and Osaka's (1994) test was administered.

**Lexical Decision Task** After a fixation point was presented for 1s, first prime, second prime and target were presented successively. The SOA of primes was 500ms. Subjects were requested to judge whether the target was word or non-word as accurately and quickly as possible.

## Results and Discussion

Subjects with reading spans of 3.0 or greater were considered to be high-span readers; those with spans of 2.5 or less, to be low-span readers.

The main results are presented in Figure 1. High span readers took longer time in inconsistent condition than in neutral condition, whereas low span readers showed no significant difference between neutral condition and inconsistent condition (Figure 1). The number of meanings had no significant interaction with consistency and reading span. High-span readers could inhibit irrelevant meanings though meanings increased, whereas low-span readers couldn't even inhibit one irrelevant meaning. We conclude that the number of meanings didn't effect to inhibition in working memory.

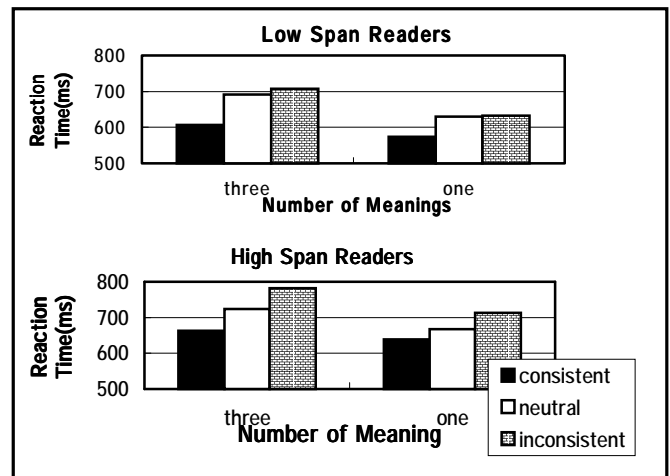


Figure 1: Reaction time of low span readers and high span readers.

## Reference

- Kobayashi, Y. , & Takano, Y. (1999). The mental process of homograph recognition : The examination of inhibition in working memory. *Proceedings of the 2nd International Conference on Cognitive Science*, 507-510. (In Japanese)
- Miyake, A., Just, M.A., & Carpenter, P.A. (1994). Working memory constraints on the resolution of lexical ambiguity : Maintaining multiple interpretations in neutral contexts. *Journal of Memory and Language*, **33**, 175-202.
- Osaka, M., & Osaka, N. (1994). Capacity related to reading – Measurement with the Japanese version of reading span test. *Japanese Journal of Psychology*, **65**,339-345.

# Attentional Perseveration after the Inverse Base–Rate Effect

John K. Kruschke, Mark K. Johansen, and Nathaniel J. Blair

kruschke@indiana.edu, mjohanse@indiana.edu, nblair@indiana.edu

Department of Psychology; Indiana University

Bloomington, IN USA 47405-7007

We report new results that demonstrate that selective attention to features is learned in the inverse base-rate effect. The inverse base-rate effect (Medin & Edelson, 1988) is found after participants have learned categories with different base rates (frequencies of occurrence). When tested with conflicting cues, participants tended to non-normatively respond with the low frequency category, suggesting that they were ignoring base-rate information.

The top of Table 1 (Training 1) shows the category structure for producing the inverse base-rate effect in disease diagnosis. The common diseases (C1 and C2) occur three times more frequently than the rare diseases (R1 and R2). One symptom (I1 or I2) is shared by two diseases and is an imperfect predictor. The other symptoms are perfect predictors, that is, they are associated with one and only one disease. In the testing phase (in the middle of Table 1) when shown I1 alone, participants tended to respond with C1. But, when tested with PC1&PR1 simultaneously participants tended to respond R1. The normative response, however, would have been to use the 3:1 base–rate information and respond C1.

Kruschke (1996) hypothesized that the inverse base-rate effect occurs because participants rapidly shift attention to reduce error while learning. Specifically he argued that participants tend to learn C1 before R1 because it occurs more frequently, and encode C1 in terms of both I1 and PC1. When subsequently learning R1, participants shift attention away from I1 and toward PR1 to avoid incorrectly responding with C1 and to protect what they have already learned about C1. Hence R1 tends to be encoded primarily in terms of PR1. Kruschke (1996) formalized this hypothesis in a connectionist model called ADIT which provides an extremely accurate account of the inverse base-rate effect data.

If the attentional account of the inverse base-rate effect is correct, it suggests that attention to the symptoms should *perseverate* into a subsequent learning task. To test this hypothesis, we added two different conditions after the test phase (the bottom of Table 1). The first condition (Training 2: “EASY”) was designed to be easy to learn because PR is relevant for correct diagnosis, just as in previous training. This should have been easy because subjects should have already learned to shift attention away from I1 and toward PR1 (or away from I2 and toward PR2).

The second condition (Training 2: “HARD”) was designed to be hard to learn because PR is irrelevant, unlike previous training. This should have been hard because while subjects should have previously learned to attend to PR1 (and

Table 1: Design of the experiment

Training 1:	I1&PC1 → C1 (3×)	I2&PC2 → C2 (3×)
	I1&PR1 → R1 (1×)	I2&PR2 → R2 (1×)
Testing:	I1? (→ C1)	PC1&PR1? (→ R1)
	I2? (→ C2)	PC2&PR2? (→ R2)
Training 2:	EASY: <b>PR</b> is relevant	HARD: <b>I</b> is relevant
	I1&PR1 → R1	I1&PR1 → R1
	I2&PR1 → R1	I2&PR1 → R2
	I1&PR2 → R2	I1&PR2 → R1
	I2&PR2 → R2	I2&PR2 → R2

PR2) and ignore I1 (and I2), I1 and I2 now were essential to learning the new diagnoses.

Participants learned the “EASY” condition in phase II significantly faster than they learned the “HARD” condition. These results, along with others involving I and PC, support our hypothesis that learned attention to features perseverates into later learning. These results cannot be explained by an eliminative inference account, such as that presented by Juslin, Wennerholm, & Winman (1999).

Rapid attention shifts have also been implication in probabilistic learning tasks (Kruschke & Johansen, 1999). Perseveration of learned attention has also recently been implicated in the classic learning phenomenon of blocking (Kruschke & Blair, 2000).

## References

- Juslin, P., Wennerholm, P., & Winman, A. (1999). Mirroring the inverse base-rate effect: the novel symptom phenomenon. In *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society* Mahwah, NJ: Erlbaum.
- Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 22, 3–26.
- Kruschke, J. K. & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, 00, 000–000. In press. Available from <http://www.indiana.edu/~kruschke/kb99.html>.
- Kruschke, J. K. & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 25, 1083–1119.
- Medin, D. L. & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, 117, 68–85.

# Do Readers Make Predictive Inferences about Conversations?

**R. Brooke Lea (lea@macalester.edu)**

Department of Psychology, Macalester College  
1600 Grand Avenue, Saint Paul, MN 55105 USA

**Patrick A. Kayser (pkayser@macalester.edu)**

Department of Psychology; Macalester College  
1600 Grand Avenue, Saint Paul, MN 55105 USA

**Elizabeth J. Mulligan (mulligan@psych.colorado.edu)**

Department of Psychology, University of Colorado, Boulder  
Boulder, CO 80309 USA

**Jerome L. Myers (jlmyers@psych.umass.edu)**

Department of Psychology, University of Massachusetts, Amherst  
Amherst, MA 01003 USA

Do readers make predictive inferences about what protagonists in a story are talking about? Lea, Mason, Albrecht, Birch, and Myers (1998) showed that when two protagonists part and then reunite, information associated with the protagonists is reactivated by their reunion via a low-level memory process (resonance). We used Lea et al.'s passages to test whether this reactivated information is then used to make predictive inferences about what the protagonists talk about after their reunion. In an example passage, Gloria tells her roommate Jane that she is going out and that they will meet later. In the intervening interval, Gloria has dinner with her cousin, while Jane makes dinner at home. Later in the passage, Gloria returns home (reunion), and they "chat for a while." Previous work has shown that the cousin is significantly more active after the reunion sentence than before it (the "reunion effect") — we wondered whether readers then use that activated information to infer what Gloria and Jane are chatting about.

In Experiment 1 we measured activation of COUSIN after two types of discussion sentences and compared them to a no-discussion control. A sentence like "They chatted for a while." was used in the Discussion condition (D); a sentence like "Just wait until you hear this." appeared in the Urgent-Discussion (UD) condition; and the No-Discussion (ND) control passage described a situation in which the protagonists reunited but no discussion was possible (e.g., because one of them was asleep). In all three conditions the target character (e.g., COUSIN) should be reactivated after the reunion, but if readers infer that the cousin is part of the discussion then its reactivation should be potentiated by the discussion sentences. We found that both discussion sentences lead to significantly faster recognition times than the control. The UD passages produced faster response times than the D passages, but the difference was not significant. Thus it appears that readers were making predictive inferences about the topic of the protagonists' discussion.

An alternative explanation for the results of Experiment 1 is that the activation difference reflects a difference in the reunion sentences, not a difference in the discussion sentences. All three versions contained reunion sentences in which both protagonists were mentioned, however, the no-discussion control passages required different reunion sentences in order to create a convincing no-discussion-possible situation. Resonance theory (e.g. Myers &

O'Brien, 1998) would not predict a difference between the two reunion types, but the possibility remains that a "linguistic" reunion like "Jane was asleep when Gloria returned home" does not reactive COUSIN to the same degree that a "physical" reunion such as "Jane was still awake when Gloria returned home" does. So in Experiment 2 we probed immediately after both types of reunion and used a before-reunion probe position control. If reunion-type makes a difference, then we should find a differential before-after reunion effect. However, we found that the target character was reactivated equally after both linguistic and physical reunions, thereby supporting the conclusion that Experiment 1's results are due to the discussion sentence manipulation and not to a difference between the reunions.

Experiment 3 was a paper-and-pencil experiment in which subjects were presented with printed versions of the passages that ended with the discussion sentence, and they were instructed to write a sentence or two about what they thought would be a likely continuation of the story. We conducted this off-line experiment to obtain converging evidence that readers were in fact making an inference that the target character was being discussed. As predicted, subjects were significantly more likely to mention the target character after the discussion sentences compared to the no-discussion control. Interestingly, the UD condition lead to significantly more mentions than the D condition, a difference that was only a trend in Experiment 1.

Together, the three experiments demonstrate how low-level, memory-based text processing can work in concert with more expectation-driven processing. In our passages, reintroducing a protagonist reactivated that target character with whom she was associated and, once reminded, the reader exploited the availability of that information to make a forward inference about the likely topic of the protagonists' conversation. Future work will explore further the collaboration between bottom-up processes like resonance, and more top-down reading processes such as predictive inference.

## References

- Lea, R.B., Mason, R.A., Albrecht, J.E., Birch, S., & Myers, J.L. (1998). Who knows what about whom: What role does common ground play in accessing distant information? *Journal of Memory and Language*, 39, 70-84.
- Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, 26, 131-157.

# A Model of Prefrontal-Hippocampal Interactions in Strategic Recall

Jean C. Lim, [lim@curie.psychology.mcmaster.ca](mailto:lim@curie.psychology.mcmaster.ca)

Suzanna Becker, [becker@mcmaster.ca](mailto:becker@mcmaster.ca)

Department of Psychology, McMaster University, 1280 Main St. West, Hamilton, ON, Canada.

May 24, 2000

Retrieval of episodic memories may be aided by the prefrontal cortex either by its providing contextual, temporal source cues or by serving an executive role - strategically organizing information into chunks, categorizing, and separating lists (Stuss, 1986). The hippocampal region is also believed to play a role in encoding episodic memories (Zola-Morgan et. al., 1990). Modelling human performance on free recall tasks that involve strategic organization of items is difficult because temporal memory of events, delayed rewards and learning in the absence of external reinforcement are required. For example, in the California Verbal Learning Test (CVLT) the task is to study a list of 16 words (four words each from four different semantic categories) and recall the list over five repeated trials (Delis et. al., 1987). Young healthy subjects typically use a semantic clustering strategy to recall the list. Elderly and frontal lobe damaged patients fail to subjectively organize such words and show poorer recall performance (Hultsch, 1975).

To simulate the hippocampal component of our model, we used a Hopfield network (Hopfield, 1982), because of its rapid learning, pattern association, recall and recognition capabilities. For our prefrontal module, we used a recurrent network trained with a reinforcement learning rule similar to that proposed by Barto and Sutton (1986) which can detect correlations between traces of past inputs and changes in outputs. For our list-learning task, a positive reward signal was used to strengthen relevant prefrontal weights during study. During retrieval, recall of non-studied items resulted in an internally generated negative signal.

The hippocampal module consisted of 400 recurrent, symmetrically connected units with no self-feedback connections. Bidirectional and symmetrical weights connected each unit of the Hopfield network to two layers: (1) the prefrontal cortex - a layer of 10 units; and (2) an input/output layer of 52 units - localist representations of the vocabulary words.

Weights to each localist word unit were pre-trained with a Hebbian update rule. This established the network's pre-experimental vocabulary of 52 word patterns. Sixteen of these words, drawn from four different semantic categories (four words from each), were the study list words. Of the remaining 36 words, 20 were semantically similar to the study list, eight were drawn from two new semantic categories, and eight were sparse random vectors. The representation of a word consisted of a distributed vector of 400 semantic features, with semantically related words having more highly correlated feature vectors. The simulation consisted of 5 study and recall trials. During study, the network was trained on the 16 list word patterns. During retrieval, the prefrontal acti-

vations served as cues to recall the list words. Learning in both phases took place in the connections between the prefrontal and hippocampal modules. A recalled word could be "correct", a perseveration (repetition) or an intrusion (non-list word). Clustering performance was determined by comparing the observed semantic clustering score to the expected clustering score.

We modelled a frontal lesioned network by freezing its prefrontal weights at zero. Calculated ratios of observed to expected cluster scores showed that the lesioned network did not cluster above chance whereas the normal network did. The lesioned model also produced, in order of frequency, more perseverative errors, similar intrusions and random intrusion errors than the normal model. Simulation results suggest that an elderly or frontal-lobe damaged subject, modelled by a hippocampal system operating independently of the prefrontal layer, is capable of pattern recognition and recall when given external guidance and cues. To perform more complex tasks such as free recall and provide context-rich source cues, an additional layer, represented in our model as the prefrontal cortex, is needed. With temporally predictive reinforcement learning, the prefrontal module was able to detect semantic clustering of word patterns, and use this to generate retrieval cues to recall the list items optimally.

- Delis, D., Kramer, J., Kaplan, E., & Ober, B. (1987). *The California Verbal Learning Test*. San Antonio, Tx: Psychological Corporation.
- Hopfield, J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences*, 79:2554-2558.
- Hultsch, D. (1975). Adult age differences in retrieval: Trace-dependent and cue-dependent forgetting. *Developmental Psychology*, 12, 83-84.
- Stuss, D., Benson, F. (1986) *The Frontal Lobes*. Raven Press. New York.
- Sutton, S., Barto, A. (1981) Toward a Modern Theory of Adaptive Networks: Expectation and Prediction. *Psychological Review* 88:2:135-170.
- Zola-Morgan, S., Squire, L. (1990). The neuropsychology of memory: parallel findings in humans and non-human primates. In A. Diamond (Ed.), *The Development of Learning*. New York: New York Academy of Sciences.



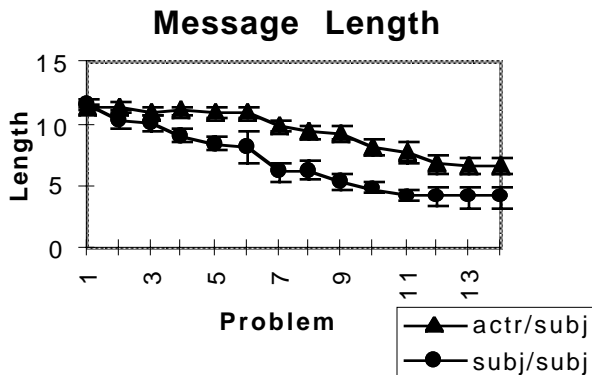
# An Adaptive Model of Simple Communication

Michael Matessa & John R. Anderson (matessa@cmu.edu)

Department of Psychology; Carnegie Mellon University  
Pittsburgh, PA 15213 USA

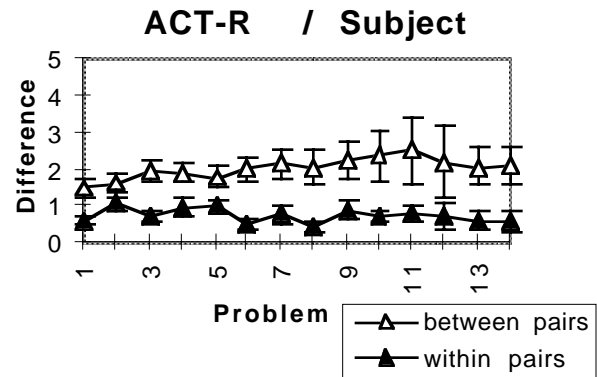
When people communicate they try to establish mutual knowledge. Garrod and Anderson (1987) proposed that a way to minimize effort during this process would be to follow a “output/input coordination” principle, where output to a partner is formulated according to the same principles of interpretation as those needed to interpret input from a partner. A computational model of establishing mutual knowledge efficiently can be given in the ACT-R architecture (Anderson & Lebiere, 1998) where goals that are completed successfully can be retrieved and used later. Applied to communication, goals of presenting and accepting information include semantic and syntactic representations of that information, and these goals can later be retrieved to provide templates for the creation of new utterances. Results from an ACT-R model communicating with human subjects show similar performance to that of human subjects communicating together.

