UNIVERSITY OF CALIFORNIA
SANTA CRUZ

THE FEATURES WE USE AND THE FEATURES WE LOSE: ENCODING, MAINTENANCE, AND RETRIEVAL

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

LINGUISTICS

by

Stephanie K. Rich

June 2024

The Dissertation of Stephanie K. Rich
is approved:

Professor Matthew W. Wagers, Chair

Associate Professor Amanda Rysling

Assistant Professor Jess H.K. Law

Associate Professor Brian Dillon

Peter F. Biehl
Vice Provost and Dean of Graduate Studies

Contents

List of Figures **vi**

List of Tables **viii**

Abstract **xi**

Acknowledgments **xiv**

1 Introduction **1**
- 1.1 Central questions 2
- 1.2 Roadmap 5

2 The Theoretical Landscape **7**
- 2.1 Dependency resolution 7
  - 2.1.1 Subject-verb dependencies 7
  - 2.1.2 Processing gaps 9
  - 2.1.3 Avenues for interference 15
- 2.2 Similarity-based interference 16
- 2.3 Encoding and retrieval 19
- 2.4 Explorations in this dissertation 24

3 Encoding interference **25**
- 3.1 Isolating encoding interference 25
- 3.2 Design 28
- 3.3 Predictions 30
- 3.4 Norming 31
  - 3.4.1 Norming: Similarity 31
  - 3.4.2 Norming: Plausibility 33
  - 3.4.3 Final itemset 35
- 3.5 Experiment 1: Encoding interference in self-paced reading 35
  - 3.5.1 Procedure 36
  - 3.5.2 Data processing and analysis 37

3.5.3 Experiment 1a 38
3.5.4 Experiment 1b 41
3.5.5 Experiment 1 discussion 44
3.6 Experiment 2: Encoding interference in eye-tracking while reading 45
3.6.1 Methods 46
3.6.2 Initial reading time results 50
3.6.3 Rereading results 52
3.6.4 Discussion 57
3.7 Discussion 59
3.7.1 Similarity 60
3.7.2 Proximity: Open questions 63

4 Encoding contexts **66**
4.1 Experiment 3: Manipulating contiguity and not structural similarity 67
4.1.1 Experiment 3a 68
4.1.2 Experiment 3b 72
4.1.3 Discussion 77
4.2 A closer look at the Temporal Context Model 79
4.2.1 Encoding contexts in list-memory research 80
4.2.2 Connections to language 83
4.3 Motivating Experiments 4 and 5 83
4.4 Experiment 4: Sensitivity to linguistic boundaries during recall . . 85
4.4.1 Methods 87
4.4.2 Predictions 91
4.4.3 Results 92
4.4.4 Discussion 96
4.5 Experiment 5: Sensitivity to linguistic boundaries during recognition 99
4.5.1 Methods 101
4.5.2 Predictions 104
4.5.3 Results 105
4.5.4 Discussion 105
4.6 Discussion 108

5 Availability of features **111**
5.1 Experiment 6: Filler gap encoding 112
5.1.1 Methods 116
5.1.2 Data processing 119
5.1.3 Predictions 121
5.1.4 Question accuracy 122
5.1.5 Question response time 123
5.1.6 Reading time results: Correct trials only 125
5.1.7 Analysis with incorrect trials 127

5.1.8 Discussion 129
5.2 Parallelism 136
5.3 Experiment 7: Parallelism in syntactic structure 140
5.3.1 Methods 141
5.3.2 Data processing and analysis 143
5.3.3 Predictions 143
5.3.4 Results 144
5.3.5 Discussion 145
5.4 Experiment 8: Parallelism in thematic roles 146
5.4.1 Methods 146
5.4.2 Data processing 148
5.4.3 Predictions 148
5.4.4 Results 149
5.4.5 Discussion 150
5.5 Experiment 9: Parallelism in information structure 150
5.5.1 Methods 152
5.5.2 Results: All items 154
5.5.3 Results: Probable repeated events 156
5.5.4 Discussion 157
5.6 Discussion 159
5.6.1 Experiment 6 and overlapping activation 159
5.6.2 Activated features in the reanalysis of gap positions 159

6 Conclusion and future work **162**
6.1 Summary of findings 162
6.1.1 Similarity 162
6.1.2 Closeness 163
6.1.3 Closeness as similarity 164
6.2 Implications and future work 164
6.2.1 Broader discourse influences on encoded features 165
6.2.2 Predictive consequences of encoding interference 168
6.2.3 (Potentially) unexpected sources of encoding interference . 171
6.2.4 Applications beyond neurotypical populations 173
6.3 Final conclusion 173

References **175**

List of Figures

3.1 Mean reading times across Experiments 1a and 1b, starting from the pre-critical region (the last region within the RC), up until the spillover region (the first region following the main verb). Near conditions are represented on the left panel, Far conditions on the right. Experiment 1a mean reading times are displayed in the top panels, Experiment 1b on the bottom. 43

3.2 First fixation durations on the auxiliary had been region, demonstrating a marginal interaction between Similarity and Proximity. 54

3.3 Second pass reading times on the auxiliary had been region, demonstrating an interaction between Similarity and Proximity. 56

3.4 Second pass reading times on the main verb region, demonstrating a penalty for Medium similarity over High similarity. 57

4.1 Mean reading times in the Early conditions in Experiment 3a in the extended retrieval site regions and spillover regions. 71

4.2 Mean reading times in the Late conditions in Experiment 2a in the extended retrieval site regions and spillover regions. 71

4.3 Mean reading times in the Early conditions in Experiment 3b in the extended retrieval site regions and spillover regions. 75

4.4 Mean reading times in the Late conditions in Experiment 3b in the extended retrieval site regions and spillover regions. 76

4.5 Example (63) from Potts (2002) demonstrating the attachment of an *as*-parenthetical. 79

4.6 Participant view of the recall phase. 90
4.7 Recall accuracy following a correct response for the reactivation question. 93
4.8 Recall accuracy following an incorrect response for the reactivation question. 94
4.9 Recognition decision judgment scale. 103
4.10 Receiver operating characteristic curves for Within and Across conditions, scaled against Mismatch conditions. 106

5.1 Response times to comprehension questions targeting the retrieved subject of the continuation (_ *was too old*). 125
5.2 Summed reading times across the filled gap region *the kayaks/carts next to*, comparing the Post-intervener conditions to the No gap conditions. 126
5.3 Reading times (ms) for the critical region following later reactivation, *was too old* and spillover *for the camp*, correct trials only. . . 127
5.4 Summed reading times across the filled gap region *the kayaks/carts next to*, comparing the Post-intervener conditions to the No gap conditions, faceted by incorrect (0) vs correct (1) responses. . . . 128
5.5 Reading times (ms) for the critical region following later reactivation, collapsing across gap position, faceted by incorrect (0) vs correct (1) responses. 130
5.6 Reading times (ms) for the critical region following later reactivation, *was too old* and spillover *for the camp*, faceted by incorrect (0) vs correct (1) responses. 131
5.7 Reading times for Experiment 7. 145
5.8 Reading times for Experiment 8. 149
5.9 Reading times for all items in Experiment 9. 156
5.10 Reading times for probable repeated events in Experiment 9. . . . 157

List of Tables

3.1 Sample item set illustrating the Similarity x Proximity manipulation in the relative clause region and the extended retrieval site containing two auxiliaries (*had been*) and an adverb (*recently*) preceding the main verb (*sharpened*). 30
3.2 Norming results for the final set of Experiment 1 items, M (SE). . 35
3.3 Question accuracy by condition for Experiment 1a, M (SE). . . . 39
3.4 Question accuracy by condition for Experiment 1b, M (SE). . . . 42
3.5 Means and standard errors (M (SE)) in milliseconds of analyzed eye-tracking regions and measures that are indicative of initial reading. 51
3.6 Means and standard errors (M(SE)) of analyzed eye-tracking regions and measures that reflect rereading, represented in milliseconds or probability of regression. 53

4.1 Accuracy data per condition for Experiment 3a, M (SE). 69
4.2 Accuracy data per condition for Experiment 3b, M (SE). 73
4.3 Output for brms mixed effect model for the correct trial analysis. 92
4.4 Recall accuracy following an incorrect response for the reactivation question. 94
4.5 Intrusion data: Percent of errors that were intrusions, by position. In the following two columns, of the intrusions, the breakdown of extra- and intra-item intrusions. 95

4.6 Percent of intrusions (out of all intrusions in each position) that displayed the following characteristics: CC, an intrusion from a contiguous chunk; CC_{for}, CC intrusions from the following (forward) chunk; CC_{back}, CC intrusions from the preceding (backward) chunk; CP_{AC}, intrusions from a contiguous position across a chunk boundary; CP_{WC}, intrusions from a contiguous position within the same chunk. 96
4.7 Results of pROC Signal Detection Theory model. 105

5.1 Characteristics of similarity and length of the interveners, in comparison to the filler, Mean (SE). 116
5.2 Sample item set illustrating the Gap x Similarity manipulation. The Similarity manipulation is demonstrated in-line in brackets, with the High similarity noun first followed by the Low similarity noun. 118
5.3 Question accuracy for Experiment 6, means (M) and standard error (SE). 123
5.4 Sample item set illustrating the Instrument x Gap manipulation. The Instrument manipulation here refers to whether instrument PPs are found in both clauses. The critical region (*the dessert*) in the second clause is bolded. 142
5.5 Sample item set illustrating the Instrument x Gap manipulation. The Instrument manipulation here refers to whether an instrument is mentioned in the first clause. The critical region (*the dessert*) in the second clause is bolded. 147
5.6 Sample item set illustrating the Instrument x Gap manipulation. The Instrument manipulation here refers to whether the instrument has a prominent role in terms of information structure (manipulated via definiteness). The critical region (*the dessert*) in the second clause is bolded. 153

5.7 A set of examples of items for Experiment 9 that conveyed possible and improbable repeated events. 154

Abstract

The features we use and the features we lose: Encoding, maintenance, and retrieval

by

Stephanie K. Rich

This dissertation explores the role of memory in language processing, and specifically how interference during lexical encoding can result in downstream interference during retrieval. The dissertation merges insights from both the sentence processing literature as well as the study of memory in non-sentential contexts and focuses on two factors which have been shown to contribute to interference: semantic similarity and proximity during encoding.

Experiments 1-2 in Chapter 3 examine interference on the basis of semantic similarity between a subject and intervener at a later retrieval site that initially does not provide any semantically selective information. Differential reading times prior to semantically selective information provide evidence for interference not just during retrieval, but during encoding. Experiments 3-5 in Chapter 4 examine proximity during encoding, and demonstrate in both recall and recognition that linguistic boundaries (e.g. clause boundaries) can serve to delineate encoding contexts in memory: reactivation of one word prior to recall boosts recall of surrounding words only within a clause, and sensitivity to changes in noun order is greater if a change occurs across a clause boundary. Experiments 6-9 in Chapter 5 explore lexical encoding in filler-gap dependency processing. Experiment 6 investigates the filled gap effect as a moment of overlapping activation between the filler and the encountered word, which may increase the chances of encoding interference. Experiments 7-9 examine whether a highlighted thematic role in the

first clause of a biclausal sentence may influence the position of predictively postulated gaps. This is discussed in terms of the initial featural representation of a wh-filler word.

The dissertation concludes by connecting the themes discussed throughout the work to other areas of interest, such as predictive processing more generally.

для бабули

Acknowledgments

One of the greatest joys of finishing graduate school is the opportunity to publicly thank the people who shared this journey with me. I have been very fortunate to work with so many people who have inspired me, mentored me, supported me, and provided dear companionship along the way. One of the horrors is knowing I won't be able to capture the full extent of my gratitude here.

I'd like to begin by acknowledging my committee members. I feel very fortunate to have been advised by Matt Wagers throughout my time at UC Santa Cruz. He has supported me through various interests and ideas and, more importantly, challenged me to connect them all. In so doing, he has helped me gain a better understanding not only of the field, but of the work that I had done. Thank you, Matt, for guiding me through all of the important things: experimental design, theories of memory, spreadsheet organization, how to generate novel slack emoji, the list goes on. Thank you, too, for creating such a joyful place to do psycholinguistics, a place where I could be my authentic person, even if who that is is someone bad at taking care of plants. For much of graduate school, I felt that the dissertation was an end point that I was working towards. As I sit down to finish it, I find myself hoping that it is only the beginning.

I'd like to thank the rest of the committee for many thoughtful questions, and general encouragement, along the way. I am grateful that Jess Law joined the department in the time that I've been here, and that I've had a chance to work with her in a number of different ways. I have many fond memories from mapping out distributivity in the little music shed to creating possible syllabi for classes that may or may not happen, and I have always benefited from her feedback. I first came to know Brian Dillon at UMass Amherst, and while he was not ever officially my mentor or teacher there, he welcomed me into lab meetings and was always happy to answer questions. I am so grateful that brief instances of hallway mentorship have turned into broader theoretical discussions as I developed the work in this dissertation. It's hard to believe Amanda Rysling was one of my first mentors in linguistics, over ten years ago. She shaped my understanding of linguistics and academia at a time I was just coming into the field, encouraged me to pursue this path, and gave me the tools to do it. I don't think I would have pursued graduate school without her mentorship early on - thank you.