The ACT-R model incorporates current theories of collaborative communication which fit naturally into the architecture. These theories include the creation of common ground by way of successful goals of presentation and acceptance (Clark & Schaefer, 1989), the use of dialogue acts to represent actions performed by speech (Core & Allen, 1997), the use of communicative obligations to motivate conversation (Traum & Allen, 1994), and the use of input from a partner to formulate output to that partner (Garrod & Anderson, 1987).



Subjects in a communication task were found to use fewer words to solve problems over time. An ACT-R model

interacting with subjects also used fewer words over time because previous utterances from its human partner were used as templates to create new utterances.



This behavior can be shown to be partner-dependent by showing the difference in message length within pairs is less than the difference between pairs. This was true for both the ACT-R model interacting with subjects and subjects interacting with other subjects.

Work in progress includes the creation of a model that purposely formulates output that is different than the input from a partner's speech to test the effect of non-accommodation on communicative efficiency.

## Acknowledgments

This research has been supported by grants N00014-96-I-0491 from the Office of Naval Research and SBR-94-21332 from the National Science Foundation.

## References

- Anderson, J. R., and Lebiere, C. 1998. *The Atomic Components of Thought*. Hillsdale, Mawhaw, NJ: Erlbaum.
- Clark, H., & Schaefer, E. (1989). Contributing to Discourse. *Cognitive Science*, 13, 259-294.
- Garrod, S. & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27, 181-218.
- Traum, D., & Allen, J. (1994). Discourse Obligations in Dialogue Processing. In *ACL94, Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 1-8.

# Morphological Influences on Phonetic Categorization

Kerstin Mauth (kerstin.mauth@mpi.nl)  
Max-Planck-Institute for Psycholinguistics, Wundtlaan 1  
6525 XD Nijmegen, Netherlands

Listeners' phonetic decisions about ambiguous sounds are influenced by lexical information as well as syntactic and semantic information from sentences. A series of experiments was constructed to test whether preceding sentential context can modulate the perception of inflectional morphemes. The morpheme under investigation was the verbal 3rd person singular marker -t in Dutch. Listeners were presented with two different types of sentences: (A) *Vraag jij of Jan morgen gaat?* 'Are you asking whether Jan leaves tomorrow?' (B) *Zie jij nog wel eens een plaat?* 'Do you see a record now and then?' The final words were chosen so that they were semantically not highly predictable from the context. But they were syntactically predictable: the first sentence had to end with a verb while the second had to end with a noun. The final consonant in each sentence was a stop that varied along a place of articulation continuum from [t] to [k]. The [k]-endpoints always formed nonwords. Sentences like (A) and (B) were compared with matched control sentences ending with [t]- and [k]-final nonwords (e.g., *snaat / snaak*).

Listeners were required to categorize the final consonants, which were clear instances of [t] and [k] at the respective endpoints and ambiguous between the two in six intermediate steps. The main question was whether the shift in the categorization function towards the [t] would be any different for the verbal context (A) as compared to the nominal context (B). Would people benefit from the fact that the phoneme [t] is more or less predictable on the basis of the verbal context because of its morphological status? If yes, this might indicate the operation of a morphological decomposition process.

As shown in Figure 1 the bias towards [t] was indeed greater in the verbal context than in the nominal context. Separate analyses within RT ranges (fast, medium and slow) also showed different patterns over time for verbal and nominal contexts. While the shift in the identification

function towards [t] in the verbal context was largest in the fastest reactions it was not reliable in this RT range in the nominal context. In the medium RT range the shift weakened in the verbal condition but built up in the nominal condition. No significant shifts were observed in the slowest responses.

But how far can this result be attributed to the listeners' processing of the context? Would a similar difference between the functions for verbs and nouns show up when these words are presented in isolation? The results (as shown in Figure 2) of a follow-up experiment where the final words and nonwords of the previous experiment were presented in isolation showed that this was not the case.

In an overall analysis listeners did not give significantly more [t]-responses in word contexts than in nonword contexts. This result with high quality materials is congruent with previous findings by McQueen (1991) who found a significant lexicality effect for word final phonemes only when the material was degraded. In the fast RT range of the present data, however, there was a lexicality effect, which was larger for nouns than for verbs.

These results suggest that people do benefit from the morphological status of the last phoneme in a word when not only the word but also its inflection is predictable from the context. As the largest shift in the identification function in the verbal context appears in the fastest RTs this morphological decomposition process is a very rapid one. Without preceding context, however, an inflectional morpheme is not treated differently from phonemes that are part of the word's stem.

## References

McQueen, J. (1991). The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 433-443.

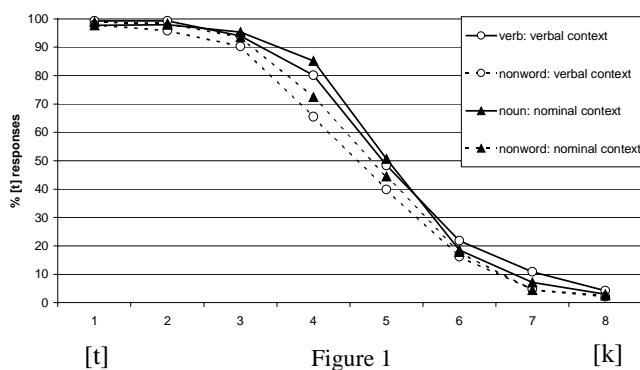


Figure 1

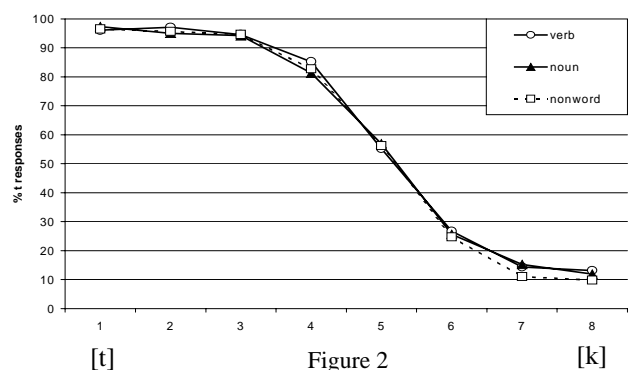


Figure 2

# Unique Entropy As A Model Of Linguistic Classification

Toben H. Mintz (tmintz@usc.edu)

Department of Psychology, SGM 501; University of Southern California  
Los Angeles, CA 90064-1061 USA

Several researchers have proposed that young children could make use of statistically weighted distributional information as a significant source of information about the categories of words in their language (Cartwright & Brent, 1997; Mintz, Newport, & Bever, 1999; Redington, Chater, & Finch, 1998). Most of these analyses result in a hierarchical cluster analysis (HCA) which clusters words together based on their distributional similarity. HCAs do not produce categories, but rather graded clusters based on similarity. A similarity threshold must be chosen such that words in clusters which exceed the similarity threshold are said to belong to the same category. Finding a deterministic method for selecting the categorization threshold which results in optimal linguistic categorization, and which does not rely on *a priori* knowledge of the correct linguistic categories, has been problematic. In this poster, I propose a deterministic solution for choosing categorization thresholds in HCAs. I present the notion Unique Entropy which, when applied to linguistic corpora, yields an optimal categorization of words into grammatical categories.

One can characterize the notion "best categorization point" for a HCA on formal, information-theoretic grounds. Specifically, the similarity threshold which yields the highest Entropy (Equation 1,  $l$ =number of groups), will provide the categorization level which maximizes the intrinsic information carried by the resulting category structure. "Best categorization" in this sense means "best" in terms of the amount of information inherent in the resulting category structure, independent of whether it best approximates the linguistic categories being sought. It is an empirical question, whether the best information theoretic classification results in the best linguistic classification. I now demonstrate that it does, at least for the four corpora analyzed in Mintz et al. (1999).

$$(1) E_l = -\sum_1^l \log(p(i))p(i), \quad p(i) = \frac{\text{number of elements in cluster } i}{l}$$

$$(2) UE_l = E_l - ((-\log(\frac{m-(n-1)}{m}))(\frac{m-(n-1)}{m})) + (n-1)(-\log(\frac{1}{m})(\frac{1}{m}))$$

The Entropy, or Information, in a set of categories is affected in two ways by the structure of the set. 1) For a set of a given number of categories, information contained in the category structure (Entropy) will be higher when categories contain the same number of items than when items are unevenly distributed among categories. 2) All else being equal, having more categories results in greater Entropy. In selecting an optimal categorization point based on maximum Entropy, one only wants to consider sources of Entropy that are due to the specific characteristics of the HCA in question and not which are due to merely having a certain number of

categories. Therefore, to determine the unique information provided by a HCA of  $m$  items at a given categorization threshold,  $l$ , which yields  $n$  categories, one must subtract out the base information that would come merely from having  $n$  categories. The result of this subtraction I call Unique Entropy (UE, Equation 2).

Figure 1a plots Unique Entropy by number of categories for the distributional analyses of four corpora presented in Mintz et al. (1999). Mintz et al. reported that the best linguistic categorization in their HCAs was obtained when members were divided into about 30 groups. This is shown by the vertical bar in Figure 1a, and corresponds to the regions with the highest Unique Entropy points for each corpus. Thus, it appears that the best linguistic classification for these corpora is achieved by selecting the classification level with the highest Unique Entropy.

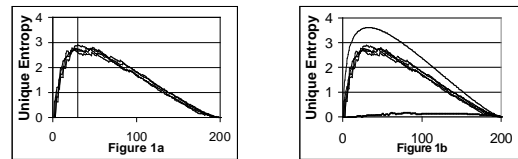


Figure 1b shows that the specific character of the UE curves produced by the distributional analyses of child directed speech is not a necessary consequence of performing such an analysis on any corpus. The lowest line plots the average UE of 10 pseudo-corpora generated by randomly ordering the words in one of the four Mintz et al. corpora. This UE curve shows that any information inherent in the random pseudo-corpora HCAs is due simply to having a given number of categories. The top curve in Figure 1b shows the upper bound for UE when classifying 200 items into  $n$  categories. The four corpus based curves are repeated in Figure 1b. The structure of the actual corpus based HCAs are nearly maximally informative by this measure.

Further research will explore the implications of this finding for psycholinguistics, as well as investigate how it extends to other areas of human categorization. Perhaps humans have evolved categories which are structurally the most informative.

## References

- Cartwright, T. A., Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition*, 63, 121-170.
- Mintz, T. H., Newport, E. L., & Bever, T. G. (1999). The Distributional Structure of Grammatical Categories in Speech to Young Children. Ms. under review.
- Redington, M., Chater, N., & Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science*, 22, 425-469.

# Analogical priming in a word naming task

**Robert G. Morrison and Keith J. Holyoak**

University of California, Los Angeles  
Department of Psychology; Franz Hall  
Los Angeles, CA 90095-1563 USA  
(morrison@psych.ucla.edu)

**Barbara A. Spellman**

University of Virginia  
Department of Psychology; 102 Gilmer Hall  
Charlottesville, VA 22903 USA

## Introduction

Research on semantic memory has often tacitly treated semantic relations as simple conduits for spreading activation between associated object concepts, rather than as integral components of semantic organization (e.g., Quillian, 1968). Yet conceptual relations, and the role bindings they impose on the objects they relate, are central to such cognitive tasks as discourse comprehension, inference, problem solving, and analogical reasoning (see Holyoak & Thagard, 1995, for review). The present study addresses the question of whether semantic relations and their bindings can influence access to semantic memory.

## Method

The experiment we report investigated whether, and under what conditions, presenting a prime pair of words linked by one of 10 common semantic relations would facilitate processing of a target pair of words linked by the same relation. For instance, the prime pair bird/nest is bound by the semantic relation “lives in”. If bird/nest is presented as a prime pair then naming bear/cave should be faster relative to a target pair bound by a different relation, (e.g., razor/shave—“used to”). Primes and targets were presented as shown in Figure 1. In Experiment 1a participants were instructed to read each word silently as it appeared and then to say out loud the word printed in all capital letters. Naming latencies were measured from the time the second word in the target pair appeared. In Experiment 1b, participants were also instructed to “note and use” the semantic relations.

## Results and Discussion

No effect was observed when participants merely read the prime pair ( $F(1,27) < 1$ ); however, under instructions to note and use the semantic relations, participants were significantly faster at naming target pairs after same relation primes ( $M = 833$  ms) than after different relation primes ( $M = 847$  ms),  $F(1, 27) = 4.45$ ,  $p < .05$  and  $F(1, 119) = 5.50$ ,  $p < .05$  in the item analysis.

Although the full set of conditions under which analogical priming may occur remains unclear, we have shown the importance of instructions when facilitation is achieved by a single same relation pair of words. McKoon and Ratcliff (1995) have demonstrated a similar effect through the context of target words; however, it is yet unclear whether the version of the effect obtained in their study may be the re-

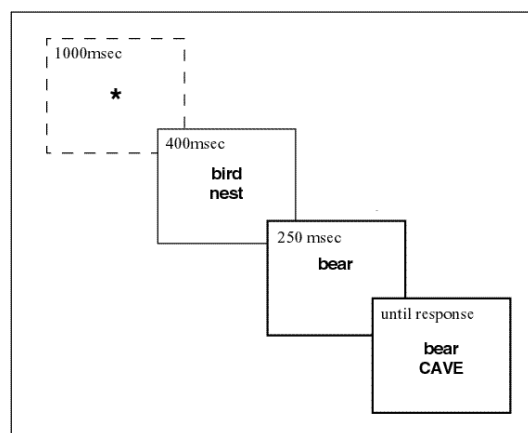


Figure 1: Analogical priming naming task

sult of an implicit strategic set similar to that imposed by our instructions. Although many questions about the nature of analogical priming remain unanswered, the phenomenon may prove central in providing theoretical linkage between basic mechanisms for accessing semantic memory and mechanisms for comprehension and reasoning.

## Acknowledgments

This research was supported by NSF Grant SBR-9729023. We thank Sondra Bland for creating a pilot version of the naming task, Krystal Long for excellent technical assistance, and Bao Truong, Ofilio Vigil, and Tannaz Sassoni for running participants.

## References

- Holyoak, K. J., & Thagard, P. (1995). Mental leaps: Analogy in creative thought. Cambridge, MA: MIT Press.
- McKoon, G., & Ratcliff, R. (1995). Conceptual combinations and relational contexts in free association and in priming in lexical decision and naming. Psychonomic Bulletin & Review, 2, 527-533.
- Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), Semantic information processing. Cambridge, MA: MIT Press.

# Finding Common Ground in Children's Referential Communication

Aparna Nadig (Aparna\_Nadig@brown.edu)

Department of Cognitive and Linguistic Sciences; Brown University Box 1978  
Providence, R.I. 02906

Julie Sedivy (Julie\_Sedivy@brown.edu)

Department of Cognitive and Linguistic Sciences; Brown University Box 1978  
Providence, R.I. 02912

For referential communication to be effective, it must be made with respect to the discourse context shared by the interlocutors--their **common ground**. It appears that young children may have particular difficulty incorporating common ground information in their production and processing, as they often fail to adapt their speech to a listener's perspective. For example, they often make ambiguous references and frequently fail to establish the antecedents of pronouns and definite noun phrases in their speech (Warden, 1976; Warren & Tate, 1992). This apparent communicative egocentrism could stem from an inability to ascertain or employ what information is shared in common ground.

Results from language processing studies suggest that even adults show evidence of difficulty with integrating common ground information. Keysar and colleagues compared two conditions which tested whether adults completely exclude information *not* shared in common ground from initial consideration. The authors found that privileged information is not completely excluded from initial consideration, and propose a two-stage model in which common ground information is used late in processing (Keysar, Barr, Balin & Paek, 1998). The present study investigated to what extent and when preschool children *do* rely on common ground information in their production and comprehension.

5 to 6 year-old children's ability to identify a unique referent with respect to common ground was tested 1) in an elicited production task, and 2) by the analysis of their eye movements, obtained from a head-mounted eye-tracking system, as they interpreted instructions in a comprehension task.

In both tasks, children viewed a vertical display containing four objects, one of which was hidden from an experimental confederate's view. Three conditions were compared: in the Contrast condition the target object and a competitor object that differed from the target with regard to a scalar feature (e.g. a big cup and a small cup, respectively) were visible to both participants; in the Contrast-Obscured condition the competitor object was available in the child's privileged view but obscured from the confederate's view; and in the No Contrast condition the competitor object was replaced by an unrelated object.

In the elicited production task children had to instruct their adult partners to pick up the target object. Children used modification in their description of the target object significantly more frequently in the Contrast condition, when both the target and competitor object were visible to both participants (requiring additional modification to distinguish between them), than in either of the other two

conditions, indicating the use of common ground in their production.

The on-line comprehension task using eye movement monitoring showed particularly striking use of common ground information. Children were instructed to pick up the target object and their eye movements were monitored as they interpreted this instruction. The description of the target object was always in the form of the head noun (e.g. the cup), regardless of condition. The eye movement data from the Contrast-Obscured condition showed no evidence of interference of the competitor object when it was hidden from the confederate's view, even from the very earliest moments of processing a target description. The time children took to identify the target object was not significantly different in the Contrast-Obscured and No Contrast(baseline) conditions. However, when the competitor object was in common ground, massive interference effects were found. Although the competitor object was visible to children in both conditions, it only impacted their processing of the instruction when it was part of the common ground information they shared with their interlocutor.

These results suggest that, in a sufficiently simple task, common ground information can be used in the earliest moments of processing, even by young children. This finding corroborates research done with adult subjects by Hanna et al. (1998) and Arnold et al. (1999), which found common ground information to be used as a partial constraint on initial interpretation.