I'd like to thank the Santa Cruz faculty generally for building such a vibrant research community that values collaboration and interdisciplinary research. I'd like to thank Maziar Toosarvandani for taking in a psycholinguist for a non-experimental QP, for mentorship around writing, for rewarding meetings even when I was struggling, and for now sharing an office wall. I've been lucky to interact with a number of faculty members on committees prior to the dissertation committee: thank you to Adrian Brasoveanu for excited encouragement and exciting reading groups; Ivy Sichel for discussions about features, a syntax project that I wish I hadn't abandoned in the throws of a pandemic, and advice that came exactly when I needed it; Pranav Anand for helping create a bridge over which I could explore NLP (and then retreat back over); and Grant McGuire for asking a question in my QE defense that I should have known the answer to, and so haven't forgotten the answer since. I'd like to thank Junko Ito, Armin Mester, and Ryan Bennett, who welcomed me and my interest in prosody into SPOT meetings and AMP planning meetings alike. I'd also like to thank Roumi Pancheva for several important conversations, ranging from academia to not. I feel grateful for the teaching-related support I've received from the faculty as well. Thank you Jorge Hankamer, Jim McCloskey, and Sandy Chung, for showing how to make the best of a stressful quarter. I'd like to thank Jaye Padgett for sharing his pedagogical goals and ideas with me, and supporting mine as well, and Rachel Walker for the joyful coincidence of teaching at the same time in neighboring buildings. I'd also like to recognize the departmental staff over the time I've been here, who contribute so much to our community. Thank you to Maria Zimmer, James Funk, Ashley Hardisty, and Logan Roberts for so much support over the years. Thank you, Sarah Arantza Amador, for supporting my writing in all forms, for making the halls of Stevenson such a jovial place to be, and putting on your mean manager face when it was too jovial (sorry!). Thank you, Gwyn Vandevere, for your careful and passionate support of our undergraduate students, for your insight into both teaching and teaching bureaucracy, and for giving me a fellow dissertator to remotely cheer for. I would also like to acknowledge the staff of the Teaching and Learning Center, especially Megan McNamara for support in developing EquiL, and Kendra Dority for being so actively involved in my development as a teacher, mentor, and person who applies to jobs.

I am so grateful to the network of mentorship and support I received prior to coming to Santa Cruz. In many ways, I feel I grew up as a psycholinguist in the UCLA Language Processing Lab, and owe to the careful mentorship of Jesse Harris too many things to count, from an understanding the inner workings of a lab, to psycholinguistic theory, to how to navigate a spreadsheet quickly, and, importantly, how to properly appreciate googly eyes. Thank you, Jesse, for long discussions of life and psycholinguistics, and supporting me through both. Thank you, too, to the graduate students and undergraduate RAs in the lab, who joined me on so many occasions for working or not working. And I will always be grateful to the UMass Amherst linguistics department (the faculty who taught me, the graduate students who welcomed me, and my fellow undergrads). UMass was where I learned what linguistics was, and that it was a thing I could do (thank you, Tom Maxfield, for telling me I was the right kind of weird to be a linguist, and that that should make me very afraid). Thank you to John Kingston, for being the first to welcome me into a lab. When I left, you gave me the advice that in order to do well in grad school, I needed to do two things. I don't remember what they were, but I hope I have done them. Thank you to Chuck Clifton for fun meetings in the windowless rooms of Tobin, and to Kyle Johnson for failing to persuade me to stay an English major. And to Lyn Frazier, thank you for sharing your infectious excitement for language processing, for long meetings about theory and pressing me to form a clear hypothesis. Thank you for holding space for my ideas when I was taking my first steps into this field, it meant the world to me.

In this moment of looking back at where my path has wound through, I am also grateful to see where it will extend in the future. Thank you Roberto de Almeida and Gary Libben for giving some direction to my future path, and for giving me such a good reason to wrap up this document.

I once held the belief that mentorship was unidirectional, received from those further along. I have been grateful to unlearn this idea, and find myself exchanging mentorship with my graduate school peers. To this end, I could not be more thankful for my first year cohort SHAJaMM: myself, **H**yein Jeong, **A**nelia Kudin, **Ja**ck Duff, **M**orwenna Hoeks, **M**ax Kaplan. Thank you for hours of collective problem solving, joint overleaf accounts, and pushing and inspiring me to always do better. To the PhD continue-ers, the men to my never ending ocean, thank you for teaching me more than you each individually know. Morwenna, thank you for overleaf templates, showing me math on a brewery napkin, and being an all around great fellow surfer office girl. Jack, the peaceful harp to my melodic techno, thank you for being on the journey alongside me from the UMass phonetics lab, to conference travels, to the great outer wilds of space, and to the desk right

beside me in Stevenson 232. Max, thank you for sharing your joys with me, for advising meetings of all kinds, for teaching me (perhaps without realizing) how to prioritize physical health, and for offering some peace of mind on a lonely avenue. It truly has been an honor to travel with you three through grad school.

I'd like to non-chronologically front an acknowledgment to a different kind of cohort: those who kept me sane, fed, and on task in the last few weeks of finishing this. Thank you for writing appointments, walks to the Slug Shop, theoretical discussions and sanity checks, a well-timed off-hand comment, jointly kept to-do lists, writing sprints where I won by only one word, late nights in the lab, lots of tea and juice, and being famous musicians who posted hour-long sets on YouTube - you all know your pairwise connections, in no particular order: Lalitha, Richard, Yağmur, Duygu, Niko, Dori, Above & Beyond, and Giolì & Assia.

When I joined the department, I felt so lucky to be welcomed by the friendship of more advanced graduate students who, whether they knew this or not, I also looked up to tremendously for their incisive research and warm community building. Thank you to Margaret Kroll, Steven Foley, and Jed Pizarro-Guevara for modeling what it was to be a Santa Cruz psycholinguist. Jed, thank you for many distanced meetings, whether they were in my driveway during April-June of 2020, or over zoom across the country. Thank you, Tom Roberts, great refractor, for being a fearless instructor for two first-time TAs and for sharing the wisdom of your great great (and so on) grand shoodtha. Thank you, Kelsey Sasaki, for long car rides to LA, and long walks along a (hopefully growing) number of bodies of water, from the Monterey Bay to Bahía de La Concha. Thank you Nick Van Handel, from our humble beginnings in a sprinkle pit until now, it was such a joy to do the work of one person together, like the one person that we were. I could always count on being excited about the same research questions as you (and un-

excited about the same research questions). Thank you Lisa Hofmann, for many meaningful wanderings of the streets of Riga. I'd like to spend a full mentat phase appreciating Richard Bibbs: thank you for embarking on the long and treacherous dissertation-writing path with me, for games and floats and dances. I'd like to thank Ben Eischens for - hey, Ben, can I ask you a question? (I'll assume you've said yes) Can I mention how grateful I am for you in my acknowledgments? (I'll wait for your response.) And thank you to Jérémie Beauchamp, for bringing such a warm and joyful energy to the department.

To those who joined after me, and have then been traveling alongside me on this path since, thank you. To Duygu Demiray, my father and my son: I hope to make you proud. To Yağmur Kiper, for understanding how my brain works, and for walking ever so quickly beside me. To Mandy Cartner - would you *will* accept my appreciation for you? I am dancing at you, always. To Elif Ulusoy, for talking through Experiment 6 with me, and watching me intently as I made small, inconsequential decisions. To Vishal Arvindam, for the music, and facing the elements as a neighbor and TA. To Yaqing Cao, for joining me in several once-in-a-grad-school experiences. To Eli Scharf, for joining the lab and conversations about memory for a bit. To Matthew Kogan, for sharing not just the path of a Santa Cruz psycholinguist, but also the path directly from Russian-speaking child to Russian-speaking grandparent.

To Lalitha Balachandran, thank you for our giant discourse tree, beginning circa 2016 when I was your lab potato, and running all the way to the present. Our collaborations, in research and course design, are some of what I'm proudest of in graduate school, and I cannot imagine doing this without regularly spending evenings charlie-and-the-chocolate-factory-ing with you. To Allison Nguyen, so much of what it has meant to be in graduate school for me is interwoven with

precious moments of our friendship, from our fateful first meeting to walking through neighborhoods with you. You've given me advice on everything from job applications to mentorship to how to not panic while knitting. When you left Santa Cruz, I felt a little piece of Santa Cruz leave with you, but I know we will always be just a beacon-lighting away. To Niko Webster, thank you for journeys of all kinds. It hasn't mattered where we were going, to Coffeetopia, to Berkeley, to a switch in roles mid-tanda, or off to the clouds, I've learned and grown so much from the conversations we've had along the way.

Thank you to you all, and those not mentioned here whose time overlapped with mine in some shape or form at UCSC. You have truly all shown me what support, community, and solidarity look like.

To grad-school-adjacent friends: thank you Ryan Johnson, for many joyous feasts and crafting a Stephanie-friendly bike route; thank you, Alice Shen, for feeding my love of linguistics in undergrad, and being the best possum since; thank you, Evan Morris, for welcoming me into the astronomy department, into your pandemic bubble, and into whatever future adventures we might have; and happy birthday to you, Dean Thomas and Lena Nahorna, each and every day.

Santa Cruz itself has been such a special place to spend the last few years. To others in Santa Cruz who made this place feel like home: thank you Kara, Ally, Fran, Anjelica, Will, Cricket. One precious thing I am grateful to experience in graduate school is the feeling of being at home in someone else's home. Thank you to cin aliler and the BFS for giving me this feeling. And thank you to Ally, Fran, Dulyn, and Mirella: I have always felt so warm and welcomed sitting around a dinner table with you. To Mark Hollingshead, thank you for sharing your wisdom, on all matters of life and combat. I'm happy to call you a friend, and I hope you don't mind me calling you a role model. To the baristas of Coffeetopia, especially

Jay, thank you for a reason to look forward to Monday mornings. To the dancers of Abrazo Queer Tango, thank you for holding me. Karen, you teach me something new each week about what it means to be in a community. Long live queer tango! And to the nature: I am so grateful to every California poppy along my evening walk, every unique landscaping choice, the damp evening air of the marine layer coming in, and every redwood tree that's offered me shade.

To Doriane Weiler, my non-romantic domestic partner, I don't think I could have done any of this without you. This being: a dissertation, surviving a pandemic, moves, fires, and not knowing what to eat for dinner. You've taught me so many different ways to build a home, and how to (begrudgingly) accept change. There's still more to explore - thank you for exploring with me.

To my friends who supported me from afar, thank you: Mady Klein, Brinna Michael, Jacquie Marques, Jeannine Cullen. To Mariam Taleb, my twelve, thank you for counting my dragonflies.

And finally, thank you to my family. To my grandparents, Manny, GG, and Glikeria, who have each encouraged my pursuit of education in their own way. To my parents, David and Larisa, and sister, Isabelle, thank you for cheering me on, whether I'm running down a hallway or writing a dissertation. I love you.

Contribution acknowledgments

Experiments 4 and 5 are joint work with Lalitha Balachandran. Any future submissions of this work for publication will be jointly first-authored.

A number of undergraduate research assistants assisted with data collection and stimulus creation. I'd like to acknowledge Ashley Ippolito for data collection assistance in Experiments 1-3; Noa Nevo and Olivia Aurtenechea Eguiron for data collection in Experiment 2; Cal Boye-Lynn and Sadira Lewis for data collection in Experiment 3; and Jenna Maas and Josh Lieberstein for their assistance in creating stimuli for Experiment 4.

Versions of the work presented here have been presented at the Conference on Human Sentence Processing (2020, 2021) and at Architectures and Mechanisms for Language Processing (2023), as well as locally at lab meetings at UC Santa Cruz. I'd like to thank the audiences for their feedback.

Research throughout this dissertation was supported by the NSF GRFP #1842400.

Chapter 1

Introduction

A number of cognitive processes rely on fast and rapid access to representations stored in memory. As it pertains to the processing of language, we rely on both the long-term storage of a vast mental lexicon, as well as fast on-the-fly processes in memory to successfully comprehend incoming language incrementally. For example, a reader encountering the sentence in (1) should be able to quickly realize, at the verb *is*, that this sentence has failed to follow Mainstream American English subject-verb number agreement. Indeed, readers pause upon noticing this discrepancy, which informs us of (at least) two facts: there is some systematic representation of number information in the mental representations of the encountered noun and verb, and that there is a rapid process of establishing how well-matched the noun and verb are on the basis of that number representation.

(1) The *dogs* that I saw at the park *is* chasing some squirrels.

This dissertation explores both of these interactions with memory, exploring how notions of similarity arise from the way words are represented in memory and how encoding and retrieval processes may be disrupted or facilitated during incremental language processing.

1.1 Central questions

The role of memory in incremental language comprehension will be explored from two angles: from the study of sentence processing, looking at how linguistic properties give rise to interactions in memory, and from the study of memory broadly, looking at how memory for non-linguistic objects or unstructured word lists can inform our understanding of encoding, maintenance, and retrieval processes during sentence comprehension.

We will look at instances of retrieval during sentence processing in which there is one grammatically possible retrieval target, separated from the retrieval site by one or more similar distractors, as in the case in (2). The example in (2) involves both subject-verb dependency processing, in which *the boys* must be understood as the subject of the matrix verb *drink*, and dependency processing within a relative clause (RC), in which *the boys* must be understood as the object of the RC verb *befriended*. More on each of these dependency types, including a brief survey of previous literature on the resolution of such dependencies, can be found in Section 2.1. To return to the specific example at hand, the intended retrieval target for each dependency is the matrix subject *the boys*, but in each case there is a similar intervener, *the girl*.

(2) The boys who the girl befriended last week drink tea.

Retrieval during sentence comprehension is often modeled using an associative, cue-based model of memory in which information is represented in bundles of encoded features (McElree, Foraker, & Dyer, 2003; Van Dyke & Lewis, 2003). A search in memory is then launched via retrieval cues which make reference to those features. In the case of (2), the two DPs may be encoded with features that represent grammatical number information as well as animacy information. The speed and accuracy of subsequent retrieval would depend on the strength of association between the retrieval cues and the target item; the strength of association, however, is weakened as the number of items associated with a particular cue increases (Watkins & Watkins, 1975; Lewis & Vasishth, 2005; Nairne, 2006; Van Dyke & McElree, 2006). In this case, their featural representations would match along the dimension of animacy and mismatch along the dimension of grammatical number. If the intervener had instead been *the girls*, such that the target and intervener matched in both animacy and number, the strength of association between the retrieval cues and the target item would be weaker.

A model like this provides a framework to capture effects that arise due to similarity between a target and intervener along a dimension that is directly referenced at the retrieval site. However, as this dissertation will explore, effects of similarity may also arise when the particular dimension of similarity manipulated is not referenced at all by the retrieval site. It is argued that such effects are therefore driven by interference during an earlier process, encoding. To understand the underlying mechanisms for such interference, the dissertation will overview several different theories of how encoding interference may arise. The goal of the work presented here is not to definitively determine that one theory of the underlying mechanism uniquely describes all encoding-like effects in sentence comprehension; in fact, these theories are not necessarily mutually incompatible. Instead, I hope to explore which theories are relevant in specific contexts during sentence comprehension.

The theories of memory examined by this dissertation differ in how they account for encoding-like effects. One group of theories, which I will refer to as Feature Change theories, derive such effects from some kind of disruption to the featural representations themselves (feature overwriting: Oberauer & Kliegl, 2006; Oberauer, 2009; superposition: Oberauer, Farrell, Jarrold, Pasiecznik, & Greaves, 2012). Another group of theories, which I will refer to as Feature Interaction theories, model the encoding process in such a way that features not referenced by retrieval can create interference (intra-feature association: Logačev & Vasishth, 2011; the Temporal Context Model: Howard & Kahana, 1999, 2002; and Self-Organized Sentence Processing: Smith, 2018; Smith, Franck, & Tabor, 2018). The specifics of these theories are expanded upon in Section 2.3. One overarching theme, though, across many of these theories is the role of the following two factors in generating interference between two or more items: i) similarity in the featural representation of those items, and ii) some form of closeness during the initial encoding process, whether we understand that as temporal contiguity, simultaneous activation, or a shared encoding context. This dissertation aims to explore both of these factors, and explore the extent to which we can relate findings that are well-established in non-linguistic memory to sentential contexts. At the highest level, the experiments in this dissertation address the following two questions:

1. What forms of similarity give rise to interference?

2. How can we understand the concept of closeness in a linguistic context?

To address the first question, I explore the nature of featural representations, looking for the most part at semantic features in instances where a later retrieval site would not select for those particular features. To address the second question, I explore different configurations that result in the encoding of two nouns in close temporal proximity, follow insights from research on memory generally to make specific predictions about patterns of recall and recognition in sentential contexts, and explore the role of unpronounced reactivation in the memory for linguistic elements.

1.2 Roadmap

In Chapter 2, I overview the crucial findings of interference both in sentence processing and the study of memory generally. I also introduce the linguistic dependencies that will be used in many of the experimental designs in this dissertation (subject-verb dependencies and gap processing) and several specific theories of how interference can arise (both Feature Change and Feature Interaction, as described above), along with the implications of these for linguistic contexts.

In Chapter 3, I present two reading studies which illustrate a distinction between encoding and retrieval interference by showing interference at a later retrieval site on the basis of semantic features, when that retrieval site does not carry any semantically selective information.

In Chapter 4, I further explore the nature of closeness, testing what linguistic cues can serve as boundaries of encoding contexts. First I present a follow-up reading study to the studies presented in Chapter 3, which manipulates closeness differently from Experiments 1 and 2. Then, I provide further literature review on the study of context in list-memory, and present two studies (recognition, and recall) which seek to replicate classic findings in list-memory research using sentential stimuli. This chapter presents preliminary evidence that clause boundaries can serve to delineate encoding contexts in memory.

In Chapter 5, I examine whether interference may occur on the basis of semantic similarity during co-activation in dependency resolution. I then explore what features may be used predictively when forming predictions about an upcoming wh-dependency, specifically thinking about the encoding of thematic information.

Finally, in Chapter 6, I conclude the overarching findings about both similarity and closeness. I then lay out a potential framework for future work in this area and how this research may connect to other research areas, for example thinking about how similarity-based interference may have consequences in the study of predictive processing, and how encoding interference may occur in unexpected ways in various experimental paradigms.

Chapter 2

The Theoretical Landscape

Successful comprehension of language relies on a number of processes occurring simultaneously or in quick succession, many of which rely on the ability to form and later access representations in memory. In this dissertation, I will focus on the creation and resolution of dependencies between syntactic elements during incremental sentential comprehension as an area that can reveal properties of the memory mechanisms that underlie language processing. The theories of memory that will be addressed here make specific predictions about instances of shared activation and proximity during encoding. Dependencies that may span long distances and rely on retrieval mechanisms for resolution therefore are ideal case studies.

2.1 Dependency resolution

2.1.1 Subject-verb dependencies

The first form of dependency resolution that will be examined in this dissertation is the process of subject-verb dependency resolution. In particular, this dependency will be examined as a process that relies heavily on accurately encoding and then successfully retrieving the featural representation of the subject upon reaching a verb. Over the course of this dissertation, two types of features will be addressed. First, grammatical features such as number or grammatical gender, and second, semantic features that address the inherent properties of an entity or the ways in which the entity can interact with the world around them. I make a distinction here as there is reason to think they may be encoded differently. For example, a feature like grammatical number, will have a finite, small number of possible values (in English, it is a binary distinction). For semantic properties, such as physical characteristics, the range of possible values is much greater.

In languages where the form of the verb depends on grammatical features of the noun, during the comprehension some form of feature checking or tracking must occur, as indeed readers are quick to notice mismatching features and spend longer reading the verb in (3b) compared to the verb in (3a), as demonstrated by Wagers, Lau, and Phillips (2009).

(3) a. The old key$_{SG}$ unsurprisingly was$_{SG}$ rusty from years of disuse.

b. * The old key$_{SG}$ unsurprisingly were$_{PL}$ rusty from years of disuse.

In addition to rapidly assessing the correspondence between grammatical features such as number or grammatical gender, integration with the verb is necessary to form a representation of the meaning of the sentence: understanding the thematic role of the subject and forming a conceptual representation of the event being described. Just as feature matching allows for fast detection of errors in subject-verb feature agreement, thematic and semantic processing allows for fast detection of implausible or semantically anomalous sentences. Rayner, Warren, Juhasz, and Liversedge (2004) specifically examined the reading of implausible vs. anomalous sentences, using stimuli as the sentences in (4) to distinguish between instances of unlikely events which are possible but unlikely (4b) and events which are read as anomalous because at least one thematic relationship cannot be possible under normal circumstances (4c), finding that a penalty for anomaly emerges earlier than a penalty for implausibility.

(4) a. John used a knife to chop the large carrots for dinner. (plausible)

b. John used an axe to chop the large carrots for dinner. (implausible)

c. John used a pump to inflate the large carrots for dinner. (anomalous)

This pattern of findings can be understood as a gradient effect reliant on the severity of a plausibility violation, or, as will be explored in this dissertation, a reflection of categorically different types of implausibility.

If we imagine that semantic properties are encoded alongside grammatical features, we could imagine that something about their affordances, or ways to be interacted with, or possible thematic roles, may be encoded. Therefore, an anomaly as in (4c) can be detected by virtue of feature matching, similarly to the verb-agreement findings in (3). On the other hand, (4a) and (4b) are thematically well formed, affordances respected, and the comprehender must draw on world knowledge more broadly to know that a knife is a more common carrot-cutting instrument than an ax.

The matter of how these specific properties might be encoded will be returned to in Section 2.3, where specific mechanisms for the encoding and retrieval of lexical features will be revisited.

2.1.2 Processing gaps

The second form of dependency resolution that will be examined in this dissertation is the process of associating a syntactic element (a filler) with a distant position that is not pronounced (a gap). For the purposes of this dissertation, two aspects of gap processing will be important in creating instances of possible interference: i) (re-)activation of the filler at a gap site; ii) active maintenance and eager attempts at resolving the dependency prior to establishing the final, correct gap position.

The processing of gaps will be addressed in two forms: wh-fillers associated with a position from which they have been dislocated (5), and relative clause constructions in which the relative clause head is associated with a position within the relative clause (6).

(5) *Which shoes* did everyone favor _ at the dance event on Saturday?

(6) I remembered *the shoes* that everyone favored _ for the dance event on Saturday.

Association of the dislocated DP filler with its corresponding gap site is crucial to generate the proper thematic interpretation for the sentence, e.g. in both (5) and (6), *the shoes* must be interpreted as the direct object of *favored* consequently receiving a stimulus thematic role. In the following two subsections, I will briefly review the current understanding in the literature about this association process in both relative clauses and wh-filler-gap dependencies.

Relative clause processing

In relative clauses (RCs), the associated gap occurs within a clausal modifier of a relativized noun, the RC head.

Psycholinguistic evidence demonstrates that during this association process, the representation of the filler is reactivated, such that, for a sentence like (7), semantic associates for a relative clause head (*boy*) were only primed following the relative clause verb (*accused*, probe 2) and not prior (probe 1) in cross-modal lexical priming (Swinney, Ford, Frauenfelder, & Bresnan, 1988; Nicol & Swinney, 1989). This suggests that reactivation includes semantic information about the filler. Other priming studies of reactivation at gap sites demonstrate priming effects for rhymes of the filler, suggesting that the phonological representation, too, is reactivated (Tanenhaus, Carlson, & Seidenberg, 1985).

(7) The man saw the boy$_i$ that the crowd at the party **1** accused t_i **2** of the crime **3**

One well-established finding in head-initial languages such as English is a preference for subject relative clauses (SRCs) compared to object relative clauses (ORCs), as evidenced by both reading times and question accuracy following interpretation (e.g. Wanner & Maratsos, 1978; Ford, 1983; Just & Carpenter, 1992). This would look like a penalty in a sentence such as (8b) compared to a sentence like (8a). One reliable exception to the ORC penalty occurs in the case of inanimate RC heads (e.g. Traxler, Morris, & Seely, 2002; Traxler, Williams, Blozis, & Morris, 2005; Gennari & MacDonald, 2008). In this case, in a configuration like (8d), an ORC with an inanimate RC head and animate RC subject, there is either a reduced penalty or no penalty in comparison to its SRC counterpart (8c). The reduced ORC penalty occurs even with typically-inanimate RC heads in cases where there has been a preceding context that anthropomorphizes the typically-inanimate object as an agentive entity (Rich et al., 2022). This suggests that the animacy effect on the ORC penalty is not a matter of world-knowledge, but rather the understanding and incidental encoding of the noun in that particular moment (although it is possible subjects simply downweighted animacy as a cue in these cases).

(8) a. The reporter that _ composed the article caused a big scandal.

b. The reporter that the article bothered _ caused a big scandal.

c. The article that _ bothered the reporter caused a big scandal.

d. The article that the reporter composed _ caused a big scandal.

[example from Gordon and Lowder (2012)]

A number of explanations have been suggested for the ORC penalty: a preference to resolve the dependency as soon as possible, a penalty for interference in memory, a preference for a more frequently encountered structure, and a preference for an agentive (correlated with subject position) interpretation for the RC head, overviewed by Gordon and Lowder (2012). Evidence for the role of interference in memory will be explored further in Section 2.2. One take-away, however, from this literature, is the role of animacy as a feature which may prompt structural expectations.

In thinking of the predictive processing that occurs in the comprehension of RCs, it is worth noting that the beginning of the RC, in English at least, is not always clearly marked. For example, in (9), even two words following the noun, the continuation is not unambiguously a RC.

(9) I watched the movie the other...

a. ...day.

b. ...professor recommended.

In contrast, wh-fillers are unambiguous indicators of an upcoming gap.

Wh-elements

There are several possible strategies a parser could follow to resolve the dependency between a wh-filler and its associated gap site. The parser could wait, postulating a gap only at the end of an utterance and completing thematic integration once all pronounced elements had already been encountered. Instead, overwhelming evidence suggests that the gap postulation process is an active and eager process (Crain & Fodor, 1985), with the parser postulating a gap at all grammatically possible gap positions, formalized in the Active Filler Strategy (Frazier, 1987; Frazier & Clifton Jr, 1989). This highly-eager process is evidenced by instances of error whenever an earlier possible analysis is incorrect, and a postulated gap site is filled, resulting in a *filled gap effect*, observed by Stowe (1986) in sentences such as (10b) compared to sentences like (10a). In reading studies, this corresponds to longer reading times at the filled gap position (*us*).