## References

- Arnold, J. E., Trueswell, J. C. & Lawentmann, S. M. *Using Common Ground to Resolve Referential Ambiguity*. Psychonomics conference poster, Los Angeles, 1999.
- Hanna, J. E., Trueswell, J. C., Tanenhaus, M.K. & Novick, J. M. *Consulting Common Ground During Referential Interpretation*. CUNY conference proceedings, New Brunswick, 1998.
- Keysar, B., Barr, D. J., Balin, J. A. & Paek, T. S. (1998). Definite Reference and Mutual Knowledge: Process Models of Common Ground in Comprehension. *Journal of Memory and Language*, 39, 1-20.
- Warden, D. (1976). The influence of context on children's use of identifying expressions and references. *British Journal of Psychology*, 67(1), 101-112.
- Warren, A.R. & Tate, C. (1992). Egocentrism in Children's Telephone Conversations. In R. Diaz and L. Berk (Eds.), *Private Speech: from social interaction to self-regulation*. New Jersey: Lawrence Erlbaum Associates.

# If robots make choices, are they alive?: Children's judgements of the animacy of intelligent artifacts

Milena K. Nigam (mkoziol@andrew.cmu.edu)

David Klahr (klahr@cmu.edu)

Department of Psychology; Carnegie Mellon University  
Pittsburgh, PA 15213 USA

## Introduction

Much of the research on children's developing concepts about the natural world has focused on how they distinguish between living and non-living entities (Carey, 1985). In addition to the well-established cue of autonomous movement, Richards and Siegler (1986) found that six- and seven-year-olds also include mental states as attributes of living things. However, there are different types of mental states (e.g., thoughts and emotions), and in today's technological environment, where even preschoolers have experience with "intelligent" artifacts such as computers and robots, children's understanding of the complex relationship between mental states and animacy judgments remains to be explored (Turkle, 1984).

We considered three types of mental states: (a) cognition (thinking), (b) emotion (having feelings) and (c) volition (having desires/goals). We expected that children's attributions of volition and emotion would be associated with animacy judgments, whereas attributed cognition, on its own, would *not* be associated with positive animacy judgments of computers and robots. (Note that we are making no claims about the causal direction of any such associations. This issue will be explored in future research.)

## Method

We tested children in three age groups spanning the period in which adult-like judgments of animacy emerge (Carey, 1985): 14 preschoolers and kindergartners, 14 second graders, and 11 fourth graders. Children were shown color photographs of three classes of entities: (a) natural kinds (person, monkey), (b) intelligent artifacts (robot, computer), and (c) simple artifacts (doll, TV, hammer). For each entity, children were asked whether it was silly or OK to say a particular statement about that entity. (E.g., "Is it silly or OK to say: 'A robot can think.?'") (Cf. Keil, 1979.) Children were asked to make judgments concerning (a) the entity's animacy status (alive or not alive) and (b) its mental state capabilities. The presentation of mental states and animacy status was counterbalanced across entities.

## Results and Discussion

Robot was the only entity where we found a substantial variation in animacy responses. We dropped fourth graders from this analysis because they all said that a robot was not alive. The distribution of mental state attributions related to robot animacy judgments is presented in Table 1.

Table 1

Covariation Matrix for Animacy and Mental State Attributions

	<u>Cognition</u>		<u>Emotion</u>		<u>Volition</u>	
	Yes	No	Yes	No	Yes	No
<b>Alive</b>	.26	.04	.30	0	.26	.04
<b>Not Alive</b>	.18	.53	.28	.42	.11	.60

To test which mental state attributes were most predictive of positive animacy judgments, we ran a stepwise logistic regression. It revealed that volition was the strongest predictor variable (odds ratio 1.8 [95% CI 1.2 to 2.9]  $p = .004$ ) for robot animacy judgments. (This analysis yields the odds of saying that robot is alive given a judgment that it is volitional.) Note that nearly 30% of responses attributed emotion to a nonliving robot, and nearly 20% of responses attributed cognition to a nonliving robot. However, only 10% of responses attributed volition to a nonliving robot. Thus, this study reveals the emergence of children's early understanding of the nature of complex intelligent artifacts and its relation to the concept of animacy.

## Acknowledgements

This research was supported in part by a graduate research fellowship from NSF to the first author and by a grant from NICHD (HD 25211) to the second author.

## References

- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.
- Keil, F.C. (1979). *Semantic and conceptual development: An ontological perspective*. Cambridge, MA: Harvard University Press.
- Richards, D., & Siegler, R. (1986). Children's understanding of the attributes of life. *Journal of Experimental Child Psychology*, 42, 1-22.
- Turkle, S. (1984). *The second shelf: Computers and the Human Spirit*. New York: Simon and Schuster.

# Effects of Visualization on Familiar Motion Problems

Matia Okubo

matia@srt.L.u-tokyo.ac.jp

Department of Psychology

University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

## Introduction

Kaiser, Jonides, and Alexander (1986) claimed that people can reason more appropriately about the curvilinear motion problems when they are related to familiar experiences than when they are not. It is, thus, predicted that visualizing the familiar experience of the motion will lead to the correct response for the curvilinear motion problem. However, Hubbard (1996) hypothesized that the visualization strategy for the curvilinear motion problem leads to the incorrect curvilinear impetus response.

To differentiate these two opposite theoretical predictions, the effects of visualization have to be subjected to empirical test. The problem used by Kaiser et al. (1986) was modified and three instruction groups were prepared: The water group predicted a path of water spouting from a spiral tube. Besides the path prediction, the hose-analogy group was reminded of the experience of using a garden hose, and the visualization group visualized the scene in which water spouted from a garden hose.

## Method

### Participants

Eighty-four female college students without college-level physics education were randomly and evenly assigned to one of three instruction groups (i.e. water, hose-analogy, and visualization groups).

### Materials and Procedure

Each participant received a booklet, where a schematic diagram of a spiral tube and one of the three instructions were printed to describe the problem. Nine alternative paths were also printed. Among those paths, one was the correct straight path. Four were curvilinear impetus paths that curved inwardly, and the other four were centrifugal force paths that curved outwardly. Participants selected the path that matched their prediction.

## Results and Discussion

Performance significantly differed across the three groups ( $\chi^2 L(4) = 17.31, p = .002$ ). Percentage of curvilinear impetus responses in the visualization group was smaller than those in the other two groups. This finding clearly disagrees with Hubbard's hypothesis (Hubbard, 1996). The visualization and water groups responded more correctly than the hose-analogy group. Although there was little difference between the visualization and water group, the results might agree

with the prediction by Kaiser et al. (1986). Because percentage of correct responses in the visualization group was larger than those in the hose-analogy group, and it might be higher than those in previous studies which reported that 37 - 46 % of participants without formal physics training predicted the correct path for the abstract curvilinear motion problems (e.g., Kaiser et al., 1986; McCloskey & Kohl, 1983). The difference between the visualization and hose-analogy groups suggests that not recalling but visualizing the familiar experience is responsible for the effects of visualization.

In conclusion, we can say that visualizing the familiar experience leads to the correct response rather than the incorrect curvilinear impetus response for the curvilinear motion problems.

Table 1: Percentage of Participants Choosing Correct, Curvilinear Impetus, and Centrifugal Force Responses.

Instruction	Response		
	Correct	Curvilinear impetus	Centrifugal force
Water	64	18	18
Hose-analogy	36	50	14
Visualization	68	11	21

Note. There were 28 participants in each group.

## References

- Hubbard, T. L. (1996). Representational momentum, centripetal force, and curvilinear impetus. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 22, 1049-1060.
- Kaiser, M. K., Jonides, J., & Alexander, J. (1986). Intuitive reasoning about abstract and familiar physics problems. *Memory & Cognition*, 14, 308-312.
- McCloskey, M., & Kohl, D. (1983). Naive physics: The curvilinear impetus principle and its role in interactions with moving objects. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 9, 146-156.

# The Problem of Relevance in Blended Mental Spaces

David Paxman

([david\\_paxman@byu.edu](mailto:david_paxman@byu.edu))

English Department, Brigham Young University

3136 JKHB BYU, Provo, UT 84602 USA

Herbert Simon's 1994 essay "Literary Criticism: A Cognitive Approach" proposed that cognitive science could ground a more unified, less volatile method of literary study. Interpretation is to be seen as a process by which meanings are evoked in readers' minds when readers select actual meanings from among potential meanings, induce contexts, invoke archetypes, and utilize local knowledge derived from the text. Many respondents alleged that Simons merely showed how problematic was the question he thought he was answering: how do we determine relevance from among myriad potential mental representations, associations, and infinite relations among them?

I don't have an answer to this problem, but I have, by way of speculation, a suggestion that may comprise part of the solution. The brain may have a discrete number of default modes in which it processes and stores a given concept. Just as visual images are processed and stored by different parts of the brain specializing in size, color, shape, motion, and proximity, so less physical concepts may be processed and stored in a series of modes. I use the common-sense term *aspects* to name these modes.

A minimum number of aspects available for any concept would include:

- Image: the concept as icon, prototype image, or gestalt
- Agent: the concept as organism capable of action
- Structure: the concept as a form with relationships among parts
- Hierarchy: the concept's place in subordinate and superordinate classes
- Use or purpose: the ends to which the concept is applied
- Phase or stage: the concept as a specifiable part of a process
- Action: the concept as an action
- State or condition: the concept as a state of being
- Cause or effect: the concept as a result or cause of some thing or state

The concept "war," for example, can be apprehended as an image or images, as an agent, a structure with parts (conflict with combatants, etc.), a hierarchy (more specific than "conflict" but less so than "WWII"), a purpose, a phase or stage, an action, a state of being, or the cause or effect of other states. Because all aspects are potentially available when any concept occurs, any one aspect may be used as metaphor or metonymy for any other.

Something like these default aspects must exist to

explain how readily we create blends such as "boat house" and "house boat" and know which aspects of the concepts *boat* and *house* to blend in each case. Turner and Fauconnier have shown that "house boat" blends conceptual counterparts (things common to both concepts) such as place of residence, sleeping spaces, and kitchens; while "boat house" blends the same two domains by recruiting boat as an occupant of "house," which here is seen in its aspects of gestalt and purpose—a building meant to house and shelter people. To accomplish this feat, the mind must have algorithms for purpose and likely outcomes as it recruits potential elements from each domain for specific purposes.

Literary examples will be offered to show that a complex blended space can be explained as a series of concepts appearing in their relevant aspects.

## References

- Simon, H. A. (1994). Literary Criticism: A Cognitive Approach. *Stanford Humanities Review, Supplement, 4* (1), 1-26.
- Turner, M. & Fauconnier. (1995). Conceptual Integration and Formal Expression. *Metaphor and Symbolic Activity, 10* (3), 183-204.



# Interpreting Eye-Movement Protocols

**Dario D. Salvucci** (dario@cbr.com)  
Cambridge Basic Research; Four Cambridge Center  
Cambridge, MA 02142

**John R. Anderson** (ja+@cmu.edu)  
Department of Psychology; Carnegie Mellon University  
Pittsburgh, PA 15213

Eye movements reveal a great deal about our thoughts and intentions. Exploiting this benefit, researchers have utilized eye movements increasingly as a tool for understanding human behavior at a fine-grained level. However, the popularity of eye-movement data has been tempered by the difficulty of analyzing these data, which typically contain a great deal of individual variability and equipment noise. Researchers must often choose between analyzing a small number of protocols by hand or analyzing a larger number of protocols with very coarse, aggregate measures.

We have developed a class of methods that automate the analysis of eye-movement protocols (Salvucci & Anderson, 1998; Salvucci, 1999). The methods analyze, or interpret, these protocols by means of *tracing* — mapping the observed sequence of eye movements to the sequential predictions of a cognitive model. The tracing process begins by running the cognitive model and generating sequence(s) of predicted thoughts and actions. The tracing process then determines the correspondence between an observed protocol and the predicted sequence that best matches the protocol.

Our tracing methodology includes three methods of varying complexity and accuracy. The simplest method, *target tracing*, performs tracing using a sequence-matching algorithm popularized for user protocol studies (Card, Moran, & Newell, 1983). The two more sophisticated methods, *fixation* and *point tracing*, utilize hidden Markov models, powerful statistical tools that have been applied with great success in speech and handwriting recognition (see Rabiner, 1989). All three methods provide fast and accurate interpretations and are robust in the presence of noise and variability. The tracing methods have been implemented into a working system, EyeTracer, that provides an interactive environment for manipulating, replaying, viewing, and analyzing protocols.<sup>1</sup>

We have rigorously tested the tracing methods in three illustrative domains: equation solving, reading, and “eye typing”. In the equation-solving domain, we collected protocols from students solving equations of a particular form and compared the interpretations of the tracing methods to those of expert human coders. Results showed that the tracing methods interpreted the protocols as accurately as the human experts in significantly less time (at least an order of magnitude difference). We also applied the tracing methods with a “trace-based methodology” (Ritter & Larkin, 1994) to develop a cognitive model of student behavior in the task. The tracing methods facilitated both exploratory and

confirmatory analysis of the protocols and resulted in a successful model of student behavior.

In the reading domain, we evaluated the ability of the tracing methods to compare cognitive models with respect to their sequential predictions. For this purpose, we used two competing models of eye-movement control in reading, E-Z Reader 3 and E-Z Reader 5 (Reichle et al., 1998); these models produced similar predictions of non-sequential measures, but E-Z Reader 5 produced qualitatively better predictions of sequential measures. By tracing a reading data set using these models, the tracing methods provided quantitative evidence that E-Z Reader 5 was indeed the better model. The tracing methods also significantly cleaned up the data and facilitated analysis of aggregate duration and fixation probability measures.

In the “eye-typing” domain, computer users typed words by looking at letters on an on-screen keyboard. Unlike earlier eye-typing interfaces, our interface had no restrictions on how long users needed to fixate letters; this feature facilitated fast input but complicated interpretation of user eye movements. Provided with a model of user input, the tracing methods greatly facilitated data analysis and resulted in faster, more accurate user input than was possible using earlier analysis methods.

## Acknowledgments

This work is based on the first author’s dissertation and was funded in part by Office of Naval Research grant N00014-95-10223 awarded to John R. Anderson.

## References

- Card, S., Moran, T., & Newell, A. (1983). The psychology of human-computer interaction. Hillsdale, NJ: Erlbaum.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, 257-286.
- Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, 105, 125-157.
- Ritter, F. E., & Larkin, J. H. (1994). Developing process models as summaries of HCI action sequences. *Human-Computer Interaction*, 9, 345-383.
- Salvucci, D. D. (1999). Mapping eye movements to cognitive processes. Doctoral Dissertation, Department of Computer Science, Carnegie Mellon University.
- Salvucci, D. D., & Anderson, J. R. (1998). Tracing eye movement protocols with cognitive process models. In *Proceedings of the Twentieth Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

<sup>1</sup> EyeTracer is publicly available on the World Wide Web at < <http://www.cbr.com/~dario/EyeTracer> >.

# Learning to Learn by Modular Neural Networks

**Akio Sashima** (sashima@etl.go.jp)

Electrotechnical Laboratory  
1-1-4 Umezono, Tsukuba, Ibaraki, 305-8568, JAPAN

**Kazuo Hiraki** (hiraki@idea.c.u-tokyo.ac.jp)

Department of Systems Science, The University of Tokyo  
3-8-1 Komaba, Meguroku, Tokyo, 153-8902, JAPAN

## Introduction

Encountering a stream of learning tasks, humans learn not only knowledge of current task but also biases of learning future tasks. Thrun (1998) insists that modeling this human ability is one of the promising new approaches in the area of machine learning research and call this approach “Learning to Learn”(LTL).

Although some LTL algorithms have been proposed by machine learning researchers, little is known about the relation between LTL and representation of mind.

In this paper, we discuss LTL in the context of modular representation of cognitive system. We hypothesize that modules of a cognitive system are building blocks for learning new tasks (Hiraki, 1998). If each module learns a reusable basic function at the initial task, mixture of modules can learn various complex functions at future tasks. That is, generality and reusability of each module enables the ability of LTL. We implement this hypothesis using modular neural networks and examine it with a function approximation task.

## Function Approximation Task

**Functions** The networks are trained to approximate the following functions:

### Function A

$$f(x, y) = \begin{cases} \cos(x) & \text{for } y = 1.0 \\ -\cos(x) & \text{for } y = -1.0 \end{cases}$$

### Function B

$$f(x, y) = \begin{cases} |\cos(x)| & \text{for } y = 1.0 \\ -|\cos(x)| & \text{for } y = -1.0 \end{cases}$$

Where  $|x|$  represents function that computes absolute value.

**Training procedure** The training procedure is divided into two consecutive stages to examine the effects of learning function A at initial stage. We compare the approximations to function B in the following two tasks:

- **AB task** The networks learn function A first, and then learn function B.
- **BB task** The networks learn function B twice.

Training times of each stage at two tasks are the same.

**Modular Network Architecture** We implement our model using multiple forward models that are part of the architecture recently proposed by Wolpert & Kawato (1998). The networks have 2 expert modules and 1 gating module.

## Result and Discussion

By all trials (20 times), the networks at AB task can correctly approximate function B, but the networks at BB task cannot approximate it. As an analysis of output of each module, we find that the training procedure makes difference in module formation of function B.

For the BB task, one expert module captures  $|\cos(x)|$ , another captures  $-|\cos(x)|$  and the gating module switches between the output of these two modules based on  $y$ . A single module network cannot correctly approximate the absolute value function, so the networks fail to approximate function B. Alternatively, for the AB task, one expert module captures  $\cos(x)$ , another captures  $-\cos(x)$  and the gating module switches between these modules based on  $x$  and  $y$ . These expert modules were already formed during the initial stage to approximate function A. So the networks have only to learn a mixture of these modules to approximate function B.