(10) a. My brother wanted to know if Ruth will bring us home to Mom at Christmas.

b. My brother wanted to know *who* Ruth will bring us home to _ at Christmas.

What information is maintained in the representation of the filler, and what can be used to guide the gap filling process? It has been established that readers do not postulate gaps within islands, demonstrating an adherence to grammatical constraints on extraction. Many other attempts to 'turn off' active gap filling have been made, with evidence instead pointing to an active process that will postulate an object gap even following an intransitive verb (Omaki et al., 2015).

One area of research has surrounded a study of what aspects of the filler are actively maintained. If the semantic properties of the noun (and for that matter, the verb) are not used for the initial postulation of the gap, what information can we say with certainty is maintained during this process, versus what may be reactivated at the gap site?

Lee (2004) demonstrated that the syntactic category of the filler is maintained: there is no filled-gap penalty in the DP subject position (*Irene*) for a PP filler (11b) compared to a DP filler (11a).

(11) a. That is the laboratory [which]$_{DP}$, on two different occasions, Irene used a courier to deliver the samples to $\_{DP}$.

b. That is the laboratory [to which]$_{PP}$, on two different occasions, Irene used a courier to deliver the samples $\_{PP}$.

Furthermore, the animacy of the filler is represented, and affects the likelihood of postulating a subject gap (Wagers & Pendleton, 2016). Expanding on the DP/PP manipulation of Lee (2004), Wagers and Pendleton (2016) manipulate the animacy of the filler and find a subject gap penalty only in the case of the animate filler (12).

(12) DP conditions

a. The scholar looked to his aging mentor who, only recently, the academic community owed much of their findings to $\_{DP}$

b. The scholar looked to the controversial text which, only recently, the academic community owed much of their findings to $\_{DP}$

Animacy information has also been shown to affect the likelihood of gap postulation on the basis of animacy-related selectional restrictions of the verb. Ness and Meltzer-Asscher (2019) demonstrate that the plausibility effect, in which a penalty for an implausible object suggests that gaps are postulated regardless of verb plausibility, emerges specifically in cases where the verb does not have an animacy restriction. This suggests that animacy information is maintained during the active gap filling processing and can actively guide the postulation of gaps at a rather fine-grained level.

Finally, number information is clearly maintained. Kim, Brehm, Sturt, and Yoshida (2020) find an earlier and larger number-mismatch penalty at a plural verb *are* if the subject gap associated with the verb is the initial site of dependency resolution (13a) compared to an instance where that subject gap is a reactivation of the filler following an earlier resolution of the dependency (13b). Notably, this difference cannot be attributed to linear distance, as in the relevant subject gap is an equivalent distance from the filler in all conditions.

(13) Ungrammatical-SG conditions of Experiment 3 in Kim et al. (2020)

a. Which mistake$_{SG}$ in the program that will be disastrous for the company $\_{SG}$ are$_{PL}$ harmful for everyone involved.

b. Which mistake$_{SG}$ in the program $\_{SG}$ will be disastrous for the company and $\_{SG}$ are$_{PL}$ harmful for everyone involved.

2.1.3 Avenues for interference

In sum, both subject-verb dependencies and filler-gap dependencies rely on encoded featural representations of the subject or the filler, respectively. This dissertation will explore the ways in which these dependency resolution processes are susceptible to interference, both at the level of the initial encoding of features and later retrieval on the basis of those features.

The reference of features during language processing has often been explained by an associative, cue-based model of memory. In this model, linguistic information is encoded as bundles of features which are accessed in a content-addressable fashion (McElree et al., 2003; Van Dyke & Lewis, 2003) . These feature bundles encode a variety of grammatical or extragrammatical properties, which can be leveraged by retrieval cues that reference those features. Errors may arise, however, due to imperfections during either encoding or retrieval, especially in the presence of a similar distractor. The next two subsections overview examples of similarity-based interference in the literature, and more concrete examples of how such interference can be modeled.

2.2 Similarity-based interference

Evidence for similarity-based interference can be found across sentence structures and tasks, and as either facilitatory or inhibitory.

To begin, we can revisit the case of the ORC penalty. In addition to animacy, the ORC penalty is sensitive to the similarity between the RC head and the RC-internal arguments, emerging when the relativized argument (*banker*) and the internal argument are similar, e.g. a penalty for (14b) compared to (14a). When they are more distinct, as in (14d), the ORC penalty is much reduced (Gordon, Hendrick, & Johnson, 2001, 2004; Gordon, Hendrick, Johnson, & Lee, 2006) .

(14) a. The banker [that _ praised the barber | a barber] climbed the mountain.

b. The banker [that the barber | a barber praised _] climbed the mountain.

c. The banker [that _ praised Joe | you | everyone] climbed the mountain.

d. The banker [that Joe | you | everyone praised _] climbed the mountain.

In (14), the appropriate notion of similarity is something like the referential type of the DP co-arguments or their size, but similarity can lead to difficulty when it is more abstract. Van Dyke and Lewis (2003) discovered that similarity in syntactic position (or abstract case) could cause difficulty in subject-verb dependency formation. In (15), the entire complex DP, *the student who...* must be integrated as the subject of the predicate given by the TP *was standing in the hallway*.

(15) a. The secretary forgot the student [who was waiting for the exam] was standing in the hallway.

b. The secretary forgot the student [who knew the exam was important] was standing in the hallway.

In (15a), this integration is easier compared to (15b), as judged by reading times and comprehension measures. In (15b), unlike in (15a), the complex DP contains another subject-like constituent, *the exam*, occupying the syntactic position of the subject of a tensed clause. Van Dyke and Lewis (2003) argued that integration at the verb is guided by cue-based retrieval in a content-addressable memory. When the cues provided by the verb do not uniquely point to a single constituent, because multiple constituents are similar to the target, then greater difficulty is encountered. This subject-verb integration effect has been replicated for related structures (Van Dyke, 2007; Van Dyke & McElree, 2011; Arnett & Wagers, 2017). Moving beyond the relationship between arguments and predicates, evidence for similarity-based interference has been in a wide range of dependency relationships, such as agreement dependencies (Wagers et al., 2009; Dillon, Mishler, Sloggett, & Phillips, 2013), antecedent-reflexive dependencies (e.g. Badecker & Straub, 2002; Chen, Jäger, & Vasishth, 2012; Dillon et al., 2013; Jäger, Engelmann, & Vasishth, 2015; Patil, Vasishth, & Lewis, 2016), antecedent-reciprocal dependencies (D. W. Kush, 2013; D. Kush & Phillips, 2014), negative polarity items (Vasishth, Brüssow, Lewis, & Drenhaus, 2008; Parker & Phillips, 2016), and sluices (Harris, 2015), to name a few.

Apart from what constructions are involved, there is still debate over what types of similarity are capable of causing interference. Many core grammatical features have been explored, including number (Wagers et al., 2009), grammatical gender (Sturt, 2003; Badecker & Kuminiak, 2007; Slioussar & Malko, 2016;

Villata, Tabor, & Franck, 2018; Lawn, 2020), case (Fedorenko, Babyonyshev, & Gibson, 2004; Logačev & Vasishth, 2011; Arnett & Wagers, 2017; Avetisyan, Lago, & Vasishth, 2020), and syntactic position (Van Dyke, 2007; Arnett & Wagers, 2017). Semantic features have been considered as well, such as animacy (Van Dyke, 2007), although this could also be considered a grammatical feature, and type of referring expression (Gordon, Hendrick, & Levine, 2002).

What causes similarity-based interference? In some configurations, increased similarity among constituents creates interference when one dependent needs to access another, as a species of retroactive interference. Here the cues used to retrieve a target constituent will be more or less effective if there are recent intervening distractors that are similar to that target. But similarity can also affect encoding, when the target and the distractor are first encountered and integrated into context. If similarity between a target and a nearby distractor can weaken or distort the representation of the target, the effects could nonetheless manifest during later retrieval. Even if cues used at retrieval uniquely and strongly matched only the target, its reduced strength or quality could influence the speed or success of integration. It can therefore be challenging to distinguish between retrieval and encoding interference.

Agreement attraction seems to present a clear case for retrieval interference (Wagers et al., 2009). But it may also reflect encoding interference. The marking and morphing theory (Eberhard, Cutting, & Bock, 2005) supposes that an ambiguous or faulty representation of number on the head noun phrase is what drives agreement attraction. This theory thus proposes that an early representational degradation results in difficulty downstream. Some recent modeling evidence suggests that both retrieval and encoding interference may be at play in agreement attraction (Yadav, Smith, Reich, & Vasishth, 2023). In the following section, we will address specific models of encoding interference in greater detail.

2.3 Encoding and retrieval

In a content-addressable system of memory, successful retrieval depends on the strength of association between a retrieval cue and the intended retrieval target, which is a function of how well and how precisely the retrieval cue matches the target (Nairne, 2006). As the number of items with shared features increases, the strength of the retrieval cue for the target decreases, a state of affairs termed cue overload (Watkins & Watkins, 1975). For this reason, the presence of similar interveners decreases the probability that the intended target will be retrieved, and any processes that rely on retrieval will consequently take longer or be less accurate. One way to model this effect is via the construct of activation, as in the cognitive modeling architecture Adaptive Character of Thought-Rational (ACT-R; Anderson & Lebiere, 1998; Anderson, 2005). Cue overload reduces the activation of the target item relative to distractors, and this is mapped either to longer retrieval latencies for the target (Lewis & Vasishth, 2005) or to decreased retrieval probabilities (McElree, 2006), or both. For example, thematic integration of a complex subject and verb is harder if there are multiple 'subject-like' encodings in memory to compete with the target (Van Dyke & Lewis, 2003), as in (15) above. In that example, the target is the matrix subject *the student* and the distractor is the embedded subject *the exam.* The match of the distractor to the cues used at retrieval not only leads to activation of the distractor, but also decreases the activation of the target, which leads to longer reading times at the verb in a sentence like (15b) compared to a sentence like (15a), where the distractor is a less good match.

Cue overload is not the only cause of difficulty at retrieval, however. Similarity can also affect the formation of item representations during encoding, where it can impact how distinctively nearby items are encoded or whether features are bound to the right items. Such encoding interference can be conceptualized in a number of ways. Some models hold that, when two similar items are processed together, their featural composition is directly altered or changed in a way that decreases their retrievability. "Feature Change" models include feature overwriting (Oberauer & Kliegl, 2006; Oberauer, 2009) and superposition (Oberauer et al., 2012). Other models hold that encoding interference can be driven by featural interactions that don't directly change the underlying item representations. Instead of disruption in the featural representation of any given item, encoding interference is a consequence of the nature of the encoding process. "Feature Interaction" models include models of intra-feature association (Logačev & Vasishth, 2011), the Temporal Context Model (Howard & Kahana, 1999, 2002) and aspects of Self-Organized Sentence Processing (Smith et al., 2018; Smith, 2018).

Feature Change models rely on distributed representations, where items consist of sets of features. When an item is activated, that activation is understood as the activation of the corresponding features. In feature overwriting (Oberauer & Kliegl, 2006; Oberauer, 2009), when two items are processed in close temporal proximity, they compete for the activation of any overlapping features, such that each shared feature may be activated by one item, and not the other. This results in a degraded representation, where both items can suffer feature loss. The proportion of remaining features for each item (Pr_i) is given by the equation in (16), which depends on two variables: C, or the proportion of features shared by the items, and n, the number of items in working memory at the same time. The insight of this equation is that any given item is expected to interact with $n - 1$ other co-active items and in each interaction it will lose half of its shared features

(C/2).

(16) $\Pr_i = (1 - C/2)^{n-1}$

The proportional feature loss is then hypothesized to be equivalent to the loss of activation for the affected items in memory. Feature overwriting therefore results in overall lower activation for the affected items, which would consequently hinder subsequent retrieval processes.

Another important facet of distributed item representations is that each item only activates a limited number of features, necessarily a subset of all potential features available. Therefore, subsequent encodings are understood as different patterns of activation across the same network of features in quick succession. Superposition describes interference that results from these rapid and distinct patterns of activation (Oberauer et al., 2012). Under this view, the representation of any item is distributed across a weight matrix. Subsequent encodings are also distributed across the same weight matrix, with similar items strengthening pre-existing associations and dissimilar items creating disruptions by creating new associations to the same position, thus weakening pre-existing associations. High-similarity items are predicted to lead to lessened superposition interference, because similar items strengthen one another. However, they can also lead to more erroneous intrusions during retrieval, as one feature is effectively shared by multiple positions. In a model such as this one, it becomes clear that encoding interference could affect not only the quality or strength of the featural representation for an encoded item, but also the precision with which a given feature is linked to the intended position.

In contrast to the Feature Change accounts of encoding interference, Feature Interaction models capture encoding-driven interference effects via interactions between stable featural representations during encoding or retrieval. In one of these accounts, dissimilar interveners can also lead to interference under the assumption that feature-to-feature associations are encoded as well as feature-to-item associations (Hommel, 1998; Logačev & Vasishth, 2011). Under this view, interference arises in cases of conflicting bindings: instances where a single feature (e.g. animate) is bound to one feature in one item (e.g. plural) and a mutually incompatible feature in another (e.g. singular). This theory therefore predicts that the greatest level of interference will occur in the case of two items with partially-matching features, more so than for a case of two items with fully overlapping features. As a consequence, the relationship between featural overlap and similarity-based interference is non-monotonic.

In the Temporal Context Model, items are encoded along with the temporal context in which they are encountered, a context which includes neighboring items. Then, during retrieval, the full encoding context, not just the individual item, is retrieved, thus reactivating neighboring items as well (Howard & Kahana, 1999, 2002). Evidence for this model comes from a wide range of unstructured list memory studies that demonstrate a robust effect of presentational contiguity. After recalling an item from list position i, participants are most likely to next recall the item from position $i + 1$ (Healey, Long, & Kahana, 2019). Semantic similarity enhances this effect: highly similar pairs are more likely to be recalled in succession than low-similarity pairs. This effect weakens significantly if the presentation interval between items is increased, suggesting that this effect of similarity is sensitive specifically to temporal proximity and not only presentation order (Howard & Kahana, 2002). If two similar items are encountered in close proximity in a sentence, in a joint context, it may be difficult to retrieve the first item without also retrieving that context, thus activating the second item. Furthermore, by virtue of being encoded in close proximity, i.e. in the same encoding context, two items will be more featurally similar compared to an instance where the same two items were encoded at a greater distance. The important insight of this model is that similarity is not only dependent on features determined by grammatical or semantic features, but also on features determined by the circumstances of encoding.

Lastly, other models do not distinguish between encoding and retrieval as separate processes. In Self-Organized Sentence Processing (Smith et al., 2018; Smith, 2018), both encoding and retrieval are captured by one single process: a structural building process in which a parse gradually emerges through local attachments between words. Each basic unit of structure building, a syntactic treelet, contains syntactic and semantic featural information and potential attachment sites that reflect selectional preferences. Attachments are formed on the basis of featural similarity, with units that are more similar competing for attachment. Smith, Franck, and Tabor (2021) demonstrate this with complex DP subjects such as the canoe by the kayaks, finding longer reading times at a later singular verb (*was damaged*) in the case of a semantically similar second noun (*kayaks*) compared to an unrelated noun (*canoes*). To capture effects such as these, effects of similarity at choice points that shouldn't reference semantic cues, the model makes several assumptions. One assumption is that representations of semantic features are multi-dimensional, such that words like canoe and kayak would share multiple features that reflect boat-like-properties. Another assumption is that upon encountering a subject like canoe, the parser would expect a verb that would be compatible with boat-like properties. As a result, similarity-based interference could emerge on the main verb in cases where the verb is compatible with boat-like properties, even if the verb does not strictly-speaking require a boat-like subject.

The studies in this paper aim to highlight patterns of effects that necessitate a driving role of interference from encoding, rather than or in addition to interference that stems solely from retrieval. As both encoding interference and retrieval interference could influence the retrieval process, disentangling the source of observed interference at the retrieval site becomes a non-trivial task. The current study aims to do just that, looking specifically at similarity across semantic features which are not referenced during retrieval.

2.4 Explorations in this dissertation

Many, but not all, of the models described above that distinguish encoding from retrieval emphasize the role of contiguity in creating encoding interference. Temporal contiguity plays a critical role in creating the conditions for feature overwriting during co-activation, creating conditions necessary for interference by superposition, or heightening similarity on the basis of a shared encoding context in the Temporal Context Model. In expanding from studies of word lists to full sentences, there are several possibilities for how hierarchical structure and linguistic boundaries may contribute to an understanding of proximity during encoding. This will be examined in a number of ways across the length of the dissertation.

Chapter 3 will first examine encoding interference from the perspective of the featural representations, and will take temporal proximity as a starting point from which to address contiguity. Chapter 4 will examine more closely how linguistic boundaries can inform the creation of encoding contexts in memory through the Temporal Context Model. And finally, Chapter 5 will address the instances of unpronounced reactivation, and how such activation may contribute to interference.

Chapter 3

Encoding interference

This chapter presents a series of reading studies to test whether an earlier instance of encoding interference would cause difficulty at a downstream retrieval site. One central aim of this chapter is to contribute to the body of literature that demonstrates that while both encoding and retrieval interference may both occur, the effects of each are distinguishable and identifiable.

3.1 Isolating encoding interference

Definitively determining that observed similarity-based interference is a result specifically of encoding interference is no trivial task. In many cases, models both of encoding interference and cue-based retrieval interference predict some form of difficulty at the retrieval site. Therefore, careful experimental manipulations are required to distinguish between the potential origins of such interference.

Studies designed to directly investigate encoding interference in language comprehension must introduce an intervener with a high degree of featural overlap to the target, but also ensure that the manipulated features are not those referenced by the retrieval process. If similarity in features orthogonal to the relevant retrieval cues results in interference at the retrieval site, it is reasonable to suspect that the retrieval process alone cannot account for the observed interference. The example of reduced ORC interference with referential dissimilarity, as discussed for (14), is a plausible example of this: the difference between an DP subject with descriptive content, like the barber, versus a pronoun, you, is potentially relevant at encoding but it isn't the kind of property that a RC verb in English is sensitive to, as a requirement on its dependents. Villata et al. (2018) make this point more forcefully by manipulating grammatical gender in object retrieval in Italian, using sentences such as (17).

(17) The dancer that the waiter has surprised _ drank a cocktail with alcohol.

The authors manipulated the grammatical gender of both the dancer and the waiter, resulting in two gender match conditions (MM, FF) and two gender mismatch conditions (MF, FM). Crucial to their design was the fact that the verbs were not marked for grammatical gender, making it implausible that gender would be used as a retrieval cue. Therefore, if gender match conditions were more difficult than gender mismatch conditions, the interference must have stemmed from an earlier process: during encoding. That is, in fact, what the authors find. Reading times were longer on the verb for the gender match condition than the gender mismatch condition, despite gender being unavailable as retrieval cue.

The results from Villata et al. (2018) demonstrate a pattern of interference wherein encoding interference on the basis of grammatical features becomes clear only at a later retrieval site, even as the relevant retrieval cues do not probe grammatical gender. Another pattern is also possible. D. Kush, Johns, and Van Dyke (2015) manipulated the phonological similarity between lexical items in a memory load and the subject of the following test sentence. They found a consistent pattern in which the grammatical subject (e.g. boat) was read slower following a memory load that contained rhymes (coat, vote, note) than phonologically dissimilar items (table, sink, truck). However, reading times at the retrieval site of the subject, the matrix verb, did not differ across memory load conditions, and performance on later comprehension questions, too, was not influenced by the presence of rhymes in the memory load. The evidence from this study demonstrates the possibility for encoding interference without any downstream effects for speed or accuracy during retrieval. The authors posit that retrieval interference as a result of phonological similarity may only arise when a phonological representation is necessary to distinguish semantically and syntactically similar constituents, but is otherwise orthogonal to resolving syntactic dependencies such as subject-verb dependencies.

The contrast in findings between overlap in grammatical cues and overlap in phonological similarity presents an important question as to what types of features may create encoding interference and which might then create later retrieval interference. The experiments presented in the current study examine semantic features which are then not referenced by later retrieval cues.

The role of semantic relatedness has been studied so far in agreement production. In a series of sentence completion tasks constructed to elicit agreement attraction, Barker, Nicol, and Garrett (2001) found that semantic relatedness in both semantic characteristics (e.g. boat-hood) and animacy influenced the likelihood of producing a plural verb following the sentences in (18), such that (18a) resulted in a greater number of plural responses.

(18) a. The canoe near the sailboats _

b. The canoe near the cabins _

This pattern of results can be interpreted as evidence of interference during encoding; the semantic features referenced are not ones that would be predicted to be used during retrieval. The current study expands on this, demonstrating how encoding interference driven by semantic featural overlap may affect later processing, specifically during comprehension in reading.

3.2 Design

To isolate the effect of encoding interference, it is necessary to establish three components: (i) Similarity: two DPs are similar along some dimension, (ii) Proximity: the two DPs occur in a context susceptible to encoding interference, (iii) Unselective Retrieval: a retrieval site which triggers retrieval of one of the impacted DPs without referencing the dimension of similarity established in (i).

In the current study, the featural overlap of two critical DPs was manipulated along lexical-semantic dimensions. We employed three levels of similarity between the two DPs: (i) **Low Similarity**, in which the two DPs, e.g. *knife-shirt*, are alike only in grammatical features such as number and animacy, (ii) **High Similarity**, in which the two DPs are alike along many physical, material, and/or functional dimensions, e.g. *knife-sword*, and (iii) **Medium Similarity**, in which the two DPs, e.g. *knife-stick*, are alike only along some dimensions, and only in specific contexts, in this case as sharpenable objects.

Then, we created a context in which encoding interference is likely to occur by introducing two DPs in close proximity. We did this by constructing RCs in which the target noun was always the head of the RC, and in which the intervening DP occurred in one of two RC-internal positions. To create Near Proximity conditions, the intervening DP occurred in the RC subject position. To create Far Proximity conditions, the intervening DP occurred in the direct object position. In all conditions, the intervening DP was a theme such that items were nearly synonymous across proximity conditions. For example, a Near Proximity condition would include an RC such as *the knife that the sword was placed by*, and a Far Proximity condition would include its analog, *the knife that someone placed the sword by*. To realize the grammatical subject position in the RC in the Far Proximity conditions, a local pronoun (*I*, *we*, *you*) or indefinite pronoun (*someone*) was used. ORCs with these grammatical RC subjects have shown highly reduced ORC penalties, as discussed for the example earlier in (14).

Finally, to create an unselective retrieval site, we constructed a complex verbal region, containing a sequence of auxiliary verbs (*had been*), followed by an adverb (e.g. *recently*), prior to the main verb (e.g. *sharpened*), to serve as an extended retrieval site. The key facet of the design is that the auxiliaries serve as a trigger for retrieval of the grammatical subject, without providing cues that reference the manipulated dimension of similarity. Semantically selective information is only encountered later at the main verb. Therefore, any differences in reading time prior to the main verb cannot be attributed to cue overload, but rather must reflect interference during encoding. Focusing on effects in this extended retrieval site ensures that our study focuses on the aftermath of encoding, and not on difficulties encountered during the encoding process itself.

Combining the three components of similarity, proximity, and unselective retrieval resulted in a 3x2 design manipulating Similarity (High, Medium, Low) and Proximity of the target and intervener (Near, Far). An example item is given in Table 3.1. Initially 48 sextets matching the design described above were created. Several other aspects of the stimuli were controlled for. The targets and interveners were length- and frequency-matched to avoid differences in reading or retrieval due to low-level lexical characteristics. The animacy of the targets and interveners was held constant within an item, such that the target always had the same

		Preceding context	
		It seems that the **knife** that	
		Relative clause	
Similarity	***Proximity***		
High	*Near*	the **sword** was placed near	
	Far	someone placed near the **sword**	
Medium	*Near*	the **stick** was placed near	
	Far	someone placed near the **stick**	
Low	*Near*	the **shirt** was placed near	
	Far	someone placed near the **shirt**	
		Extended retrieval zone	**Spillover**
		had been recently sharpened	in the kitchen the other day.

Table 3.1: Sample item set illustrating the Similarity x Proximity manipulation in the relative clause region and the extended retrieval site containing two auxiliaries (*had been*) and an adverb (*recently*) preceding the main verb (*sharpened*).

level of animacy as the potential interveners, and the items were blocked such that roughly half of the items contained animate targets and interveners, and half inanimate. Within the animate items, some items contained human targets and interveners, while others contained non-human animals. This, too, never varied within an item, only across items. The initial set of 48 items was paired down to 42 through two norming studies, described below.

3.3 Predictions

In looking for interference during subject-verb integration, we predicted that interference could be detected upon encountering the left edge of the verb phrase, at the auxiliary *had*. By design, no semantically selective information is available at this juncture. If heightened semantic similarity increases the chance for encoding interference, we expected to see increased reading times at the retrieval site in conditions with semantically similar interveners (High, Medium, compared to

Low). If temporal contiguity plays a critical role in creating this form of interference, we expected these differences to be larger in the Near conditions compared to the Far conditions. Once the reader reaches the main verb, which is semantically selective, we expected that patterns of interference might more closely reflect classic retrieval interference findings.

3.4 Norming

We conducted two norming studies to validate the effectiveness of our Similarity condition. In the first norming study, we asked participants to judge the similarity of the RC head noun to the RC internal distractor. In the second norming study, we asked participants to judge the plausibility of nouns as the subject of the main verb. Our primary goal was to identify experimenter-generated itemsets which did not realize the design. Critical to our predictions is a gradient effect of similarity in the extended retrieval site before selective cues are provided by the main verb. Therefore, we wanted to be able to differentiate how similar two nouns are when considered on their own, before introducing the argument structure of the verb.

3.4.1 Norming: Similarity

One key facet of the design is the relative similarity between the RC head and the intervener. Initial item construction relied on experimenter intuition, following systematic notions of shared properties, such as shared physical likeness, size, or, in the case of animates, shared occupation. To confirm these judgments, a similarity rating task was conducted.