This result shows that learning a stream of tasks with modular representation is strongly affected by task order. We consider that modular representation is one of the key factors for LTL.

We believe our model of LTL can help us to understand relation between developmental process and module formation. Karmiloff-Smith (1992) argues that humans show the developmental stages that correspond to module reformation. Humans may need each developmental stage to learn simple skills that enable more complex one for later stages. We will test for these effects in developmental tasks, such as learning arm control and eye movement.

## References

- Hiraki, K., Sashima, A., & Phillips, S. A. (1998). Maturational Biases and Encapsulation in Spatial Development, *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 1226). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press/Bradford Books.
- Thrun, S., & Pratt, L. (Eds.) (1998). *Learning to Learn*. Kluwer Academic Publishers.
- Wolpert, D.M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11, 1317-1329.

# Modeling Embodied Cognition in a Complex Real-Time Task

Michael J. Schoelles (mschoell@gmu.edu)

Wayne D. Gray (gray@gmu.edu)

Human Factors & Applied Cognition  
George Mason University  
Fairfax, VA 22030 USA

The interaction between perception and cognition is an important component of human performance in complex dynamic tasks. In time critical situations we propose that subjects develop microstrategies (Gray, Schoelles, & Fu, 1999) that manipulate these interactions to improve performance. In this paper, we report on our effort to model these interactions. The model in its current state performs a complex dynamic decision making task in a scaled world simulation of a radar operator (Argus Prime). The ultimate goal of the model is to predict changes in performance as the cognitive and perceptual workload of the task changes.

The task in the Argus Prime experimental environment requires a mix of perceptual and cognitive actions. The task involves four subtasks. For target selection, the user attends to icons on the screen (perception), decides to process an icon (cognition), and selects it (motor). In information retrieval the user reads the raw data values for this object (perception). Score calculation entails mapping raw data to target score (cognition), mapping score to threat value (cognition), selecting a threat value (perception and motor), and entering the decision (motor). Finally, feedback processing consists of perceiving feedback (perception) and processing the feedback (cognition). As this brief task analysis illustrates, each subtask combines cognitive, perceptual, and motor operators. Less apparent from this overview is when the actions can proceed in parallel and when they constrain each other.

The cognitive architecture on which the model is built is ACT-R/PM. The ACT-R/PM architecture combines ACT-R's theory of cognition (Anderson & Lebière, 1998) with modal theories of visual attention (Anderson, Matessa, & Lebière, 1997) and motor movement (Kieras & Meyer, 1997). ACT-R/PM explicitly specifies timing information for all three processes as well as parallelism between them. The software architecture facilitates extensions beyond the modal theory of visual attention and motor movements. Our current efforts are taking advantage of this architectural feature to match the modeling effort with the issues raised by the analytic and empirical research in the Argus effort. In particular, we are working on three extensions, one for eye movements, tracking objects, and perceptual support for working memory.

*Eye Movements.* For the analysis of the eye tracking data shows we have incorporated Eye Movements and Movements of Attention extension (EMMA) (Salvucci, 2000) into the model. EMMA provides multiple eye movements per attention shift and provides encoding time for objects based on frequency of attending to the same object and the object's distance or eccentricity from the current point-of-gaze

*Tracking Objects.* We are currently incorporating into the target selection task a theory of multiple object tracking. Sears and Pylyshyn (in press) have applied the FINST model to multiple object tracking. This theory hypothesizes a stimulus driven mechanism that individuates objects in the environment by pointing to them; that is, assigning an index. The indexing precedes object identification and the index remains bound to the object even if characteristics of the object change. In particular, if the location of the object changes continuously then the index can still be used to point to the object. Attention can be directed to the *object* with the index as its argument. The dynamic environment of Argus Prime seems well suited to modeling this theory as a possible mechanism used by subjects in the target selection phase.

*Perceptual Support for Working Memory.* ACT-R/PM provides for both external and internal sources of activation for memory retrieval. Currently the amount of external source activation is a free parameter. Our current efforts are involved with quantifying *how* the level of external source activation varies with task conditions and what microstrategies subjects develop to optimize retrievals by controlling the mix of internal and external source activation.

## Acknowledgements

The work reported was supported by a grant from the Air Force Office of Scientific Research AFOSR#F49620-97-1-0353.

## References

- Anderson, J. R., & Lebière, C. (Eds.). (1998). *Atomic components of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., Matessa, M., & Lebière, C. (1997). ACT-R: A theory of higher-level cognition and its relation to visual attention. *Human-Computer Interaction*, 12(4), 439-462.
- Gray, W. D., Schoelles, M. J., & Fu, W.-t. (1999). Modeling microstrategies in a continuous dynamic task. *Manuscript submitted for publication*.
- Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction*, 12(4), 391-438.
- Sears, C. R., & Pylyshyn, Z. W. (in press). Multiple object tracking and attentional processing. *Canadian Journal of Experimental Psychology*.
- Salvucci, D. D. (2000). A model of eye movements and visual attention. In Proceedings of the International Conference on Cognitive Modeling (pp. 252-259). Veenendaal, The Netherlands: Universal Press.

# Does Human Memory Reflect the Environment of Early Hominids?

Lael Schooler  
(ljs24@psu.edu)  
Department of Psychology  
Pennsylvania State University  
University Park, PA 16802

Juan Carlos Serio Silva  
(serioju@ecologia.edu.mx)  
Departamento de Ecología Vegetal  
Instituto de Ecología A.C.  
km 2.5 antigua carretera a Coatepec  
ap 63 cp 91000  
Xalapa, Veracruz, Mexico

Ramon Rhine  
Department of Psychology  
University of California  
Irvine, CA 92521

The rational analysis of memory (Anderson, 1990) proposes that human memory has evolved to cope optimally with the informational demands that the environment places on people. We have shown that human memory performance reflects patterns with which environmental stimuli (e.g., words) occur and reoccur (Anderson, & Schooler, 1991; Schooler & Anderson, 1997). Because the human cognitive system did not evolve in our modern environment, Shettleworth (1998) has questioned the validity of our previous analyses: "to consider Anderson and Schooler's results relevant to pressures that have caused memory to evolve, one has to assume that headlines in late 20<sup>th</sup>-century newspapers reflect a general and enduring property of events in the world."

It is, of course, impossible to study the informational demands that the environment placed on early hominids. About the best we can do is study the informational demands placed on animals whose current ecological niches share something in common with the ecological niches in which hominids evolved. The question, then, is which animals fill the appropriate ecological niches. Dart (1926) argued that time spent on the savanna was critical in the development of intelligence. More recently Milton (1981) has pointed out that hominids evolved first in tropical forests, where "the extreme diversity of plant foods in tropical forests and the manner in which they are distributed in space and time have been a major selective force in the development of advanced cerebral complexity in higher primates." She argues that "to understand the origins of mental complexity, one must look not only at life in the savannas but also life in tropical forests." Thus, studying how primates move through forests and savannas represent good starting points for understanding the informational demands that shaped early hominid evolution.

We have analyzed existing data on the ranging patterns of howler monkeys through forests, and baboons through savanna. Serio-Silva, using the focal animal method, recorded the identification numbers of the trees the howlers were visiting. Rhine's group used the focal animal method as well. They recorded the positions of the baboons in terms of quadrats measuring 720 m<sup>2</sup>. It appears that the visitation patterns of howlers and baboons match up with the statistics of the modern environments. Our analyses show that there are statistical properties shared among domains as diverse as

word usage in the New York Times and the ranging patterns of howler monkeys in trees. These analyses suggest that there are "general and enduring" properties shared between modern and early hominid environments.

Beyond demonstrating the feasibility of performing environmental analyses of primates in natural environments, these analyses have implications for the kinds of memory mechanisms to explore. By exploiting enduring statistical properties of the environment, the memory system could rely on mechanisms that need only infer the *parameters* of known functions. This is a far simpler task than trying to infer what these functions might be. This supports a general-purpose memory system that just estimates parameters for various memory traces.

Some have argued that such general-purpose mechanisms are unlikely to evolve. "General-purpose mechanisms can't solve most adaptive problems at all, and in those cases where one could, a specialized mechanism is likely to solve it more efficiently." Cosmides & Tooby (1994). Implicit in their argument is the assumption that diverse domains do not share fundamental features in common. To the extent that a variety of domains do share statistical properties a general-purpose memory system would be efficient and evolvable.

## References

- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2, 396-408.
- Cosmides, L. & Tooby, J. (1994). Beyond intuition and instinct blindness: toward an evolutionarily rigorous cognitive science. *Cognition*, 50, 41-77.
- Dart, R. (1926). Taungs and its significance. *Natural History*, 26, 315-327.
- Milton, K. (1981). Distribution patterns of tropical plant foods as an evolutionary stimulus to primate mental development. *American Anthropologist*, 83, 534-548.
- Schooler, L. J. & Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology*, 32.
- Shettleworth, S. J. (1998). *Cognition, Evolution, and Behavior*. Oxford: Oxford University Press.

# **Knowledge Construction Links: Cues and Trajectories as Prior Experience and Knowledge**

**Kathy L. Schuh** ([kathy-schuh@uiowa.edu](mailto:kathy-schuh@uiowa.edu))  
Instructional Design and Technology, 361 Lindquist Center  
Iowa City, IA 52242 USA

## **Background and Purpose**

The constructivist perspective asserts that new knowledge is based on individuals' prior experience and knowledge, and is, therefore, idiosyncratic. Although constructivist instructional strategies (e.g., problem based learning, cognitive apprenticeship, and anchored instruction) have shown promise in classrooms, little research is available on why these methods are helpful at a cognitive level (i.e., how and why learners construct their unique knowledge). This study begins to address this foundational omission using an emergent research methodology to identify the use of trajectories and cues in knowledge construction.

## **Methodology**

This instrumental collective case study reports incidents of learners' use of prior knowledge and experience into their new learning opportunities as a component of describing knowledge from a constructivist perspective. Participants were students in three sixth-grade classrooms (N=74) that differed in learner-centeredness, a characteristic of constructivist classrooms. Learner-centeredness was determined by students' perceptions as measured through the Learner-Centered Battery. Data were gathered through observation, interviews, and a writing activity during subject units that spanned two to seven weeks. Knowledge was operationalized as links (verbally or in writing) through which learners included information that was often tangential to the current topic (i.e., prior experiences and knowledge). Interview participants were selected based on these comments or through random selection. In the open-ended independent writing activity, students were asked to begin their writing with the subject matter topic but also told that they could follow tangential connections.

Given the instrumental nature of the study, the knowledge construction links were first identified. Further analysis proceeded as in qualitative studies, seeking emergent trends in the data, i.e., characteristics of the knowledge construction links as well as the environment in which they were embedded.

## **Findings and Interpretation**

Emerging in the data were cues, those stimuli that prompted learners' to link new information with their prior experience and knowledge. Cues were singular or multiple and often led to tangential comments or questions. For example, a student described the role of vitamin C (the focus of her independent science inquiry project) in a discussion of the Middle Ages and "citrus fruit" (the cue). Once a learner encounters a cue, a trajectory may follow. Trajectories were described by the features of the prior knowledge. Given the richness of an individual's reconstructions, trajectories often contained multiple types, in particular, the type of experience and the characters who were involved.

Ten cue types (sounds like, looks like, feels like, is a, same word but different concept, same concept but different context, same concept and same context but different content, same concept and same context with same content, different concept within same context, series, and complex relationships) and ten trajectory types (acting, general acting, future, and operative experiences, family, friends, school, society, media, and affect/emotion) were identified in this data set.

These constructs captured a view of knowledge that focused on the uniqueness of knowledge: the potential and necessity of considering knowledge within learning situations as unbounded and freely crossing domains and contexts, and dynamic by necessity given each learners' prior experience and knowledge. Further, these constructs, although occurring in all three classrooms, were fostered in the most learner-centered classroom, thus providing exploratory explanations of this fundamental learning process for constructivist instructional strategies (i.e., why these strategies foster learning).

These constructs within the construction process as it occurred in these classrooms provide a variety of opportunities for describing knowledge and knowing. In this poster session I seek opportunities for collaboration with other cognitive scientists to begin to explore how these constructs can be modeled.

# A Delay-dependent Switch in the Information Children Use to Remember Locations

Anne R. Schutte (Anne-Schutte@uiowa.edu)  
Department of Psychology, E11 Seashore Hall  
Iowa City, IA 52246

John P. Spencer (John-Spencer@uiowa.edu)  
Department of Psychology, E11 Seashore Hall  
Iowa City, IA 52246

## Abstract

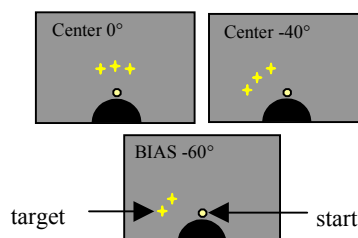
Several types of information can be used to remember the location of an object over short-term delays. The current study looked at how young children integrate three specific types of information over delays—metric information (i.e., direction and distance), reference locations (i.e., landmarks), and longer-term memories of where objects have been found in the past. Three-year-olds pointed to remembered targets in a large, homogeneous task space. The layout of the targets and how often each target appeared was varied across conditions. Three-year-olds' responses were biased toward the center of the task space and toward an average remembered location, and these biases increased as delays increased. In addition, the bias toward the average remembered location was stronger when the memory of a single location was differentially strengthened.

## Introduction

There are many ways to remember the location of hidden objects. For example, the location of a set of car keys might be remembered as being “on the desk”, in the upper left corner of the desk, or, more specifically, a few inches from the left edge. Evidence suggests that young children use three specific types of information: metric information (i.e., direction and distance), reference locations (i.e., landmarks), and longer-term memories of where objects have been found in the past (Huttenlocher, Newcombe, & Sandberg, 1994; Smith, Thelen, Titzer, & McLin, 1999). Here, we investigated how young children integrate these three types of information in memory during delays.

## Method

Thirty-six to 40-month-olds were asked to remember the location of small, spaceship-shaped lights on a large table with no salient landmarks in the task space. On each trial, a marker was moved to a start location. Then a spaceship appeared for 2s and disappeared. This was followed by a delay of 0, 5, or 10s, after which participants heard a go signal instructing them to move the marker to the



**Figure 1.** Examples of the target layouts on the table top. The child stood within the black arc.

remembered spaceship location. In each condition, targets were separated by 20°, but the number and layout of the targets was varied across conditions. In the Center 0° condition, three targets were used. These targets were positioned symmetrically with respect to the midline of the table (see Figure 1). In the Center 40° conditions, three targets on the same half (right or left) of the table were used, with the center target at 40° or -40° (see Figure 1). In the Bias 60° conditions, the participants moved to two possible targets located at 40° and 60° (or -40°, -60°). Participants moved to the 60° (-60°) target twice as often as the 40° (-40°) target to differentially strengthen this location in memory.

## Results

As the delay increased in the Center 0° condition, participants made larger directional errors toward the midline of the table when moving to the left and right targets. In the Center 40° conditions, three-year-olds' responses to the 60° and -60° targets were biased inward, toward the midline of the table, but responses to the ±20° and ±40° targets were not significantly biased. Data from the BIAS conditions clarified why these responses were not biased towards midline. In the BIAS conditions, responses to the non-biased (±40°) targets were pulled toward the biased (±60°) targets over delays. Thus, memory responses are pulled towards two types of information—the midline of the table and a longer-term memory of an average remembered target location.

## Discussion

Results from the present study demonstrate that there are systematic delay-dependant biases in how young children maintain location information in memory. As delays increase, children's memory responses are biased towards reference axes—the midline of the table—and towards a longer-term memory of previously moved-to locations. Three-year-olds' resultant memory errors depend critically on the delay duration and the relative strength of each type of information.

## References

- Huttenlocher, J., Newcombe, N., & Sandberg, E. H. (1994). The coding of spatial location in young children. *Cognitive Psychology*, 27, 115-147.
- Smith, L. B., Thelen, E., Titzer, R., & McLin, D. (1999). Knowing in the context of acting: The task dynamics of the A-Not-B error. *Psychological Review*, 106, 235-260.

## 19-Month-Olds' Sensitivity to Negation/Tense Dependencies

Melanie Soderstrom (melsod@jhu.edu)

Johns Hopkins University/Department of Psychology  
3400 N. Charles St. Baltimore MD 21218

Peter Jusczyk (jusczyk@jhu.edu)

Johns Hopkins University/Department of Psychology  
3400 N. Charles St. Baltimore MD 21218

Kenneth Wexler (wexler@psyche.mit.edu)

Massachusetts Institute of Technology/Department of Brain and Cognitive Sciences  
77 Massachusetts Ave. Cambridge MA 02139

Recent comprehension studies have shown that infants have early knowledge of adult syntactic relationships long before they are capable of demonstrating this knowledge in productive speech (e.g. Santelmann & Jusczyk, 1998). The current study addresses the relationship in infant grammar between negation and tense in two related contexts - the difference between adverbs and negation in their effect on the placement of tense marking, and the connection between negation and the presence or absence of tense marking.

In English, tense markings are found before a negation, but after an adverb. For instance, compare the following sentences:

- 1) Mary never goes to the store.
- 2) \*Mary not goes to the store.
- 3) \*Mary does never go to the store.
- 4) Mary does not go to the store.