Materials

The stimuli for the similarity rating task consisted of pairs of two words. The crucial comparison was between target-intervener pairs, but all potential pairs across target and interveners (e.g. intervener to intervener) were tested, for a total of six conditions. The pairs were taken from the previously described 48 items.

Participants

Sixty undergraduate students participated in the lab for course credit. Twelve participants were excluded from the analysis who reported that they learned English after the age of eight, and six more participants were excluded to achieve a fully-counterbalanced set of 42 participants. Participants for this norming task were restricted from participating in any other relevant norming studies or the full study.

Procedure

The stimuli were presented in a Latin Square design, such that each participant saw exactly one pair per item, and the same number of types of pairs (target-intervener, intervener-intervener, etc.). The experiment was hosted on IbexFarm (http://spellout.net/ibexfarm/) and conducted in a private laboratory room. Participants were presented with pairs of words and instructed to rate them based on similarity on a 7-point Likert scale. Participants were given minimal guidance in how to construe the concept of similarity, as we did not want to bias their understanding beyond how eventual naïve reading participants might construe similarity. Participants were however given four guided practice items, and seven additional practice items, for participants to adjust to the task before encountering the experimental stimuli.

Analysis

Results were analyzed using cumulative link mixed models using the function 'clmm' from the ordinal package (Christensen, 2015) in R (R Core Team et al., 2013). We aimed to exclude items such that, for the final data set, the target-high pair was rated the most similar, and target-low pair the least similar. We did not impose strict top-down expectations for the similarity rating of the target-mid pair, and as a result, this category had the largest variance in similarity ratings. In the majority of items, the target-mid pair rating fell between the target-high and target-low pairs, but in 3 of the final items it was rated more or as similar compared to the target-high pair, and in 10 of the final items it was rated less or as similar compared to the target-low pair. To address this, a separate analysis was conducted in all reading experiments using the similarity rating scores as a continuous measure of similarity.

Final item statistics can be found in Table 3.2.

3.4.2 Norming: Plausibility

A plausibility rating study was conducted to confirm experimenter judgments of plausibility of target and interveners as subjects of the main verb. This aspect of norming was crucial to confirm that the High and Medium similarity interveners were equally plausible subjects of the matrix verb compared to the target nouns, with the intention of creating the best case for retrieval interference at the main verb. The design criteria therefore warranted Low similarity interveners that were significantly less plausible than the target nouns or other interveners as subjects of the main verb.

Materials

Materials were taken and modified from the original set of 48 to consist of the target/intervener and the extended retrieval site, resulting in four conditions, given in (19).

(19) a. Target: The knife had been recently sharpened.

b. High: The sword had been recently sharpened.

c. Med: The stick had been recently sharpened.

d. Low: The shirt had been recently sharpened.

Participants

Forty-nine undergraduate students participated online for course credit. Nine participants were excluded from the final data set for reporting learning English after the age of eight or not being familiar with the vocabulary used. The excluded participants were replaced such that the final set of 40 participants were fully counterbalanced across conditions. Participants for this norming task were restricted from participating in any other relevant norming studies or the full study.

Procedure

The stimuli were presented such that each participant saw exactly one sentence per item, and the same number of types of sentences (Target, High intervener, etc.). The experiment was hosted on IbexFarm. Participants were presented with the sentences and asked to rate their plausibility on a 7-point Likert scale. Participants were encouraged to think of how likely or plausible the events described in the sentence were. They were provided four guided practice items and nine additional practice items before encountering the experimental stimuli.

Word	Similarity to target	Plausibility as subject
Target	–	5.10 (0.06)
High	5.86 (0.06)	4.87 (0.07)
Medium	3.59 (0.1)	4.84 (0.06)
Low	2.5 (0.08)	2.78 (0.08)

Table 3.2: Norming results for the final set of Experiment 1 items, M (SE).

Analysis and exclusion criteria

We aimed to exclude items such that, for the final data set, the target and two similar (High, Medium) interveners did not significantly differ in terms of plausibility as subjects, and such that the Low condition was significantly less plausible as the subject. Items were excluded if the z-scored difference in plausibility of the High and Medium interveners was greater than $|2|$.

3.4.3 Final itemset

Following the results of the two norming studies, six items were excluded, for a final item set of 42. The item statistics for the final item set can be seen in Table 3.2.

3.5 Experiment 1: Encoding interference in self-paced reading

The final item sets were run in a self-paced reading study to establish an online measure of retrieval difficulty. Regardless of encoding interference, we predicted to find some degree of retrieval interference once semantically selected information was available, at the main verb, such that more similar interveners incurred longer reading times. We predicted that evidence of encoding interference would emerge prior to the main verb, in the form of differences in reading time based on semantic similarity despite semantic retrieval cues not yet being available.

Experiment 1 was run as two experiments on different populations, using the same materials. The procedures and analyses will be reported together, while the participants and results will be reported separately as Experiment 1a and Experiment 1b.

3.5.1 Procedure

Stimuli were presented in word-by-word self-paced reading. Participants were instructed to press the spacebar to progress through the sentence. Each sentence was followed by a comprehension question with a Yes/No answer. There was an equal number of correct Yes and No responses across the full experiment. Questions on experimental stimuli varied with respect to which region of the sentence they probed, evenly distributed across the preceding context, the predicate within the relative clause, the adverbial region, and the final spillover region. For example, the question corresponding to the sample item in Table 1 targeted the preceding context (8). No questions probed whether participants had maintained the correct understanding of the subject of the main verb (e.g. *Was it the knife that was sharpened?*).

(20) NEAR: It was clear that the knife that the {HIGH: sword | MEDIUM: stick | LOW: shirt} was placed by had been recently sharpened in the kitchen yesterday.

FAR: It was clear that the knife that someone placed the {HIGH: sword | MEDIUM: stick | LOW: shirt} by had been recently sharpened in the kitchen yesterday.

Q: Was it obvious this event happened? (YES)

In Experiment 1a, participants accessed the experiment, hosted on IbexFarm, through Prolific, an online experiment-hosting platform (https://app.prolific.co/). They were instructed to complete the experiment on a desktop at a time when they could complete the study in one sitting, and when there was minimal distraction in their surroundings. The study was estimated to take 30 minutes to complete. In Experiment 1b, participants completed the experiment on a desktop in a laboratory setting. Participants were told that they could take breaks as needed throughout the experiment but were instructed not to take a break during the middle of a sentence.

3.5.2 Data processing and analysis

Data processing and analysis were consistent across Experiments 1a and 1b. Outlier exclusion was conducted via minimum and maximum hard cut-offs for RTs at 100 ms and 2000 ms respectively, such that any data points above 2000 ms or below 100 ms were excluded (Exp1a = .7% data loss, Exp1b = 0.8% data loss). The final analyses were conducted using the set of correct trial responses. Log-transformed reading times were analyzed for each region using a linear mixed effects regression model with random intercepts for subject and item (lme4 package; Bates et al., 2015). A maximal model including random slopes and intercepts was first attempted, but after not consistently converging across regions, the most maximal model that would converge was used for each region. p values were estimated using the package lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017). Question accuracy was analyzed using a generalized linear mixed effects model (lme4 package; Bates et al., 2015). Sum contrast coding was used for the two-level factor of Proximity, and Helmert coding was used for the three-level factor of Similarity. The Helmert coding contrasted the average of High & Medium with

Low similarity, to investigate effects of any similarity compared to the baseline Low similarity condition, and then contrasted High to Medium similarity. Analyses were also conducted using scaled lab-collected similarity ratings as a continuous predictor in the models, to address any variation in the relative similarity differences across items. To distinguish between the two analyses, the former will be referred to as the experimenter category analysis, and the latter the lab-collected similarity ratings analysis. Results for Experiments 1a and 1b are visualized in Figure 3.1.

3.5.3 Experiment 1a

Participants

Sixty-one participants were recruited via Prolific (https://app.prolific.co/), with the following conditions: nationality was restricted to the United States, and participants must have indicated that they had not been diagnosed with Dyslexia, Dyspraxia or ADHD, or experience any other literacy difficulties. They had to report that their first language was English, and that they had completed (at least) secondary education. Participants were compensated at a rate of $11/hr. Participants who scored below 75% accuracy on filler comprehension questions were excluded from the final dataset, leaving a total of 57 participants.

Question accuracy

Question accuracy was examined for the final remaining dataset (Table 3.3). There was a main effect of Proximity, such that overall the Near condition yielded higher accuracy than the Far condition, $z = 6.05$, $p < .001$. There was a main effect of High-Medium similarity compared to Low similarity such that Low similarity yielded higher accuracy than the averaged higher similarity conditions,

Similarity	Proximity	
	Near	*Far*
High	0.81 (0.004)	0.78 (0.004)
Medium	0.77 (0.005)	0.77 (0.004)
Low	0.82 (0.004)	0.81 (0.004)

Table 3.3: Question accuracy by condition for Experiment 1a, M (SE).

z = -5.08, $p < .001$. There was a significant difference between the High and Medium similarity conditions themselves, with High similarity conditions resulting in higher comprehension question accuracy than the Medium conditions, z = 5.33, $p < .001$. This effect was qualified by an interaction, in which the High similarity conditions were more accurate than the Medium conditions more so in the Near Proximity than Far, $z = 2.09$, $p < .05$.

For the final analysis, incorrect trials were excluded.

Extended retrieval site prior to semantically selective information

No significant effects were found across the auxiliary regions *had been*. At the adverbial region, there were marginally longer reading times for the Far condition (M = 349 ms, SE = 6 ms) compared to the Near condition (M = 342 ms, SE = 5 ms), t = -1.72, $p = .08$. No other effects across these regions reached significance, but numerical trends in the Near conditions at the adverbial region suggested a penalty for the High similarity (M = 351 ms, SE = 10 ms) compared to both the Medium similarity (M = 336 ms, SE = 8 ms) and Low similarity (M = 338 ms, SE = 9 ms). However, this difference did not reach significance. The analysis using lab-collected similarity ratings did not yield qualitatively different results, again finding a marginal effect for Proximity at the adverbial region, t = -1.71, p = .09.

Extended retrieval site following semantically selective information

At the lexical verb, a slight numerical trend in the Near conditions suggested longer reading times for High similarity (M = 368 ms, SE = 11 ms) and Medium similarity (M = 370 ms, SE = 11 ms) compared to Low similarity (M = 357 ms, SE = 11 ms). However, in both the analysis using experimenter-determined categories of similarity and the analysis using lab-collected similarity ratings as a continuous measure of similarity, no effects or interactions reached significance at the lexical verb or in the spillover region following it.

Experiment 1a interim discussion

Experiment 1a found that the Medium similarity conditions yielded the lowest question accuracy, compared both to the High and Low similarity conditions. Such a pattern could suggest that Medium similarity conditions resulted in greater interference, leading to degraded comprehension question accuracy, despite the fact that most questions did not probe the identity of the target or intervener. The Medium penalty did not correspond to patterns in the reading data, however. Under one understanding of semantic similarity-based interference, an increase in similarity should monotonically increase the degree or likelihood of interference. The reading data demonstrated numerical trends in which High similarity was read slower prior to semantically selective information, which would be predicted if high degrees of featural overlap resulted in interference even if those featural properties could not have been addressed via retrieval cue. However, this difference was not statistically significant. A marginal difference did emerge for Proximity at the extended retrieval site prior to semantically selective information, with the Far condition resulting in longer reading times. This would be unpredicted following theories of interference in which close temporal proximity (represented here by the Near condition) increases the likelihood of interference during encoding. This finding will be explored more in Experiment 1b and Experiment 2.

3.5.4 Experiment 1b

Participants

Sixty-four undergraduate students were recruited via the UCSC Linguistics SONA Subject Pool and were granted one course credit for their participation. Participants who indicated that English was not their native language, and who didn't learn English in an immersive setting before the age of eight, were excluded from the final data set, as were participants who disclosed a reading or attention disorder. Finally, participants who scored below 75% on filler comprehension questions were excluded from the final dataset, leaving a total of 54 participants.

Question accuracy

Question accuracy was examined for the final remaining dataset (Table 3.4). Contrary to Experiment 1a, there was a main effect of High-Medium similarity compared to Low similarity such that Low similarity yielded lower accuracy than the averaged higher similarity conditions, $z = 13.17$, $p < .001$. The relationship between the High and Medium conditions themselves was in line with Experiment 1a: there was a significant difference between the High and Medium similarity conditions, with High similarity conditions resulting in higher comprehension question accuracy than the Medium conditions, $z = 5.18$, $p < .001$. Again, this effect was qualified by an interaction, in which the High similarity conditions were more accurate than the Medium conditions more so in the Near Proximity than Far, $z =$ -3.22, $p < .01$.

Similarity	**Proximity**	
	Near	*Far*
High	0.79 (0.004)	0.76 (0.005)
Medium	0.77 (0.005)	0.77 (0.005)
Low	0.74 (0.005)	0.73 (0.005)

Table 3.4: Question accuracy by condition for Experiment 1b, M (SE).