Harris and Wexler (1996) showed that the productions of children are consistent with this adult pattern as early as 1.5 years old. In Experiment 1, the Headturn Preference Procedure (HPP) was used to determine whether the preference patterns 19-month olds follow this same pattern. Infants were tested on two sets of passages that were produced using synthesized speech (Dectalk). Both sets contained sentences with verbs in 3rd person singular, present tense. In the grammatical set, the verb was preceded with "never" (see sentence 1). In the ungrammatical set, the verb was preceded with "not" (see sentence 2). Passages were played in random succession to either side of a testing booth, with playing time for each trial contingent on the infant's interest as measured by orientation of gaze to a paired light stimulus. The dependent measure was total orientation time to the paired side light. Mean scores across trials were calculated for the grammatical passages and ungrammatical passages for each infant.

Twenty-two out of 28 infants oriented longer to the grammatical passages than the ungrammatical passages. The overall mean scores were 8.5 s for the grammatical passages, and 6.8 s for the ungrammatical passages, with  $p = .028$ . Overall, these data support the notion that 19-month-olds are sensitive to the differences between negation and negative adverbs.

One striking feature of children's early production is the optional use of infinitival (not tense-marked) forms of verbs

in contexts where a tensed verb is used by adults, often referred to as the Optional Infinitive (OI) stage. So far the evidence for this phenomenon in normal acquisition is only productive in nature (but see Rice et al. (1999) for comprehension evidence of OI in children with SLI and for normal older children). Experiment 2 compared 19-month olds' preference for passages containing sentences like that in (2) with similar sentences lacking the tense marking:

- 2) \*Mary not goes to the store.
- 5) ?Mary not go to the store.

Both of these sentences are ungrammatical for adults. However, Wexler's (1994) analysis of OI productions predicts that only sentence 2 is ungrammatical for children in the OI stage, while sentence 5 is treated as grammatical.

Surprisingly, 20 out of 28 infants oriented longer to the tense-marked passages (2) than the unmarked passages (5). The overall mean scores were 7.8 s for the tense-marked passages, and 9.5 s for the unmarked passages, with  $p = .027$ . This finding is not predicted by current production-based theories of acquisition.

One explanation for the unexpected finding is that infants are not attending to the "not" in this context, although they did detect the not/never distinction in Experiment 1. We are currently exploring this possibility using nonsense words before the main verb.

### References:

- Harris, T. & Wexler, K. (1996). The Optional-Infinitive Stage in Child English: Evidence from negation. Harald Clahsen (ed.), *Generative Perspective on Language Acquisition*, John Benjamins B.V.
- Rice, M., Wexler, K., & Redmond, S. (1999). Grammaticality Judgements of an Extended Optional Infinitive Grammar: Evidence from English-speaking children with specific language impairment. *Journal of Speech, Language and Hearing Research*, 42(4), 943-961.
- Santelmann, L. & Jusczyk, P. (1998). Sensitivity to Discontinuous Dependencies in Language Learners: Evidence for limitations in processing space. *Cognition*, 69, 105-134.
- Wexler, K. (1994). *Optional Infinitives, Head Movement and the Economy of Derivations*. David Lightfoot & Norbert Hornstein (eds.) *Verb Movement*. Cambridge: Cambridge University Press.

## Scaling and Testing for Non-Euclidean Spaces

Jesse Spencer-Smith (jbspence@indiana.edu)

Department of Psychology; Indiana University; 1101 E. 10th St.  
Bloomington, IN 47405-7007 USA

Most current models or theories that posit the existence of geometric or spatial representations (*e.g.*, Valentine, 1991), or that rely on representations generated by multidimensional scaling such as the Generalized Context Model (Nosofsky, 1984), assume that the spaces in which the representations are embedded are Euclidean, or are endowed with a Minkowski power metric. Although the need for investigation of more general spaces was noted as early as 1964 (Shepard, 1964), and has been argued more recently (Townsend and Thomas, 1993), non-Euclidean spaces have been largely overlooked. A natural generalization of Euclidean space are the Riemannian spaces. The sphere (with distances measured on the surface) is an example of a Riemannian space. In the current investigation, qualitative and quantitative tests to uncover properties of perceptual spaces are developed and tested.

Locally, Riemannian spaces are well-approximated by Euclidean spaces. Accordingly, for sufficiently-restricted stimulus sets, non-Euclidean properties of the spaces (such as curvature) may not become evident. If points on a sphere are sufficiently close, the distances between those points can be approximated by assuming the points lie on a plane. While cities in the continental U.S. can be approximated as lying on a plane, regular discrepancies are apparent. Approximating cities in the western hemisphere as lying in a plane results in far greater errors, while an approximation of the cities around the globe in a similar manner would misrepresents fundamental properties of the space. Between antipodal points on a sphere, for example, there exist not one shortest path, but infinitely many.

A metric multidimensional scaling tool which assumes constant-curvature Riemannian spaces (such as the sphere, pseudosphere, or plane) based on work by Lindman and Caelli (1972; see also Indow, 1982) is implemented in Matlab and is applied to new and existing data. Qualitative tests for curvature are developed and demonstrated, and applied to new and existing data.

### Critiques of Geometrical Models

Some of the most insightful and widely-cited critiques of geometric models of similarity (*e.g.* Gati and Tversky, 1982; Beals, Krantz and Tversky, 1968) address spaces which are endowed with Minkowski power metrics, and do not apply to spaces endowed with Riemannian metrics. The specific tests which fail with certain Riemannian spaces are detailed.

### Acknowledgments

James T. Townsend, Rudy Professor of Psychology, Indiana University provided invaluable advice during the development of this work. The work has also benefited greatly from conversations with Bruce Solomon of the Department of Mathematics, Indiana University.

### References

- Beals, R., Krantz, D. H. & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review*, 75, (2), 127-142.
- Gati, I. & Tversky, A. (1982). Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 8, (2), 325-340.
- Indow, T. (1982). An approach to geometry of visual space with no a priori mapping functions: Multidimensional mapping according to Riemannian metrics. *Journal of Mathematical Psychology*, 26, (3), 204-236.
- Lindman, H. & Caelli, T. (1972). Constant curvature Riemannian scaling. *Journal of Mathematical Psychology*, 17 (2), 89-109.
- Nosofsky, R. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, (1), 104-114.
- Shepard, R. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1, 54-87.
- Townsend, J. T. & Thomas, R. (1993). On the need for a general quantitative theory of pattern similarity. In S. C. Masin (Ed), *Foundations of perceptual theory. Advances in psychology, Vol. 99*. Amsterdam, Netherlands: North-Holland/Elsevier Science Publishers
- Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 43A,(2), 161-204.



# Variation in Children's Word Production: Can 'Competence' Models deal with young Children's Truncation Patterns?

Helena Taelman  
(taelman@uia.ac.be)

Steven Gillis  
(gillis@uia.ac.be)

University of Antwerp - CNTS, Universiteitsplein 1,  
2610 Wilrijk, Belgium

Young children (often) truncate words: they omit whole syllables from multisyllabic words, as exemplified in (1):

(1) elephant /olifAnt/ [olwant] (Maarten, 1;10.19)

The truncation patterns have been extensively studied: most existing models account for truncations in terms of children's linguistic (i.e. prosodic) competence (i.a. Fikkert, 1994; Demuth, 1995; Gerken, 1996; Pater, 1997; Bernhardt & Stemberger, 1998). These models make two crucial predictions:

1. Truncation patterns are explained as a way to accommodate words into prosodic templates, which are determined by children's (limited) knowledge of the prosodic regularities of the language. The initial rhythmic template is defined as a trochaic foot. Hence, early truncations are considered to be adaptations of words to the trochaic template (Gerken, 1996).

2. Development is conceptualized as a stage-wise progression, which is determined by an elaboration of children's knowledge of the prosodic rules of the language (Fikkert, 1994).

Although metrical competence models have received empirical support (i.a. Fikkert, 1994), a comprehensive test with a large corpus of child language data is currently lacking so that the breadth and the accuracy of the metrical competence accounts of children's truncations still need to be determined.

We present a naturalistic, longitudinal, observational case study of a Dutch speaking boy (age 1;8.29 – 1;11.15). The corpus (available through CHILDES) consists of 19,960 tokens. On the basis of a fine-grained quantitative and qualitative analysis of this corpus, we will challenge the two predictions outlined above:

1. A significant portion of the child's word productions *cannot* be explained as accommodations to a (trochaic) rhythmic template. The relevant data consist of (a) truncations which result in iambic production forms, and (b) truncations of trochaic words.

We identified a number of non-prosodic factors which determine truncations, viz. segmental factors (deemed irrelevant in existing models) and 'performance' factors such as imitation (an interactional influence) and utterance length (a processing factor).

2. A stage-wise progression model is untenable because of (a) inter-word variability (contrary to the predictions, words

with the same prosodic pattern do not evolve concurrently: different truncation patterns are found at the same time) and (b) intra-word variability (contrary to the predictions, words show within-word inconsistencies: correct and various truncated variants of the same word coexist).

We identified a number of non-prosodic factors which determine the observed patterns: i.a. word age, frequency in the input, frequency in the child's own production, and truncation rate.

We conclude that current 'competence' models are unable to deal with the variations in children's actual production data and that an alternative model is called for in which the non-prosodic 'performance' factors identified in this study can be accommodated.

## References

- Bernhardt, B., & Stemberger, J. (1998). *Handbook of phonological development. From the perspective of constraint-based nonlinear phonology*. San Diego: Academic Press.
- Demuth, K. (1995). Markedness and the Development of Prosodic Structure. *NELS 27*, Amherst, MA: GLSA. Available from <http://ruccs.rutgers.edu/roa.html>.
- Fikkert, P. (1994). *On the acquisition of prosodic structure*. Doctoral Dissertation, Rijksuniversiteit Leiden.
- Gerken, L. (1996). Prosodic structure in young children's language production. *Language*, 72, 683-712.
- Pater, J. (1997). Minimal violation and phonological development. *Language Acquisition*, 6(3), 201-253.

# Is Musical Ability Related to the Prosody Learning of Second Language?

Akihiro Tanaka (tanaka@srt.L.u-tokyo.ac.jp)

Yohtaro Takano (takano@L.u-tokyo.ac.jp)

Department of Psychology, University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan

Some recent studies have reported that phonological ability for second language (L2) is correlated with phonological loop capacity (e.g. Baddeley, Gathercole, & Papagno, 1998). In addition, Slevc & Miyake (Manuscript in preparation) reported that adult learners' L2 phonological ability is correlated with their musical ability.

However, the previous studies have dealt only with phonology. The purpose of this study is to examine the prosodic aspect, especially the intonation, and its relations to musical and verbal memory abilities in L2 learning. We use Chinese language as L2 because it has an intonational property known as "four tones".

## Method

### Participants

The participants were 35 high school and undergraduate students in Japan. None of them had ever learned Chinese.

### Materials and Procedures

The experiment consisted of three parts: musical ability test, verbal memory ability tests, and Chinese learning session.

**Musical Ability Test** The frequency difference limen for pure tone was recorded. Though this pitch discrimination ability is only a part of musical ability, we refer to this measure as musical ability for convenience.

**Verbal Memory Ability Tests** Two tests were used to measure verbal memory ability. Reading span (RS) was measured by the Japanese version of the reading span task; and letter span (LS) was measured by the letter span task.

**Chinese Learning Session** Session consisted of five blocks. Each block had a learning phase and a test phase. In the learning phases, the speech sound of the Chinese words and the Japanese equivalents were presented. In the test phases,

a target and a distractor were auditorially presented after a Japanese word was presented. Participants' task was to choose a correct word in 2-alternative forced-choice form. The distractor was one of the following words: phonologically-changed words, words differing in intonation, or other words presented in learning phases. The scores for the trials with these distractor words were recorded as phonological, prosodic, and associative scores, respectively.

## Results and Discussion

Table 1 shows the correlations between the memory ability measures and the mean scores of the five test phases (from 1<sup>st</sup> through 5<sup>th</sup> blocks) in the Chinese learning session. The prosodic score significantly correlated only with RS, while the associative and the phonological scores correlated with both LS and RS. Performance in the RS task is considered to be closely related to efficiency of the central construct of working memory. Therefore, these results imply that prosody learning is free from phonological loop capacity and is related to the central construct.

Table 2 shows the improvements (correct rate of 5<sup>th</sup> block – that of 1<sup>st</sup> block) of each scores for high and low musical ability groups. There was a significant difference only in the improvements of the prosodic scores between high and low groups ( $F(1,33)=7.39, p<.05$ ). This result can be interpreted that the learners with high musical ability are able to analyze the intonational feature with higher accuracy.

In sum, the results imply that when we learn the auditory features of L2, the prosodic features are learned through musical ability, while the phonological features are learned through verbal memory ability.

## References

- Baddeley, A. D., Gathercole, S. E. & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, 41, 78-104.
- Slevc, B & Miyake, A. (Manuscript in preparation). Individual differences in second language proficiency: Does "Good Ear" really matter?

Table 1: Correlations of mean scores for all the Chinese learning session with memory ability measures

Score Type	Memory Ability	
	LS	RS
Phonological	.40*	.49**
Prosodic	.29	.58**
Associative	.52**	.58**

Note. \*  $p<.05$  \*\*  $p<.01$

Table 2: Improvements in the Chinese scores for high and low musical ability groups

Score Type (improvement)	Musical Ability	
	High	Low
Phonological	9.38	5.88
Prosodic	19.79	5.88
Associative	9.03	8.82

## **Main Idea Identification: A Functional Imaging Study of a Complex Cognitive Process**

Lêda Maria Braga Tomitch ([leda@andrew.cmu.edu](mailto:leda@andrew.cmu.edu))

Universidade Federal de Santa Catarina

Departamento de Língua e Literaturas Estrangeiras

Campus Universitario-Florianopolis-SC 88010-970 BRASIL

Marcel Adam Just ([just+@cmu.edu](mailto:just+@cmu.edu))

Patricia A. Carpenter ([carpenter+@cmu.edu](mailto:carpenter+@cmu.edu))

Carnegie Mellon University

Center for Cognitive Brain Imaging

5000 Forbes Ave Pittsburgh-PA 15213 USA

Main idea identification is at the very heart of human thinking, being a skill required in everyday situations such as reading a message, interpreting an interlocutor's utterance, listening to the news and attending a lecture. It is part of the human nature to try to integrate incoming information and build a macrostructure containing the main points of the input, so that this information can be more easily stored in memory and retrieved when needed. Despite its importance in human interaction, the process of main idea identification is yet little understood.

Cognitive brain imaging has provided researchers new possibilities for trying to unravel what happens in the human brain during the performance of various complex tasks. This study uses fMRI to investigate the amount of brain activation in a set of cortical areas in the task of main idea identification. Readers were assigned to two types of reading situations, the difficulty of processing being manipulated as follows: in an easier condition, the passages contained the main idea in an introductory topic sentence, followed by two sentences whose content was difficult to interpret in the absence of the topic setting introductory sentence. In a more difficult condition, the

two such sentences occurred at the beginning of the passage, and the topic sentence occurred last. The greater cognitive complexity in processing the two abstract sentences prior to knowing the topic was expected to translate into an increase in brain activation in the right hemisphere for the hard condition.

Results indicate that the complex task of main idea identification is associated with increased neural activity in a range of brain regions of both hemispheres, including the temporal lobe, the extrastriate cortex, the parietal lobule and the inferior frontal gyrus, regardless of the position of the main idea in the paragraph. Furthermore, particularly prominent activity is found in the temporal regions of both cerebral hemispheres when compared to the other areas.

---

This work was supported by grant BEX0300/99-3 from CAPES-Brasilia-Brasil, grant MH29617 from the National Institute of Mental Health and grant PO1NS35949 from the National Institute for Neurological Disorders and Stroke.

# Schema Acquisition and Solution Strategy in Statistics Problem Solving

David Trumpower (dtrumpow@unm.edu)  
Department of Psychology; University of New Mexico  
Albuquerque, NM 87131 USA

## Introduction

Research on multistep problem solving in knowledge-rich domains, such as physics and mathematics, has revealed several differences between experts and novices. Experts tend to classify problems according to abstract principles useful for their solution (Chi, Feltovich, & Glaser, 1981) and solve problems using a forward-working strategy (Simon & Simon, 1978), whereas novices tend to classify problems based on surface features and solve problems using a backwards-chained strategy. To explain these findings it is generally posited that experts possess domain-specific knowledge in the form of schemas. However, schemas are theoretical constructs inferred from the phenomena they are used to explain. It is proposed here that empirically derived knowledge representations using the Pathfinder scaling algorithm can provide a more direct observation of schema acquisition associated with the attainment of expertise. Pathfinder operates on proximity data to provide a network representation with the most efficient connections between concepts, and has been used as a valid assessment of classroom learning (Schvaneveldt, 1990).

## Methods

Twenty-eight students enrolled at UNM served as participants. They were asked to think aloud as they solved 17 statistics word problems (13 training, 4 test problems). All could be solved using the following equations:  $df_B = a - 1$ ,  $MS_B = SS_B / df_B$ ,  $F = MS_B / MS_W$ . Participants' solutions and verbal protocols were recorded. After solving all problems, participants rated the relatedness of all pairwise combinations of the six concepts contained in the equations above.

## Results and Discussion

Problem solutions were analyzed to determine the strategy used on each of the test problems, and relatedness ratings were submitted to Pathfinder to derive visual representations of participant's acquired knowledge structures.

In all of the problems, participants were given two of the following values and asked to solve for the other:  $a$ ,  $MS_B$ ,  $SS_B$ . Thus, all problems could be solved by first calculating  $df_B$  and then calculating the goal value. A schema useful for solving these problems, then, would involve the relationships between  $df_B$  and each of  $a$ ,  $MS_B$ , and  $SS_B$ . This "df schema" can be seen in the knowledge structures as links between those concepts (see Figure 1). Participant's knowledge structures were analyzed for the presence of this schema.

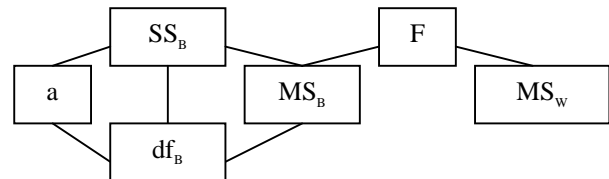


Figure 1: Sample knowledge representation

If schemas defined in this manner are associated with expert-like problem solution, then participants that possess the df schema should be more likely to solve problems in a forward manner. A solution was considered forward if the subgoal ( $df_B$ ) was calculated before consideration of an equation containing the goal; otherwise it was considered backward. Overall, ten participants solved all of the test problems using a forward strategy. Seven of these ten possessed the schema. Of seven participants, on the other hand, who solved one or fewer test problems using a forward strategy, only one possessed the schema. This discrepancy is found to be significant by a Fisher's exact test,  $p = .05$ .

In the final test problem, participants were given values for  $MS_W$ ,  $F$  in addition to  $a$ ,  $SS_B$  and asked to find  $MS_B$  (Note that participants were not trained on this type problem). Thus, they could use the forward solution described above, or they could work backwards using the equation  $F = MS_B / MS_W$  to solve for the goal in one step. It was hypothesized that participants possessing the df schema would use the former strategy while those that did not would use the latter. Nine of 13 participants possessing the df schema did solve the final problem using the forward strategy, while only three of 15 participants that did not possess the schema did so. This discrepancy is also significant,  $p = .02$ .

These results show that Pathfinder derived representations can reveal the acquisition of schemas that guide expert-like solution of statistics problems. In the absence of these schemas, problem solvers tend to rely on backwards-chained strategies, as predicted.

## References

- Chi, M. T. A., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152.
- Schvaneveldt, R. W. (Ed.). (1990). *Pathfinder associative networks: Studies in knowledge organization*. Norwood, NJ: Ablex.
- Simon, D. P., & Simon, H. A. (1978). Individual difference in solving physics problems. In R. Siegler (Ed.), *Children's Thinking: What Develops?* Hillsdale, NJ: Erlbaum.

# Management of Multiple Goals on the basis of Situational Urgency

**Takafumi Tsuchiya (tsuchiya@scs.chukyo-u.ac.jp)**

School of Computer and Cognitive Sciences, Chukyo University,  
101 Tokodate Kaizu Toyota 470-0393 Japan

The functional studies have considered human beings with multiple goals as efficient problem solving systems. While this approach revealed the situated nature of cognitive architecture, it still leaves out some important issues concerning the everyday problem solving. One such issue is that of time, and another is that of the subjective values assigned to achieving each goal. In order to deal with the multiple goals, a person ought to be efficient in setting, concentrating, suspending, discarding, and achieving some of possible goals in accordance with the person's cognitive appraisal of the urgency with which each goal presents itself.

The urgency of a goal is an important situational cognition made by a problem solving agent with a limited temporal resource. When it manages multiple goals and may achieve some of possible goals, it should appraise the subjective value of each goal to be gained. Then it would make an effort to succeed in the achievement of the important goal, and allocate its own time over activities to do so. The urgency of each goal at a given time is defined by the following three parameter values: the subjective value to be lost if the goal is not achieved, the subjective probability of achieving the goal, and the available time for doing accomplishing the goal-achievement action.

The purpose of this research is to design an autonomous agent that is required to set and achieve multiple goals with various degrees of urgency in a simple world of a video game type. For the functional study of emotional architecture, Simon (1967) discussed an interruption mechanism of ongoing processes on a serially fashioned cognitive architecture. Frijda (1986) pointed out that there were a set of mechanisms ensuring personally valuable goal satisfaction. Sloman (2000) proposed their 'three layer' model and discussed the interaction of layers. This research employs a serially fashioned architecture for coping with situations in the simplified world, and intends to specify various functions for the management of multiple goals.

The agent embedded in the world is designed to have three phases in its course of problem solving. The first phase is planning to make a better plan searched as a solution path of operators in the problem space for achieving each single goal. The second phase is goal scheduling in the face of multiple goals, whose function is to schedule how to achieve the given set of goals in what order. Note that, while the target goal is being achieved, the urgency values of other goals in queue will increase due to the decrement in the available time for their achievements. The scheduling rule by a heuristics called urgency comparison is proposed. What it is aimed to do is to reduce the sum of urgency values of all the goals. The third phase, that of action mode selection, does the switching of its action mode between the execution

mode and the deliberation mode to be done in accordance with the urgency presented by the current goal. If this urgency value is very high, the agent should allocate its time for rush execution of some operators in a plan, despite of its limited plausibility. On the other hand, if the urgency is relatively low, the agent may be able to engage in a more deliberate appraisal of the global situation.

## Poster Summary

The design of our simple world will be introduced, first. A task given for the agent is to rescue as many falling objects as possible, which appear randomly in the world. Symbolic descriptions of states constituting the problem space and a plan that the agent would generate, based on the expected utility taking into account the success probability of operation, will be given (Feldman & Sproull, 1977). After the formulation of the urgency value for a goal (Toda, 1995), the two phases in the agent's problem solving, the goal scheduling and the action mode selection, will be discussed. The goal scheduling produces a quasi-optimal goal queue in a dynamic fashion in accordance with the urgency of the current goal (Minton et al. 1992; Zilberstein 1996). The action mode selection allocates limited time for actual execution and deliberate planning. High urgency value of the current goal may make the agent stay in the execution mode for a period of the available time. The final section will describe the current level of implementation and future directions of our research.

## References

- Feldman, J. A., & Sproull, R. F. (1977). Decision theory and artificial intelligence II: the hungry monkey. *cognitive science*, 1, 58-192.
- Frijda, N. H. (1986). *The emotions*. Cambridge University Press.
- Minton, S., Johnston, M. D., Philips, A. B., & Laird, P. (1992). Minimizing conflicts : a heuristic repair method for constraint satisfaction and scheduling problems, *artificial intelligence*, 58, 161-205.
- Simon, H. A. (1967). Motivational and emotional controls of cognition. *psychological review*, 74, 29-39.
- Sloman, A. (2000) Models of models of mind, *Proceedings of the AISB'00 symposium on how to design a functional mind*(pp. 1-9), University of Birmingham.
- Toda, M. (1995). A decision theoretical model of urge operations (Tech. Rep. No.95-1-01). Toyota, Japan: Chukyo University, School of Computer and Cognitive Sciences.
- Zilberstein, S. (1996). Using anytime algorithms in intelligent systems. *AI magazine*, 17, 73-83.

# Transformational Analyses of Visual Perception

Douglas Vickers (Douglas.Vickers@dsto.defence.gov.au)  
Land Operations Division, Defence Science and Technology Organisation, PO Box 1500  
Salisbury, South Australia 5108, Australia

Adrian K. Preiss (Adrian.Preiss@adelaide.edu.au)  
Department of Psychology; University of Adelaide  
Adelaide, South Australia, 5005, Australia

## Transformations and Symmetry Optimization in Visual Perception

Visual perception readily lends itself to conceptualization as an optimization process. A primary difference between perceptual theories concerns the nature of the optimized quantity. Most theories suggest that this is either economy of coding or some form of likelihood (Palmer, 1999).

A smaller number of theorists have sought to explain perception in terms of maximizing symmetry (e.g., Leyton, 1992). On our version of this view, the perceptual system subjects image elements to multiple transformations and represents structure by the parameters of those transformations that maximize correspondence with the current sensory input (Vickers, Navarro, and Lee, in press). This paper examines two applications of this approach.

## Perception of Projections of the Platonic Solids

In an early experiment, Hochberg and Brooks (1960) showed that the tendency to see outline figures as two- or three-dimensional was a function of the number of angles and line segments required to specify them in two or three dimensions. According to a transformational approach, whichever perception is associated with more symmetry-preserving transformations will occur more readily than one associated with fewer such transformations.

To test this prediction with stimuli that are representative of major classes of geometrical objects, we asked 50 observers to rate printed examples of 16 and 18 orthographic projections, respectively, of the first two of the regular polyhedra (the Platonic solids): the cube and the tetrahedron. The projections were generated in Mathematica by systematically rotating the figures around the two axes orthogonal to the line of sight

The means and standard deviations in observers' preferences for a two- or a three-dimensional interpretation covaried in a continuous manner that was (weakly) predicted by both the discontinuous differences in the symmetries of the two- and three-dimensional figures and by a count of the number of distinguishable elements. Further analyses suggested that the data may be better accounted for in terms of subjectively perceived symmetry, either as rated by observers or as estimated by the symmetry maximizing program developed by Vickers, Navarro, and Lee (in press).

## Memory and the Perception of Process History

Leyton (1992) has argued that visual perception consists of recovering the process-history undergone by an object. According to Leyton, this recovery proceeds by progressively removing asymmetries or "distinguishabilities", so as to infer an original object that is maximally symmetric. A similar evolution towards regularity is claimed for the successive reproductions of random arrays (Giraud & Pailhous, 1999). However, there has been no quantitative investigation of either of these tendencies towards symmetry.

An experiment, modeled on Bartlett's (1932) method of serial reproduction, was carried out, in which 44 observers, tested in five groups of 4 to 13, were asked to reproduce briefly presented, irregular heptagons, drawn randomly from an original pool of 168 figures. Observers were then presented with each other's (randomly allocated) reproductions and asked to reproduce them. This process was repeated until each observer had made 20 reproductions. In agreement with Leyton's hypothesis, analysis of the 57 (or more) figures that were reproduced at least 10 times showed that observers had a progressive tendency to reproduce figures with a smaller perimeter and with more nearly equidistant vertices, as measured by a reduction by a quarter and a third, respectively, in the mean and the standard deviation (normalized for perimeter size) of the lengths of the edges.

## References

- Bartlett, F.C. (1932). *Remembering*. Cambridge: Cambridge University Press.
- Giraud, M.-D., & Pailhous, J. (1999). Dynamic instability of visual images. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 1495-1516.
- Hochberg, J., & Brooks, V. (1960). The psychophysics of form: Reversible perspective drawings of spatial objects. *American Journal of Psychology*, 73, 337-354.
- Leyton, M. (1992). *Symmetry, causality, mind*. Cambridge, MA: MIT Press.
- Palmer, S. (1999). *Vision: From photons to phenomenology*. Cambridge, MA: MIT Press.
- Vickers, D., Navarro, D., & Lee, M.D. (in press). Towards a transformational approach to perceptual organization. *Proceedings of Fourth International Conference on Knowledge-Based Intelligent Engineering Systems*.

# Use of Agent and Object-Oriented Information in Language Acquisition

Laura Wagner (wagner@psych.umass.edu)

Department of Psychology, University of Massachusetts, Amherst, MA 01003 - 7710

The purpose of this study is to investigate children's bias to focus on agency information when they map meanings onto linguistic forms. Fisher et al. (1992) have shown that children have an agency bias in construing a verb's meaning so that the most agentive participant will be the verb's syntactic subject. This bias makes verbs like "give" (the giver is the most agentive actor) very easy to learn but means that learning verbs like "take" (the giver is still most agentive) require exposure to particular argument structure cues. This study asks if the agency bias extends beyond verb learning, to the grammatical encoding of events more generally.

The domain of this investigation is the English progressive construction (be + V-ing). This construction conveys information related both to the agents and to the objects of events (cf. Smith 1991). As the imperfective grammatical aspect marker, it removes completion entailments from an event, and when relevant, from the object of the event. Thus (1) indicates that (at least potentially) the event of flower-drawing is incomplete, as moreover, is the flower. In this guise, the progressive contrasts with perfective forms (2) which entail the completion of the event and when relevant, the object of the event.

- (1) The girl was drawing a flower
- (2) The girl drew a flower

In addition to this object-oriented function, the progressive also codes for the dynamicity and engagement of the agent of the event. Thus, the difference between (3) and (4) is not one of object completion (there is no object to speak of) but rather of highlighted properties of the agent.

- (3) Jenny was sitting in the chair
- (4) Jenny sat in the chair

Previous work in children's acquisition of the grammatical aspect (e.g. Weist 1991) has claimed that children understand the grammatical aspect (i.e. object-oriented) entailments of the progressive by as young as age 2;6. Children in these studies were able to consistently match a progressive sentence (1) to a picture of an incomplete event (a girl in the midst of drawing a flower) and a perfective sentence (2) to a picture of a complete event (a girl next to a completed flower). However, these studies failed to differentiate between object and agent oriented information: in all cases, the incomplete event was signaled both by the presence of an incomplete object (a half-drawn flower) and by an engaged agent (the girl working on the flower), while the complete event was signaled by both the presence of a complete object (a fully drawn flower) and an un-engaged agent (the girl proudly considering her picture). These studies cannot, therefore, tell us whether children were using object or agent oriented information (or both) to solve this task.

Additional work on grammatical aspect that limited itself to using object-oriented cues (e.g., just the half-drawn and fully drawn flowers) found very different results, including a substantial age delay in comprehension (Wagner 1998).

This result suggests that children's success in the previous tasks may depend on the accessibility of agency information.

The current experiment explicitly manipulates the information available about the agent and object, thus allowing us to see developmentally the relative importance of each source of information. It uses the same forced choice sentence-to-picture matching task used previously, in which children are presented with two depictions of the same event and asked to match these pictures to descriptions containing either the progressive or non-progressive form of a predicate. This experiment uses four kinds of picture pairs: type (1) contains only object information (parallel to the work of Wagner 1998), type (2) contains only subject information (atelic/non-completive events are used so that the status of the event's object remains constant in both depictions), type (3) contains both object and subject information consistent with each other (parallel to Weist 1991), and type (4) contains both object and subject information but at odds with each other, so that the completed object is combined with the dynamic agent and the incomplete object with the less-engaged agent.

Adults, who presumably are able to use both agent and object oriented information, should succeed with picture types (1), (2), and (3), but should provide inconsistent responses when the two types of information are at odds with each other. A child reliant on agency information, on the other hand, would succeed with types (2) and (3), fail with type (1), and behave consistently with type (4), since this child will be insensitive to the competing cues.

Preliminary results (so far, N = 20 across three groups: adults, 5-year-olds and 3-year-olds), show that all groups are able to use both object and agent-oriented information to some degree, but the 3-year-old group is much more dependent on agent-oriented information compared to the other two groups. These results therefore suggest a larger role for the agency bias in acquisition; wherever language encodes event-related information, agency information appears to play a disproportionately large role in young children's mapping process.

## Acknowledgements

This work was funded through a UMass Faculty Research Grant. My thanks go especially to Lila Gleitman and the UMass Developmental Seminar.

## References

- Fisher, C., G. Hall, S. Rakowitz and L. Gleitman (1992) When it is Better to Receive than to Give. *Lingua* 333 - 375.
- Smith, C. (1991) *The Parameter of Aspect*. Dordrecht: Kluwer.
- Wagner, L. (1998) *The Semantics and Acquisition of Time in Language*. PhD dissertation, UPenn.
- Weist, R. (1991) Spatial and Temporal Location in Child Language *First Language* 11, 253 - 267.

# Domains, Knowledge, and Constraints on Classification

**William D. Wattenmaker (wwattenmaker@njcu.edu)**

Department of Psychology; New Jersey City University  
2039 Kennedy Boulevard; Jersey City, NJ 07305

**Kathleen A. Filak (kfilak@njcu.edu)**

Department of Psychology; New Jersey City University  
2039 Kennedy Boulevard; Jersey City, NJ 07305

**Josephine A. Mendoza (jmendoz2@njcu.edu)**

Department of Psychology; New Jersey City University  
2039 Kennedy Boulevard; Jersey City, NJ 07305

This research examined the domain-generalizability of learning and processing characteristics. Wattenmaker (1995) investigated the importance of linear separability as a constraint on categorization in object and social domains. Linear separability is a principle that has been investigated as a constraint on information integration in a number of different domains including connectionist modeling and categorization (e.g., Waldmann, Holyoak, & Fratianne, 1995). In relation to categorization, linearly separable categories are categories that can be partitioned on the basis of a weighted, additive combination of component information. Several different concepts from object and social domains were used in the Wattenmaker (1995) experiments, and in all cases linearly separable structures were easier to learn when they were represented by social descriptions (e.g., trait or behavioral features) than when they were represented by object descriptions (e.g., features of animals or human artifacts).

These results were interpreted as indicating that there are fundamental differences in the structure of domains and these differences lead to different types of background knowledge being associated with different domains. If the structure of knowledge varies with domain, then it will be difficult to formulate domain general constraints in terms of abstract structural properties such as linear separability.

In the present research, a number of implications of this knowledge-based approach to domain differences were investigated in five experiments, and the results provided several important extensions of previous research. The structural property of interest in the present experiments was Family Resemblance (FR), which is a property that is closely related to linear separability and that has been investigated extensively in categorization research. Compatibility with FR principles was examined by asking participants to divide descriptions into groups (e.g., see Medin, Wattenmaker, & Hampson, 1987).

One goal of the present research was to examine domain generalizability with a broader range of domains than was used in the Wattenmaker (1995) experiments. Thus, in addition to object and social domains, we included medical categories. In two experiments that used several different medical and social categories, many more FR categories were formed in the social domain than the medical domain. This result extends prior findings to a new domain and strengthens the

conclusion that the naturalness of structural properties will vary with domain

The results also revealed clear intra-domain differences. Previous research directly contrasted concepts from one domain with concepts from another domain. However, the knowledge view predicts that if types of concepts within a domain are associated with different knowledge, then differences should be observed within that domain. Indeed, we found clear differences within the social domain as FR structures were more compatible with social trait categories than with occupation or social event categories.