Extended retrieval site prior to semantically selective information

A main effect such that High similarity conditions were read faster than Medium similarity conditions persisted across the auxiliary regions had (High M = 413 ms, SE = 9; Medium M = 465 ms, SE = 11 ms), t = -3.55, $p < .001$, and been (High M = 396 ms, SE = 8 ms; Medium M = 420 ms, SE = 9 ms), t = -2.25, $p < .05$, as well as in the adverbial region (High M = 385 ms, SE = 7 ms; Medium M = 406, SE = 8 ms), t = -2.01, $p < .05$. Using lab-collected similarity ratings as a continuous measure of similarity, an effect of similarity emerged only at the auxiliary *had*, with higher similarity resulting in longer reading times, t = -2.27, $p < .05$. No effects of Proximity or interactions reached significance.

Extended retrieval site following semantically selective information

No effects reached significance at the main verb. In the spillover region directly following the main verb, an interaction emerged such that the difference between High similarity and Medium similarity conditions was greater in the Near condition (difference: 24 ms) than in the Far condition (difference: 7 ms), t = -2.36, $p < .05$. There was separately a main effect of Proximity in this region, in which the Near condition was overall read slower (M = 397 ms, SE = 6 ms) than the Far condition (M = 382 ms, SE = 5 ms), t = 2.02, $p < .05$. Using lab-collected similarity ratings as a continuous measure, the effect of Proximity remained, t

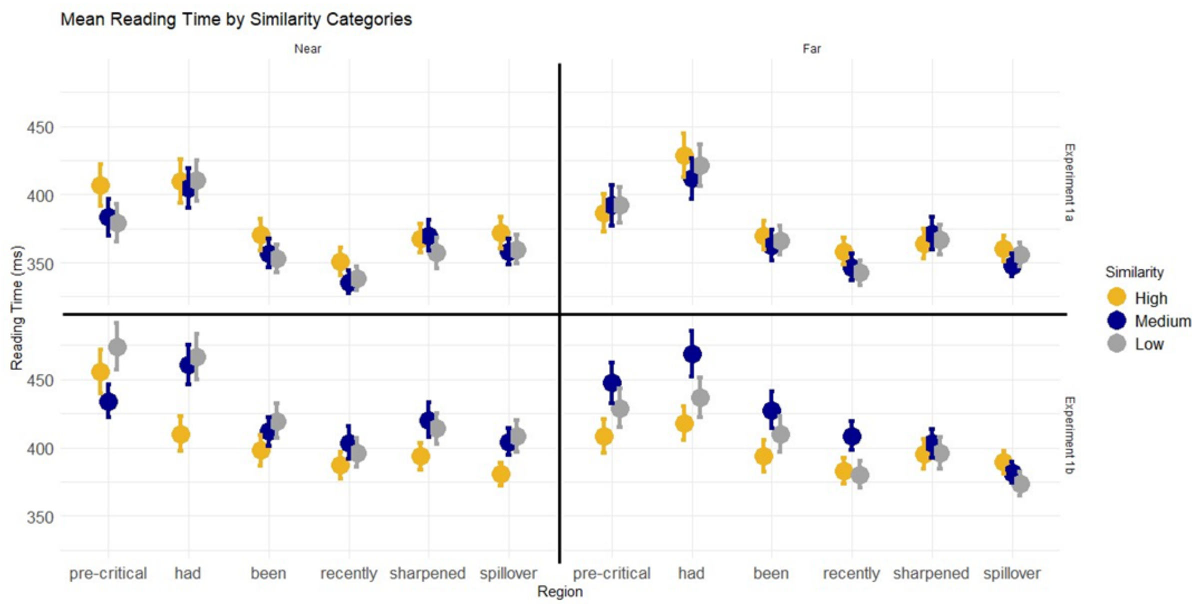

Figure 3.1: Mean reading times across Experiments 1a and 1b, starting from the pre-critical region (the last region within the RC), up until the spillover region (the first region following the main verb). Near conditions are represented on the left panel, Far conditions on the right. Experiment 1a mean reading times are displayed in the top panels, Experiment 1b on the bottom.

= 1.98, $p < .05$, and there was an interaction between Proximity and Similarity Rating, t = -2.17, $p < .05$, such that distractors rated more similar to the target were read faster in the Near condition, but slower in the Far condition.

Experiment 1b interim discussion

Experiment 1b replicated the general finding in question accuracy data from Experiment 1a, lower accuracy for Medium similarity, compared to High similarity. In this iteration of the study, this general pattern extended to the reading time data, with a penalty for Medium similarity persisting across the extended retrieval site prior to semantically selective information. This provides strong evidence that differential semantic similarity between a target and intervener can disrupt retrieval, even in cases where the retrieval process is not hypothesized to make reference to features relevant to the semantic features shared by the two nouns. The exact pattern of interference suggests that interference is greatest not in cases of maximal feature matching, but rather in cases where features partially match.

An interaction with Proximity doesn't emerge until semantically selective information has been encountered, in the spillover region following the verb. The difference between High and Medium similarity is exacerbated in the Near condition, compared to the Far condition, suggesting that interference is greatest when the target and intervener are encountered first in close temporal proximity or share more overlap in their encoded contextual states.

3.5.5 Experiment 1 discussion

This experiment was designed to test for semantic similarity-based interference at a verbal retrieval site during reading. The retrieval site was purposefully designed in such a way that at the initial point at which retrieval is triggered, no semantically selective information is available (*had been*), and therefore no semantic retrieval cues should be used. Reading patterns that diverge prior to semantically selective information on the basis of semantic similarity therefore would suggest that the interference comes not from cue overload, but from interference on the basis of encoding. Two potential patterns of interference are predicted on the basis of theories of similarity-based interference: greater interference as similarity between the target and the intervener increases (High similarity penalty), greatest interference when the associations between relevant features mismatch across the target and intervener (Medium similarity penalty). Across both iterations of Experiment 1, evidence for the former only emerges in numerical trends, where in Experiment 1a High similarity interveners resulted in numerically longer reading times prior to semantically selective information, but this difference was not significantly different. The most robust pattern of semantic effects was one in which Medium similarity interveners resulted in longer reading times across the extended retrieval site, compared to High similarity interveners, seen in Experiment 1b. This provides evidence in support of an understanding of interference in which interference emerges in the case of partially matching features between the target and intervener, potentially explainable if features are bound to one another, and conflicting bindings across different encodings results in downstream interference (Logačev & Vasishth, 2011). It is unclear why different patterns should emerge across different populations and settings; it may be that interference on the basis of highly similar interveners is a harder effect to detect, with smaller effect sizes, and potentially requiring higher power to reliably detect.

The interpretation of the origins of interference in this design rests on the localization of an effect of similarity. The self-paced reading methodology allows for certainty about when semantically selective information is available, but also allows for uncertainty in which effects may spillover from previous regions and disallows for potentially informative processes seen in natural reading such as rereading. Experiment 2 was therefore conducted to more closely examine timecourse predictions within this design.

3.6 Experiment 2: Encoding interference in eye-tracking while reading

To more closely examine the timecourse of interference effects, the design used in Experiment 1 was implemented in an eye-tracking while reading study. The interpretation of effects thus far has relied on the assumption that retrieval should be triggered at the *had been* region, but because no semantically selective information should be available at that point, no semantically-influenced retrieval cues should be referenced. Any semantic interference prior to the main verb would implicate semantic similarity-based interference not on the basis of cue overload during retrieval, but rather interference earlier in the process: during encoding. Self-paced reading studies are not necessarily ideal for a design with such tightly time-locked predictions, as the method often results in effects showing up not on the region of interest itself but variably across following regions (Forster, Guerrera, & Elliot, 2009) and the response times themselves include not just reading time, but button press planning and execution as well.

The eye-tracking replication aims to provide a more fine-grained reading profile, specifically allowing for the examination of regressions back in the text and early fixations at the critical *had been* region.

3.6.1 Methods

Materials

The materials were identical to those used in Experiment 1, but are repeated here in (21) for convenience and demonstration of analysis regions.

(21) It seems that $|_{\text{RC-head}}$ the knife | ...

a. Near-High: | that the sword was placed near...

b. Far-High: | that someone placed the sword near...

c. Near-Medium: | that the stick was placed near...

d. Far-Medium: | that someone placed the stick near...

e. Near-Low: | that the shirt was placed near...

f. Far-Low: | that someone placed the shirt near...

... $|_{\text{had-been}}$ had been $|_{\text{adv}}$ recently $|_{\text{verb}}$ sharpened $|_{\text{spillover}}$ in the kitchen $|_{\text{final}}$ the other day.

For the purpose of analysis, the auxiliary region *had been* is treated as one region, as demonstrated in (21), followed by the adverbial region, the main verb, and spillover regions. In self-paced reading, the point at which semantically selective information becomes available is unambiguously at the point of the main verb. Eye tracking with natural reading complicates this slightly. The moving window self-paced reading display doesn't allow for parafoveal preview, in which readers may extract information from word n+1 while fixating on word n, up to 14 characters to the right for English readers (Rayner, Pollatsek, Ashby, & Clifton Jr, 2012). It's quite possible that, depending on the specific location of a fixation in the adverb region, readers in the eye-tracking experiment may have had some access to semantic information about the upcoming verb in the adverb region. The *had been* region, therefore, is the only region in which definitively no semantically selective information was available. Despite the possibility of available semantically selective information at the adverb region via parafoveal preview, results in the adverbial region will still be reported alongside the *had been* region.

Participants

Sixty participants were recruited through the department Subject Pool and received course credit for their participation. Participants were excluded if they had more than 10 track losses within the critical region (*had been*) across all experimental items, leaving 50 participants remaining. Participants had normal or corrected-to-normal vision and had learned English before the age of 8. Data from the first 26 participants were collected prior to the onset of the COVID-19 pandemic in March 2020. The remaining participants were run following the re-opening of the university in the Fall of 2021 while wearing black surgical masks.

Procedure

Participants read sentences on the screen while their head was stabilized using a height-adjustable chinrest. Eye movement data was collected using a table-mounted SR EyeLink 1000 eye-tracker with the chin rest positioned such that the participant was roughly 60 cm away from the screen, with roughly 2 characters subtending 1 degree of visual angle. Data from the right eye were recorded at a sampling rate of 1000Hz. Prior to the experiment and following any breaks, a 3-point calibration was conducted. The items were presented using the program Eye-Track from the Eye Lab at UMass Amherst (blogs.umass.edu/eyelab/software/) in 11-pt black mono-spaced font on a gray background. The items were counter-balanced and randomized as in the previous experiment such that each participant saw one and only one form of each item. The experimental stimuli were preceded by 12 practice stimuli, and randomized among 58 filler items, which consisted of both fillers intended to obscure the present manipulation as well as experimental stimuli from unrelated experiments.

Participants were instructed to restrict movement during the experiment and rest their eyes between trials when possible. They were also instructed to remove any eye make-up prior to the experiment. Following the return to campus during the COVID-19 pandemic, participants were instructed to wear a mask throughout the duration of the experiment.

Data processing

Blinks and track losses were cleaned from the data using software developed by the UMass Eye Lab (https://blogs.umass.edu/eyelab/software/). All single fixations below 80 ms or above 1200 ms are excluded, first pass times over 2000 ms are excluded, and total times over 4000 ms are excluded. Participants were excluded on the basis of track loss during the critical *had been* region, resulting in 10 excluded participants, who exceeded 10 track losses in the critical region across all trials. The following reading measures were analyzed: *first fixation* duration, the duration of the first fixation on a region; *first pass time*, the sum of fixations in a region before exiting the region to the left or right; *go past time*, the sum of all fixations beginning with the first fixation in a region until the region is exited to the right; *second pass time*, the sum of re-reading fixations on a region; *total times*, the sum of all fixations on a region; *regressions in*, the probability of a launching a regression into a region; and *regressions out*, the probability of launching a regressive saccade from a region. Following initial data processing steps, the data were trimmed to exclude outliers that lay beyond three standard deviations of the mean for each analyzed region and measure.

Reading measures and analysis

The analysis of reading measures will be divided between reading measures that reflect initial stages of reading (first fixation, first pass, go past) and measures that reflect re-reading (second pass, total times, and probability of regressions in and out). Only theoretically-motivated region-measure pairs have been analyzed. For measures that reflect initial reading, analysis is restricted to the *had been* region, the adverb, the main verb, and the spillover region. Analysis of regressions out will also be restricted to those regions. Eye tracking methodology not only allows for more fine-grained measures during initial processing, but also for the measurement of naturalistic reading. While we don't predict any differences on the retrieval target, the RC head (*knife*), when first encountered, as all conditions are the same through this point, we did predict that increased interference when establishing subject-verb agreement might in turn increase the probability that participants would regress back into the RC head region and spend more time rereading. Therefore, for second pass and total times, the verbal regions are analyzed in addition to the RC head, and for the probability of regressions into a region, the RC head is analyzed.

As with Experiment 1, analysis was conducted on the basis of experimenter-defined categories of similarity (High, Medium, Low), as well as with a continuous measure of similarity, lab-collected similarity ratings. Only significant and marginally significant findings are presented by region, separated by whether the region precedes or follows semantically selective information.

3.6.2 Initial reading time results

First, the results from reading measures which reflect initial reading (first fixation durations, first pass reading times, and go past times) will be reported. The means for initial reading measures can be found in Table 3.5.