The results also supported the generality of the previous findings in that in all of the experiments FR constructions occurred much more frequently in the social than the object domain. Indeed more FR constructions occurred in the social domain even when social categories were contrasted with abstract categories such as “beautiful” and “freedom.”

In summary, the results support the finding that the structure of knowledge varies with domain and these differences in knowledge will make some strategies and processes more natural in some domains than others. This will make it difficult to specify domain general constraints in terms of abstract structural properties such as linear separability or family resemblance. Future research that directly compares different domains will allow us to converge on those aspects of structure and process that are domain-general.

## References

- Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, 19, 242-279.
- Waldmann, M. R., Holyoak, K.J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General*, 124, 181-206.
- Wattenmaker, W. D. (1995). Linear separability and knowledge structures: Integrating information in object and social categorization. *Cognitive Psychology*, 28, 274-328.



# Knowledge Effects, Conceptual Structure, and Incidental Learning

**William D. Wattenmaker (wwattenmaker@njcu.edu)**

Department of Psychology; New Jersey City University  
2039 Kennedy Boulevard; Jersey City, NJ 07305

**Josephine A. Mendoza (jmendoza2@njcu.edu)**

Department of Psychology; New Jersey City University

**Vanessa K. Nieves (vnieves@njcu.edu)**

Department of Psychology; New Jersey City University

A major theme in recent research on concepts has been the influence that theories have on conceptual structure (e.g., Keil, 1989; Murphy & Medin, 1985; Wattenmaker, 1999). The terms theories and knowledge structures refer to informal theories, mental models, and general world knowledge. A second recent theme in research on concepts has been a focus on the types of concepts that are formed as a result of different encoding tasks, especially incidental and intentional tasks (e.g., Anderson & Fincham, 1996). Research on these two issues has proceeded independently, however.

Consistent with the independence of research on knowledge and research on encoding tasks, the influence of background knowledge has only been investigated with intentional tasks. Intentional tasks represent only a small subset of possible encoding tasks, however. Indeed learners in intentional tasks tend to be highly strategic problem solvers. In natural learning conditions, however, people often develop concepts when they are not in a highly analytic problem solving mode. Thus research on knowledge effects has told us very little about how prior knowledge influences concept formation in a broad range of important encoding tasks.

The present research was designed to examine knowledge effects in incidental tasks. With incidental learning, participants perform an encoding task that is unrelated to categorization. Thus although prior knowledge has been found to have a powerful influence with intentional encoding, knowledge effects might not be as great with incidental encoding. Indeed, in the process of generating hypotheses, participants in intentional conditions often actively search for relevant information. This might lead to more pronounced knowledge effects with intentional encoding.

An alternative possibility is that the activation and application of relevant knowledge will occur automatically. If this occurs, then similar types of knowledge might be applied regardless of the encoding task.

In an initial investigation of this topic, Wattenmaker (1999) examined the ability of participants to detect conceptually related feature co-occurrences. The results of these experiments revealed that background knowledge was as beneficial in incidental as intentional conditions. The present research was designed to determine if the Wattenmaker (1999) results would generalize to situations in which the application of background knowledge required more complex processes. To accomplish this, we presented participants with descriptions that could be perfectly partitioned

into two categories if an underlying theme that was consistent with prior knowledge was activated.

The results of a control condition indicated that participants rarely formed the knowledge-based categories if they had minimal exposure to the exemplars. This control condition was compared to intentional and incidental conditions. In the intentional condition, participants were told to try to discover groups that the descriptions could be divided into. In the incidental condition, the presence of groups was not even mentioned. Instead, participants were given an unrelated task. After the encoding task, participants in both conditions were given the task that was used in the control condition: all the descriptions were presented and participants were asked to divide them into two groups.

The knowledge-based categories were formed more often in the incidental than the control condition and equally often in the incidental and intentional conditions. These results occurred even though the formation of the knowledge-based categories required that prior knowledge guide the interpretation and integration of features from several dimensions.

Even though applying prior knowledge required elaborate inferential processes, background knowledge had the same degree of influence in incidental and intentional conditions. Thus these experiments provide an important extension of the finding that many types of knowledge effects will be strategy-independent (Wattenmaker, 1999). The results underscore the pervasiveness and power of the influence of background knowledge on concept formation.

## References

- Anderson, J. R. & Fincham, J. M. (1996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 259-277.
- Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.
- Murphy, G. L. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychology Review*, 92, 289-316.
- Wattenmaker, W. D. (1999). The influence of prior knowledge in intentional versus incidental concept learning. *Memory & Cognition*, 27 (4), 658-698.

# Using Cognitive Models in the Design and Evaluation of Team Structure

Monica Z. Weiland (monica\_weiland@chiinc.com)

James L. Eilbert, PhD (jim\_eilbert@chiinc.com)

CHI Systems, Inc.

716 N. Bethlehem Pike, Suite 300

Lower Gwynedd, PA 19002-2650

The Navy is moving rapidly toward deployed systems that increase use of automation while dramatically reducing manning, as in the Surface Combatant of the 21st century (SC-21) ship. With more automation and fewer people, these future systems will place more emphasis on human cognitive performance. These changes will require new organizational designs that are optimized for the cognitive role of humans and their automated counterparts. Thus, new and more general methodologies for designing organizations and optimizing the allocation of functions to individual team members must be developed.

Organizational computational models deal with the organizational structure and its effects on decision processes and information flows within the structure. They characterize organizational decisions as the aggregate of individual agents characterized by demographic and psychological parameters (Carley & Behrens, 1999). Modeling at this level, while useful for high level analysis of organizational behavior, masks the processes that occur at the individual level. To study the relation between these high level processes and individual behavior, we have been extending the COGNET theory and computational model of individual cognition (Zachary, 1992) using concepts from team training research (Smith-Jentsch, Zeisig, Acton & McPherson, 1998). This new model of organizational cognition is termed ORGNET.

To use ORGNET for designing and evaluating new designs or redesigns, we developed the Process for Redesign of Organizations (PRO) methodology and associated toolset. PRO is a flexible environment for discovering the structure of organizations/teams already in place or for building new structures. The starting point of PRO is to build an aggregate ORGNET simulation model of the overall team as a set of interacting tasks and knowledge that can perform an overall job or mission. In addition to demonstrating the basic competence of the tasks to do the team's job, the team model provides a set of measures that are used to characterize the team's basic tasks and their interactions. The measures include complexity and workload of individual tasks, as well as the information flows and workflows that emerge from model runs against representative scenarios. The user can visualize and analyze these measures to discover the basic structure and relationships of the tasks to each other.

These measures are tied to a set of design principles, such as minimizing the overall levels of communication, or

leveling workload. These design principles give the user guidance on how the measures should be used (i.e. minimized or maximized) to optimize the team structure.

The next step in PRO is to build candidate function allocations (i.e. cluster tasks into roles) based on design principles selected by the user. The user may do this manually through a graphical interface or through optimization algorithms. These algorithms find team structures that conform to the set of principles and other constraints on the organizational design defined by the user.

After an initial team structure is found, an iterative process is carried out. Tasks associated with separating the team into distinct members, i.e. communication to maintain situation awareness, task management, and backup, are added to the model and the measures are recalculated.

The user of the ORGNET software can iteratively refine the team structure either by changing the set of principles used in clustering the tasks, or by manually moving tasks between roles and seeing the results of those changes on the conformance of the design to the selected principles. We are in the process of testing various assumptions, including the efficacy of using measures from the monolithic model to predict team performance, and the use of low-fidelity aggregate models in finding an initial team structure.

## Acknowledgements

This work is being sponsored by NAVSEA PMS-500 (DD-21 Program office), with technical oversight by NAWC-TSD Code 4961.

## References

- Carley, K. M., & Behrens, D.M. (1999). Organizational and individual decision-making. In A.P. Sage & W. B. Rouse (Eds.), *Handbook of Systems Engineering and Management*. New York: Wiley-Interscience.
- Smith-Jentsch, K. A., Zeisig, R. L., Acton, B., & McPherson, J. A. (1998). Team dimensional training. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making Decisions under Stress: Implications for Individual and Team Training*. Washington, DC: APA Press.
- Zachary, W., Ryder, J., Ross, L., and Weiland, M. (1992) Intelligent Human-Computer Interaction in Real Time , Multi-tasking Process Control and Monitoring Systems. in M. Helander and M. Nagamachi (Eds.). *Human Factors in Design for Manufacturability*. New York: Taylor and Francis, pp 377-402.

# Scene Context and Change Blindness: Memory Mediates Change Detection

Carrick C. Williams (carrick@eyelab.msu.edu)

Andrew Hollingworth (andrew@eyelab.msu.edu)

John M. Henderson (john@eyelab.msu.edu)

Department of Psychology and Cognitive Science Program  
Michigan State University, East Lansing, MI 48824

Viewers often fail to detect changes to natural scenes when the change occurs during a visual disruption such as a saccadic eye movement. This *change blindness* phenomenon has led some researchers to claim that visual representation is limited to the currently attended object (e.g., Rensink, O'Regan & Clark, 1997). This *attention hypothesis* holds that once visual attention is withdrawn from an object, no visual object representation remains to support change detection. An alternative view, the *memory hypothesis*, holds that despite the change blindness phenomenon, a relatively detailed representation is retained in memory from previously attended objects (Hollingworth & Henderson, 1999).

To test these competing hypotheses, we examined participants' ability to detect changes to the visual form of a target object. Changes were made during a saccade that took the eyes away from the target object after it had been fixated the first time (Henderson & Hollingworth, 1999). Because attention precedes the eyes to the next fixation position, the target object was not within the current focus of attention when it changed. Thus, the attention hypothesis predicts that these changes should not be detected, whereas the memory hypothesis holds that visual memory can be detailed enough to support token-change detection.

In addition, we manipulated the semantic relationship between the target object and the scene in which it appeared. Research on long-term scene memory has demonstrated that semantically inconsistent (i.e., improbable) objects are retained more accurately in memory than consistent objects (Friedman, 1979). Thus, the memory hypothesis predicts not only above-floor change detection rates, but also a detection advantage for semantically inconsistent objects.

## Method

Twelve volunteers' eye movements were monitored as they viewed 24 black-on-white line drawings of realistic scenes. In each scene a semantically consistent target object (e.g., mixer in kitchen) was chosen, and targets were swapped across scenes to create stimuli for the semantically inconsistent condition (e.g., mixer in farmyard). When a change occurred, the target was replaced with a different example of that type of object (e.g., the mixer replaced by a visually different mixer). A control condition was included in which no change occurred. Participants were instructed to view each scene to prepare for a memory test and to press a button if a change occurred.

## Results

We examined the percentage of trials on which the participant detected a change in a scene. There was a

reliable difference between the consistent (18.1%) and inconsistent conditions (35.2%),  $F(1,11) = 5.28, p < .05$ . This difference was likely due, at least in part, to the fact that gaze duration prior to the change was longer for inconsistent (628 ms) versus consistent targets (489 ms),  $F(1,11) = 7.46, p < .02$ . In addition, a significant percentage of detections (41%) was delayed more than 1500 ms after the change. Of these late detections, 94% occurred upon refixation of the target. Finally, for trials on which a change was not detected, mean gaze duration when the eyes returned to the changed object (749 ms) was longer compared to the equivalent entry in the control condition (499 ms),  $F(1,11) = 6.29, p < .05$ .

These data demonstrate that participants can detect changes to the visual form of an object that is not within the current focus of attention at the time of change. Thus, these data are consistent with the memory hypothesis but not with the attention hypothesis. The modulation of detection performance by semantic consistency provides converging evidence that inconsistent objects are preferentially retained in memory. In addition, the fact that many detections were delayed more than 1500 ms and that these detections tended to occur upon refixation suggests that visual information was often retained for a relatively long period of time and consulted only when focal attention was directed back to the changed region (Henderson & Hollingworth, 1999). Finally, the large implicit effect of change on gaze duration indicates that the explicit detection measure underestimated the extent to which visual information was retained in memory.

## Acknowledgments

This research was supported by NSF grants SBR 9617274 and ECS 9873531 to John M. Henderson and by an NSF Graduate Research Fellowship to Andrew Hollingworth.

## References

- Henderson, J. M., & Hollingworth, A. (1999). The role of fixation position in detecting scene changes across saccades. *Psychological Science, 10*, 438-443.
- Hollingworth, A., & Henderson, J. M. (1999). Transsaccadic change blindness and long-term scene memory. Paper presented at Annual Workshop on Object Perception and Memory, Los Angeles.
- Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General, 108*, 316-355.
- Rensink, R.A., O'Regan, J.K., & Clark, J.J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8*, 368-373.

# Sequential Probability as a Segmentation Cue for Cantonese

Michael C.W. YIP

Department of Psychology

Chinese University of Hong Kong, Shatin, Hong Kong

[mcwvip@psy.cuhk.edu.hk](mailto:mcwvip@psy.cuhk.edu.hk)

## Introduction

Listeners appear to use sequential probability in the segmentation of Cantonese continuous speech. Because there are some sounds appear more frequently at the beginning or ending of Cantonese syllables than the others, and these kinds of probabilistic information within syllables may cue the locations of possible syllable boundaries in continuous speech signal. Three syllable-spotting experiments were conducted to examine the role of sequential probability in recognition of Cantonese syllables in the continuous speech.

## Experiment

In the syllable-spotting experiment, listeners were presented with a nonsense syllables strings [si:l~~l~~Qj4] which involved a high SP onset consonant or [si:lkwQj4] which involved a low SP consonant onset; and then listeners were instructed to spot any real Cantonese syllables [si:l], literally means lion, embedded on the basis of the acoustic alternations and the phonological information provided by the sound strings by pressing a response key and then named aloud the spotted target syllable.

## Results and Discussion

Response latencies for each target syllables shown that listeners are actually sensitive to the sequential probability on a syllable's onset during online speech segmentation. But these effect was absent on the syllable's final portion (neither the whole rime or only the final consonant). These results implied that the likelihood of a syllable's onset seems to be more important than the likelihood of a syllable's offset, that is in line with other psycholinguistics studies whose also emphasizing the importance of a syllable's onset in the fast recognition of words in continuous speech (Connine, Blasko & Titone, 1993; Grosjean, 1980; Li & Yip, 1998; Yip, in press). In addition, the absence of probabilistic effects on syllable-final may be due to the fuzzy phonotactic structure of Cantonese syllables (Yip, 2000).

Finally, together with other related research findings from other languages (Gaygen, 1999; van der Lugt, 1999), it is argued that sequential probability is an useful source of information in the segmentation of spoken language.

## References

- Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of words have a special status in auditory word recognition? *Journal of Memory and Language*, 32, 193-210.
- Gaygen, D. (1999). *Effects of phonotactic probability on the recognition of words in continuous speech*. Doctoral Dissertation, State University of New York at Buffalo, New York.
- Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics*, 28, 267-283.
- Li, P. & Yip, M. (1998). Context effects and the processing of spoken homophones. *Reading and Writing*, 10, 223-243.
- van der Lugt, A. (1999). *The use of sequential probabilities in the segmentation of speech*. Manuscript. Max-Planck-Institute for Psycholinguistics, Nijmegen.
- Yip, M. (in press). Spoken word recognition of Chinese homophones: The role of context and tone neighbors. *Psychologia*, 43.
- Yip, M. (2000). *Probabilistic Phonotactics and the Segmentation of Cantonese Continuous Speech*. Doctoral Dissertation. Chinese University of Hong Kong, Hong Kong.

# The Effect of Languages on Children’s Use of Action Information

Hanako Yoshida (hayoshid@indiana.edu)

Linda B. Smith (smith4@indiana.edu)

Department of Psychology; 1101 E. Tenth Street  
Bloomington, IN 47405-7007 USA

## Introduction

Noun meanings, and particularly the meanings of object names, do not differ much across languages. The verbs of different languages, however, differ dramatically in the meanings they lexicalize. The developmental implications of these have generated considerable controversy. In this paper, we do not speak directly to this controversy, which has been narrowly focused on the relative frequencies of noun and verb types and tokens in the input and in children’s vocabularies. Instead, we ask a subtler question: Do the marked differences among languages in verb meanings have consequences for children’s understanding of nouns? We present evidence on this issue by examining verbs of transfer verbs in English and Japanese and their effect on young children’s attention to object properties and generalization of object names.

Verbs of transfer in English and Japanese are a good starting point for answering the question of how differences in verbs may influence children’s acquisition of nouns. Specifically, the verb “put” in English is used for all sorts of transfer events -- from putting water in the tub, a cup on a table, a ring on a finger, a hat on a head, mail into a slot, and thread through a needle. In contrast, Japanese has a set of more specific transfer verbs. Thus, for water in the tub it is “haru”, for a cup on a table, it is “oku”, for a ring on to a figure it is “hameru”, for a hat on the head, it is “kaburu,” for mail into a slot, it is “sashikomu,” and thread through a needle, it is “toosu.” These verbs in contrast to the more abstract meaning of English “put” focus attention on the objects in the event and on their relation to each other. Here, then, is the question: Do these more specific verbs of transfer in Japanese modulate Japanese children’s interpretation of object names?

## Experiments

Prior to the test phase, the child was introduced to test objects with each paired containers and how the each test object fit into or through the container. As shown in Figure 1, one test pair, the same-Fit choice, contained an object and container that matched the exemplar in the fit as described by the Japanese verb. One test pair, the same-Shape choice, contained an object the same shape as the exemplar. However, the fit of this object into the provided container would

not be instance of the Japanese verb “hameru.” The third test pair, the distracter, presented an object and a fit into the provided container that was unlike the exemplar. During the test phase, the child was presented with an exemplar and their respective containers with its appropriate action. For example, the exemplar for a “Hameru” set was demonstrated by a pushing motion how it snapped into the same shaped container, a manner outcome that would be referred to by the Japanese verb “hameru.” Then, children were asked to select one from these three choices. The specific conditions across the three experiments 1 through 3 differed in the verbal descriptions of the events, the verbal requests to make a choice and whether the exemplar had a name or not.

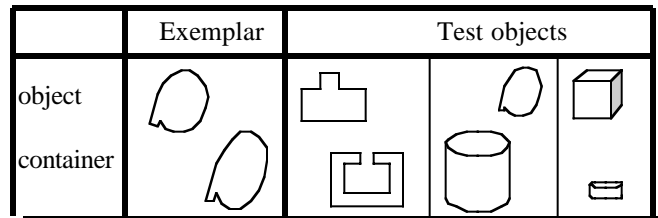


Figure 1. Actual stimuli

## Conclusion

This series of experiments suggest that language a child learns emphasize or de-emphasize action information.

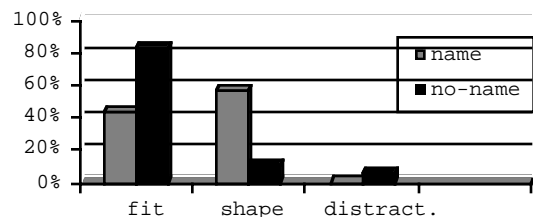


Figure 2. Japanese-speaking children’s performance

As you can see in Figure 2., Japanese-speaking children’s attention to shape of objects increased when they generalize name of the object, and decreased when they generalize object without name. Explicit action information comes to guide the naming of novel objects. However, English speaking-children did not affected by the verbal cues as much as Japanese-speaking children did.

## Author Index

- Adolphs, Ralph, 101  
Ahn, Woo-kyoung, 19  
Akamatsu, Shigeru, 322  
Albacete, Patricia L., 25  
Allen, James F., 1  
Almor, Amit, 310  
Alsop, David, 1022  
Altmann, Erik M., 31, 847  
Anderson, John R., 369, 1042, 1050  
Andonova, Elena, 693  
Andrews, Janet K., 1015  
Archambault, Annie, 585  
Aster, Denise, 1019  
Atkinson, Robert K., 393, 591  
Badecker, William, 523  
Baldwin, Timothy, 597  
Bale, Alan C., 459  
Bates, Elizabeth A., 1033  
Bay, Stephen D., 603  
Beaman, C. Philip, 37  
Becker, Suzanna, 1031, 1041  
Beller, Sieghard, 853  
Bensusan, Hilan, 609  
Bertenthal, Bennett I., 794  
Billman, Dorrit, 615  
Blair, Nathaniel J., 1039  
Blessing, Stephen, 1029  
Blok, Sergey Victor, 621  
Bloom, Jonathan E., 447  
Bongers, Raoul M., 535  
Boroditsky, Lera, 42  
Boudelaa, Sami, 48  
Bowles, Anita R., 1016  
Branstrom, Robert B., 54  
Brem, Sarah, 741  
Bridgeman, Bruce, 60  
Brook, Andrew, 65  
Brédart, Serge, 148  
Bundy, Alan, 226  
Burns, Bruce D., 541, 559, 627  
Byrne, Michael D., 71  
Campbell, Ruth, 322  
Carlson, Richard A., 1017  
Carpenter, Patricia A., 1060  
Casasanto, Daniel J., 77, 1018  
Catrambone, Richard, 591  
Cheng, Peter C-H., 776  
Chi, Michelene T. H., 699, 705  
Chiarello, Christine, 1035  
Choplin, Jesse M., 232  
Chown, Eric, 142  
Christiansen, Morten H., 83, 645  
Chuah, Johnny, 633  
Chung, He Len, 794  
Clearfield, Melissa, 639  
Cleeremans, Axel, 947  
Cohen, Leslie, 1019  
Coleman, Mike, 322  
Colgan, Sheila, 328  
Collister, David, 1020  
Colunga, Eliana, 89, 160, 1021  
Conrad, Frederick G., 447  
Constantinou, Constantinos P., 859  
Conway, Christopher M., 83  
Cooke, Ayanna, 1022  
Corley, Martin, 435  
Costello, Fintan, 95  
Cottrell, Garrison W., 101, 352, 358  
Curtin, Suzanne, 83  
Curtiss, Susan, 1024  
Cutler, Anne, 1034  
Cutler, D. L., 782  
Dailey, Matthew N., 101  
Davidson, Lisa, 1023  
de Bode, Stella, 1024  
Dennis, Martin J., 19  
Detre, John A., 77, 1018, 1022  
DeVita, Christian, 1022  
Diab, Mona, 399  
Dowman, Mike, 107  
Drake, Peter, 639  
Durbin, Michael A., 113  
Dyer, Fred C., 1026  
Earwood, Jason, 113  
Eilbert, James L., 1067  
Ellefson, Michelle R., 645  
Estes, Zachary, 1025  
Evans, Jonathan St. B. T., 119  
Evens, Martha W., 262  
Falk, Richard J., 1026  
Feeney, Aidan, 119  
Felberbaum, Michael, 481  
Ferguson, Ronald W., 125  
Ferreira, Victor, 352  
Filak, Kathleen A., 1065  
Fine, I., 131

Fodor, Janet Dean, 136  
Forbell, Eric, 142  
Forbus, Kenneth D., 286, 770  
Ford, Martin R., 901  
Fox, Susan Eileen, 651  
Frank, Robert, 523  
Freedman, Reva, 262  
French, Robert M., 148, 657  
Fu, Wai-Tat, 154, 663, 959  
Gabriel, Rosângela, 1027  
Galantucci, Bruno, 505  
Gallistel, Randy, 1  
Gaskell, M. Gareth, 48, 405  
Gasser, Michael, 160  
Gee, James, 1022  
Gentner, Dedre, 286, 621, 770  
Gerjets, Peter, 166, 441  
Gillis, Steven, 1058  
Giraud-Carrier, Christophe, 609  
Glass, Michael, 262  
Gleitman, Lila, 481  
Glosser, Guila, 77  
Gobet, Fernand, 723, 776  
Goel, Ashok, 196  
Golden, Richard M., 113  
Goldrick, Matthew, 1028  
Goldstone, Robert L., 172  
Golinkoff, Roberta M., 794  
Gosselin, Frédéric, 178, 585, 930  
Graesser, Arthur, 184, 983  
Gray, Wayne D., 154, 663, 753, 1052  
Gredebäck, Gustaf, 190  
Greeno, James G., 669  
Griffith, Todd W., 196  
Griffiths, Thomas L., 202  
Gross, Steven, 208  
Grossman, Murray, 1022  
Gunzelmann, Glenn, 1029  
Gupta, Prahlad, 812  
Hagen, Cornelius, 871  
Hagmayer, York, 214  
Hagstrom, Paul, 292  
Hahn, Udo, 429  
Hall, Rogers, 675  
Harley, Trevor A., 328  
Haselager, W. F. G. (Pim), 535  
Hayes, Matt, 936  
Healy, Alice F., 1016  
Heise, Elke, 441  
Henderson, John M., 1026, 1068  
Hendriks, Petra, 220  
Heneveld, Alex, 226  
Hicks, John, 681  
Hiraki, Kazuo, 1051  
Hirsh-Pasek, Kathy, 794  
Hogeweg, Paulien, 577  
Hollich, George J., 1030  
Hollingworth, Andrew, 1026, 1068  
Holst, Anders, 1036  
Holyoak, Keith J., 1045  
Howell, Steve R., 1031  
Huart, Johanne, 148  
Hummel, John E., 232  
Hund, Alycia M., 1032  
Hutchins, Sean, 511  
Iyer, Gowri K., 1033  
Jacobs, Robert A., 131  
Janetzko, Dietmar, 687  
Jansen, Anthony R., 238  
Janyan, Armina, 693  
Jeong, Heisawn, 699, 705  
Jimura, Koji, 711  
Johansen, Mark K., 1039  
Johnson, Andrew, 364  
Johnson, Elizabeth K., 1034  
Johnson, Todd R., 547, 633, 717  
Johnson-Laird, P. N., 759, 1000  
Jones, Gary, 723  
Jones, Gregory V., 729  
Jones, Sari, 244  
Jordan, J. Scott, 764  
Jordan, Pamela W., 250  
Jusczyk, Peter, 1030, 1034, 1056  
Juslin, Peter, 190, 244, 841  
Just, Marcel Adam, 1060  
Kacinik, Natalie, 1035  
Kako, Edward, 256  
Kalmanson, Julia, 735  
Kamachi, Miyuki, 322  
Kanerva, Pentti, 1036  
Karnavat, Ashish, 184  
Katagiri, Yasuhiro, 529  
Kaufman, David R., 5, 741  
Kayser, Patrick A., 1040  
Ke, Lan, 901  
Keller, Frank, 747  
Killgore, William D. S., 77  
Kim, Jung Hee, 262  
Kinyon, Alexandra, 268  
Kirschenbaum, Susan S., 753  
Klahr, David, 1047  
Knauff, Markus, 759, 871

Knightly, Leah M., 1037  
Knoblich, Günther, 764  
Kobayashi, Yuki, 1038  
Kokinov, Boicho, 274  
Komazaki, Hisaaki, 711  
Kovordányi, Rita, 280  
Kristoferson, Jan, 1036  
Kruschke, John K., 1039  
Krych, Meredyth, 615  
Kuehne, Sven E., 286, 770  
Kusumi, Takashi, 711  
Labieuse, Christophe, 148  
Lane, Peter C. R., 776  
Lea, R. Brooke, 1040  
Legendre, Géraldine, 292  
Le Pelley, M. E., 782  
Lepper, Mark R., 464  
Levy, Simon, 298  
Lewis, Eric, 741  
Lewis, Joseph A., 788  
Li, Ping, 304  
Lim, Jean C., 1041  
Liu, Jing, 794  
Livingston, Kenneth R., 1015  
Lo, Ya-Fen, 469, 912  
Long, Christopher J., 310  
Love, Bradley C., 316, 800  
Lowe, Will, 806  
Luce, Paul A., 1030  
Ludden, David, 812  
Luger, George F., 788  
Lyons, Michael J., 322  
MacAndrew, Siobhan B. G., 328  
Maglio, Paul P., 818  
Mahadevan, Sridhar, 1026  
Maier, Uwe H., 393  
Maldjian, Joseph A., 77  
Marcovitch, Stuart, 334  
Markman, Arthur B., 565, 735  
Marriott, Kim, 238  
Marshark, Eve, 1019  
Marslen-Wilson, William, 346, 387, 405  
Martin, Maryanne, 729  
Matessa, Michael, 1042  
Mathern, Gary W., 1024  
Matsuoka, Takashi, 711  
Mauth, Kerstin, 1043  
McDonald, Scott, 806  
McLaren, I. P. L., 340, 782, 994  
Melnik, Ofer, 298  
Mendoza, Josephine A., 1065, 1066  
Meunier, Fanny, 346  
Milostan, Jeanne C., 352  
Mintz, Toben H., 1044  
Mix, Kelly, 639  
Moher, Thomas G., 364  
Molenaar, Peter C. M., 971  
Morris, Bradley J., 823  
Morrison, Robert G., 1045  
Mozer, Michael C., 16  
Mulligan, Elizabeth J., 1040  
Myers, Jerome L., 1040  
Nadig, Aparna, 1046  
Nagy, Gabriella, 553  
Nakagawa, Masanori, 711  
Nersessian, Nancy J., 196  
Newton, Natika, 10  
Ngo, Thuy L., 883  
Nieves, Vanessa K., 1066  
Nigam, Milena K., 1047  
Noelle, David C., 358  
Nokes, Timothy J., 829  
Norris, Dennis, 1034  
Oberlander, Jon, 681  
Ohlsson, Stellan, 364, 829  
Okada, Takeshi, 835  
Okubo, Matia, 1048  
Olsson, Henrik, 244  
O'Malley, Christine J., 889  
Paxman, David, 1049  
Pazzani, Michael J., 603  
Persson, Magnus, 841  
Peterson, Margaret J., 847  
Petrov, Alexander A., 274, 369  
Pine, Julian M., 723  
Piper, Kim, 794  
Placa, Nicora, 481  
Plante, Andre, 322  
Ploetzner, Rolf, 853  
Plunklett, Kim, 1027  
Pollack, Jordan, 298  
Pomeroy, Victoria, 184  
Pomplun, Marc, 375  
Preiss, Adrian K., 1063  
Quinn, Bryan, 770  
Raftopoulos, Athanassios, 859  
Rajmakers, Maartje E. J., 971  
Ramey, Christopher H., 794  
Ramscar, Michael, 226, 381, 571, 865  
Ranney, Michael, 741  
Rapp, Brenda, 1028  
Rauh, Reinhold, 871



Reid, Agnieszka, 387  
Reingold, Eyal M., 375  
Renkl, Alexander, 393  
Resnik, Philip, 399  
Rheinberg, Falko, 541  
Rhine, Ramon, 1053  
Richardson, Daniel C., 487  
Richardson, Julian, 226  
Rodd, Jennifer, 405  
Roelofs, Ardi, 411  
Saccuman, Cristina M., 1033  
Saffran, Jenny R., 417  
Salvucci, Dario D., 1050  
Samuelson, Larissa K., 423  
Sashima, Akio, 1051  
Schauer, Holger, 429  
Scheepers, Christoph, 435  
Scheiter, Katharina, 166, 441  
Schlieder, Christoph, 871  
Schmidt, Lauren A., 42  
Schober, Michael F., 447  
Schoelles, Michael J., 1052  
Schöner, Gregor, 924  
Schooler, Lael, 1053  
Schoppek, Wolfgang, 877  
Schuh, Kathy L., 1054  
Schunn, Christian D., 883, 889, 959, 965  
Schutte, Anne R., 1055  
Schyns, Philippe G., 178, 585  
Scott, Sam, 895  
Sedivy, Julie, 1046  
Shastri, Lokendra, 453  
Shears, Connie, 1035  
Shen, Jiye, 375  
Shillcock, Richard, 681  
Shimajima, Atsushi, 529  
Shimokido, Takashi, 835  
Shipley, Thomas F., 1019  
Shultz, Thomas R., 459, 464  
Sidman, Robert D., 901  
Silva, Juan Carlos Serio, 1053  
Sloutsky, Vladimir M., 469, 475, 907, 912, 1006  
Smith, Linda B., 89, 1021, 1070  
Smolensky, Paul, 1028  
Snedeker, Jesse, 481  
Soderstrom, Melanie, 1056  
Sommerfeld, Melissa C., 493, 669  
Sougné, Jacques, 918  
Spellman, Barbara A., 1045  
Spencer, John P., 924, 1032, 1055  
Spencer-Smith, Jesse, 1057  
Spivey, Michael J., 487  
Stenning, Keith, 493  
Stevenson, Lisa M., 1017  
Stewart, Andrew J., 930  
Straub, Kathy, 523  
Strube, Gerhard, 871  
Subramanian, Devika, 499  
Suret, Mark, 340, 994  
Swilley, Angela, 615  
Tabor, Whitney, 505, 511  
Tack, Werner H., 166  
Taelman, Helena, 1058  
Taht, Kathy, 1019  
Takano, Yohtaro, 1059  
Tanaka, Akihiro, 1059  
Tanaka, Hozumi, 597  
Taraban, Roman, 936  
Taylor, Roger, 705  
Teller, Virginia, 136  
Tenenbaum, Joshua B., 16, 202, 517  
Thanukos, Anna, 741  
Thibaut, Jean-Pierre, 942  
Timmermans, Bert, 947  
Todorova, Marina, 292, 523  
Tomitch, Lêda Maria Braga, 1060  
Toth, Eva Erdosne, 953  
Trafton, J. Gregory, 959, 965  
Trickett, Susan B., 959, 965  
Trueswell, John C., 256, 481  
Trumpower, David, 1061  
Tsuchiya, Takafumi, 1062  
Tversky, Barbara, 1020  
Tyler, Melinda J., 487  
Umata, Ichiro, 529  
Vainikka, Anne, 292  
van der Henst, Jean-Baptiste, 1000  
VanLehn, Kurt A., 25  
van Rooij, Iris, 535  
Venn, Simon, 119  
Vickers, Douglas, 1063  
Visser, Ingmar, 971  
Vollmeyer, Regina, 541, 627  
Wagner, Laura, 1064  
Waldmann, Michael R., 214  
Wang, Hongbin, 547, 717  
Wattenmaker, William D., 1065, 1066  
Weiland, Monica Z., 1067  
Wendelken, Carter, 453  
Wenger, Michael J., 818  
West, Robert L., 553

Westermann, Gert, 977  
Wexler, Kenneth, 1056  
Wiebe, Muffie, 669  
Wiemer-Hastings, Katja, 184, 983  
Wiemer-Hastings, Peter, 989  
Wieth, Mareike, 559  
Williams, Carrick C., 1068  
Williams, Diane E., 375  
Wills, A. J., 994  
Winman, Anders, 190, 244  
Wulfeck, Beverly B., 1033  
Xu, Fei, 517

Yamauchi, Takashi, 565  
Yang, Yingrui, 1000  
Yarlas, Aaron S., 475, 1006  
Yarlett, Daniel, 381, 571  
Yelland, Greg W., 238  
Yip, Michael C. W., 1069  
Yoshida, Hanako, 1070  
Young, Ezekiel E., 487  
Zelazo, Philip David, 334  
Zhang, Jiajie, 547, 633, 717  
Zuidema, Willem H., 577