Prox.	Similarity	RC Head	Auxiliary	Adverb	Main Verb	Spillover
			First fixation durations			
Near	*High*	–	261 (5)	249 (5)	265 (5)	256 (5)
	Medium	–	261 (6)	250 (5)	268 (6)	254 (5)
	Low	–	260 (6)	240 (5)	270 (6)	251 (5)
Far	*High*	–	259 (5)	249 (5)	262 (6)	255 (5)
	Medium	–	266 (6)	235 (5)	270 (6)	250 (5)
	Low	–	248 (5)	242 (5)	274 (6)	255 (5)
			First pass times			
Near	*High*	–	341 (11)	288 (8)	290 (7)	439 (13)
	Medium	–	341 (11)	275 (8)	296 (8)	444 (15)
	Low	–	335 (11)	278 (8)	301 (8)	433 (14)
Far	*High*	–	326 (10)	277 (7)	292 (7)	446 (14)
	Medium	–	347 (11)	262 (7)	298 (7)	431 (14)
	Low	–	325 (10)	275 (8)	298 (8)	441 (15)
			Go past times			
Near	*High*	–	540 (28)	377 (18)	376 (16)	709 (37)
	Medium	–	566 (33)	399 (24)	399 (19)	750 (43)
	Low	–	526 (29)	364 (21)	383 (15)	7001 (37)
Far	*High*	–	511 (31)	366 (18)	372 (17)	622 (29)
	Medium	–	565 (32)	384 (22)	381 (17)	673 (35)
	Low	–	647 (39)	413 (25)	400 (18)	701 (41)

Table 3.5: Means and standard errors (M (SE)) in milliseconds of analyzed eye-tracking regions and measures that are indicative of initial reading.

Extended retrieval site prior to semantically selective information

At the *had been* region, there was a marginal interaction in first fixation durations between Similarity and Proximity such that in the Far conditions, the averaged High and Medium similarity conditions (M = 263 ms, SE = 6 ms) yielded longer first fixation durations than the Low similarity condition (M = 248 ms, SE = 5 ms), but no difference in the Near conditions, t = -1.735, p = .08 (Figure 2). At the adverb region, there was a marginal main effect in first pass times in which the High similarity condition (M = 283 ms, SE = 5 ms) was read slower than the Medium similarity condition (M = 269 ms, SE = 5 ms), t = 1.801, p = .07.

Extended retrieval site following semantically selective information

In the similarity ratings analysis, there was a main effect of Similarity at the main verb, such that higher similarity interveners resulted in faster go past times, t = -2.541, $p < .05$.

3.6.3 Rereading results

Next, measures that reflect rereading will be reported. This includes regression probabilities, as well as second pass and total reading times. The means for rereading measures can be found in Table 3.6.

RC head

In rereading measures on the RC head, there was a robust penalty for the Near position compared to the Far position. This emerged as longer second pass times for Near interveners (M = 500 ms, SE = 16 ms) compared to Far interveners (M = 407 ms, SE =14 ms), in both the experimenter category analysis, t = 5.242, $p < .001$, and the similarity ratings analysis, t = 5.194, $p < .001$. There were also

Prox.	Similarity	RC Head	Auxiliary	Adverb	Main Verb	Spillover
			Regressions out			
Near	*High*	–	25% (2)	13% (2)	13% (2)	22% (2)
	Medium	–	23% (3)	18% (2)	14% (2)	24% (3)
	Low	–	23% (3)	13% (2)	16% (2)	22% (2)
Far	*High*	–	22% (2)	14% (2)	15% (2)	18% (2)
	Medium	–	26% (3)	20% (3)	17% (2)	26% (3)
	Low	–	33% (3)	19% (3)	17% (2)	22% (3)
			Regressions in			
Near	*High*	60% (3)	–	–	–	–
	Medium	60% (3)	–	–	–	–
	Low	54% (3)	–	–	–	–
Far	*High*	48% (3)	–	–	–	–
	Medium	50% (3)	–	–	–	–
	Low	50% (3)	–	–	–	–
			Second pass times			
Near	*High*	512 (27)	243 (18)	166 (13)	160 (12)	257 (18)
	Medium	521 (29)	229 (18)	143 (13)	172 (14)	247 (18)
	Low	464 (28)	213 (17)	164 (13)	167 (14)	236 (19)
Far	*High*	400 (24)	199 (15)	166 (12)	156 (11)	257 (18)
	Medium	418 (26)	291 (20)	173 (12)	191 (13)	293 (19)
	Low	402 (25)	239 (19)	166 (13)	158 (13)	252 (20)
			Regression-contingent second pass times			
Near	*High*	235 (19)	194 (16)	156 (12)	–	–
	Medium	216 (19)	164 (15)	130 (12)	–	–
	Low	185 (18)	164 (15)	154 (13)	–	–
Far	*High*	193 (16)	155 (13)	154 (12)	–	–
	Medium	191 (17)	207 (17)	169 (12)	–	–
	Low	180 (15)	173 (16)	154 (13)	–	–
			Total times			
Near	*High*	914 (32)	665 (24)	479 (17)	480 (15)	773 (25)
	Medium	903 (33)	658 (26)	450 (17)	496 (18)	806 (27)
	Low	843 (33)	638 (24)	445 (16)	491 (19)	753 (25)
Far	*High*	812 (29)	577 (21)	475 (16)	452 (15)	782 (25)
	Medium	831 (33)	713 (26)	449 (15)	501 (17)	789 (23)
	Low	780 (30)	670 (26)	471 (19)	496 (18)	778 (27)

Table 3.6: Means and standard errors (M(SE)) of analyzed eye-tracking regions and measures that reflect rereading, represented in milliseconds or probability of regression.

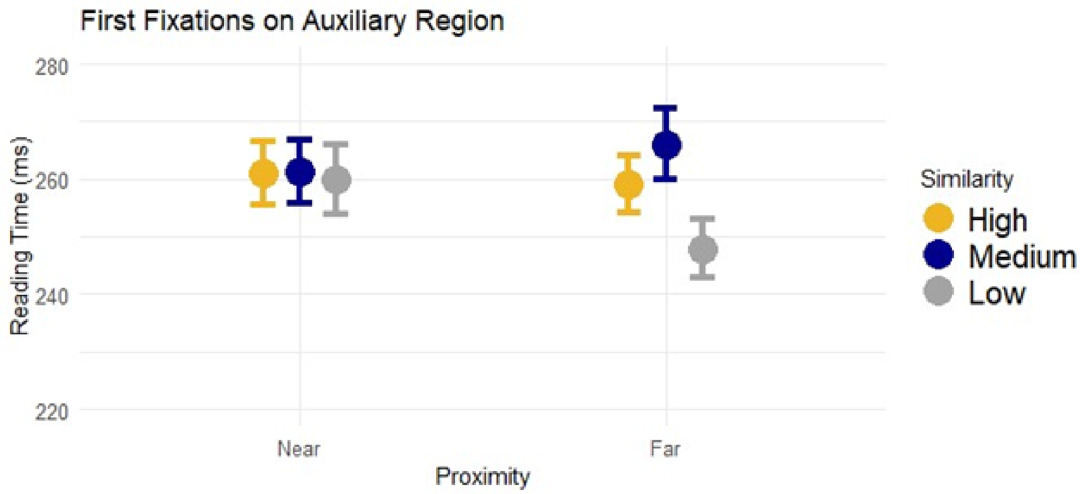

Figure 3.2: First fixation durations on the auxiliary had been region, demonstrating a marginal interaction between Similarity and Proximity.

significantly longer total times for Near interveners (M = 888 ms, SE = 19 ms) compared to Far (M = 808 ms, SE = 18 ms) in both the experimenter category analysis, $t = 3.973$, $p < .001$, and similarity rating analysis, $t = 3.954$, $p < .001$. These differences in reading time coincided with a significantly higher probability of regressions into the region in the Near condition (M = 58%, SE = 1.6%) than the Far condition (M = 49%, SE = 1.6%) in both the experimenter category analysis, $t = 4.138$, $p < .001$, and similarity rating analysis, $t = 4.069$, $p < .001$.

Extended retrieval site prior to semantically selective information

In the experimenter category analysis for rereading times at the *had been* region, there was an advantage for High similarity interveners (M = 221 ms, SE = 12 ms) in comparison specifically to the Medium similarity interveners (M = 226 ms, SE = 13 ms), $t = -2.269$, $p < .05$, which was driven by a significant interaction such that this difference was more pronounced in the Far condition (difference: 172 ms) than the Near condition (difference: 14 ms), $t = 3.177$, $p < .01$ (Figure 3.3). In the corresponding analysis of rereading measures using lab-collected similarity ratings, higher similarity interveners resulted in faster rereading times, especially in the Far condition, as a main effect of higher similarity interveners being read faster than lower similarity interveners marginally for second pass times, t = -1.675, $p = .09$, and significantly for total times, t = -2.651, $p < .01$. These main effects were qualified by interactions in which the high similarity advantage was more pronounced in the Far condition, in both the second pass, $t = 2.015$, $p < .05$, and total times, $t = 2.824$, $p < .01$. This advantage for high similarity interveners was also evidenced by a decreased probability to regress out of the region, with a marginal main effect of higher similarity, t = -1.821, $p = .07$.

At the adverb region, there was a marginal main effect of longer total times for the averaged High and Medium similarity interveners (M = 477 ms, SE = 12 ms) over Low similarity interveners (M = 449 ms, SE = 11 ms), $t = 1.751$, $p = .08$. This was paired with increased regressions out of the adverbial region for Medium similarity interveners (M = 19%, SE = 1.8%) compared to High similarity interveners (M = 14%, SE = 1.5%), t = -2.172, $p < .05$. There was also a marginal main effect of Proximity, such that there were fewer regressions out of the adverbial region in the Near condition (M = 15%, SE = 1.3%), compared to the Far condition (M = 17%, SE = 1.4%), t = -1.769, p = .08.

Extended retrieval site following semantically selective information

In second pass times at the main verb, Medium similarity interveners (M = 182 ms, SE = 9 ms) resulted in longer reading times than High similarity interveners (M = 158 ms, SE = 8 ms), t = -2.02, $p < .05$ (Figure 4). The same distinction between the two higher similarity categories emerged in total times, with Medium similarity (M = 498 ms, SE = 12 ms) having overall longer total reading times than High similarity (M = 466 ms, SE = 11 ms), t = -2.471, $p < .05$. In the

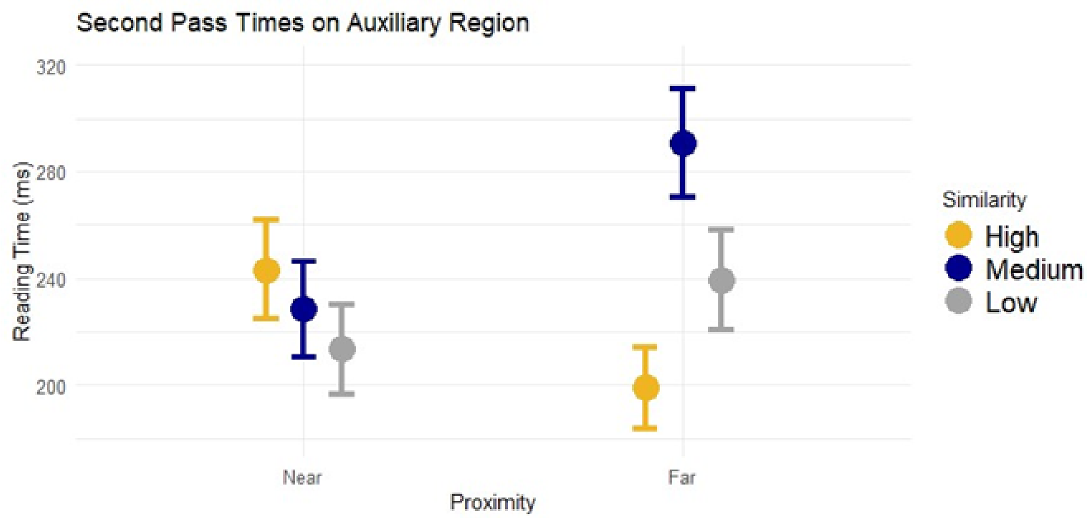

Figure 3.3: Second pass reading times on the auxiliary had been region, demonstrating an interaction between Similarity and Proximity.

similarity ratings analysis, higher similarity interveners were read faster in total times, t = -3.173, $p < .01$, with a marginal interaction in which higher similarity was read faster specifically in the Far condition, t = 1.814, p = .06.

Rereading contingent on reaching semantically selective information

The rereading times reported above indicate later reading measures, but do not distinguish between whether semantically selective information had been reached. A secondary analysis was conducted on second pass times in which the reader had already reached the main verb (sharpened). In rereading times on the RC head, the penalty for Near (M = 213 ms, SE = 11 ms) over Far (M = 188 ms, SE = 9 ms) persisted, t = 2.24, $p < .05$. In the auxiliary had been region, a crossover interaction emerged, t = 2.7, $p < .01$, such that the High conditions were read slower in the Near condition (M = 194 ms, SE = 16 ms) than the Far condition (M = 155 ms, SE = 13 ms), while the Medium conditions were read faster in the Near condition (M = 164 ms, SE = 15 ms) than the Far condition (M = 207 ms,

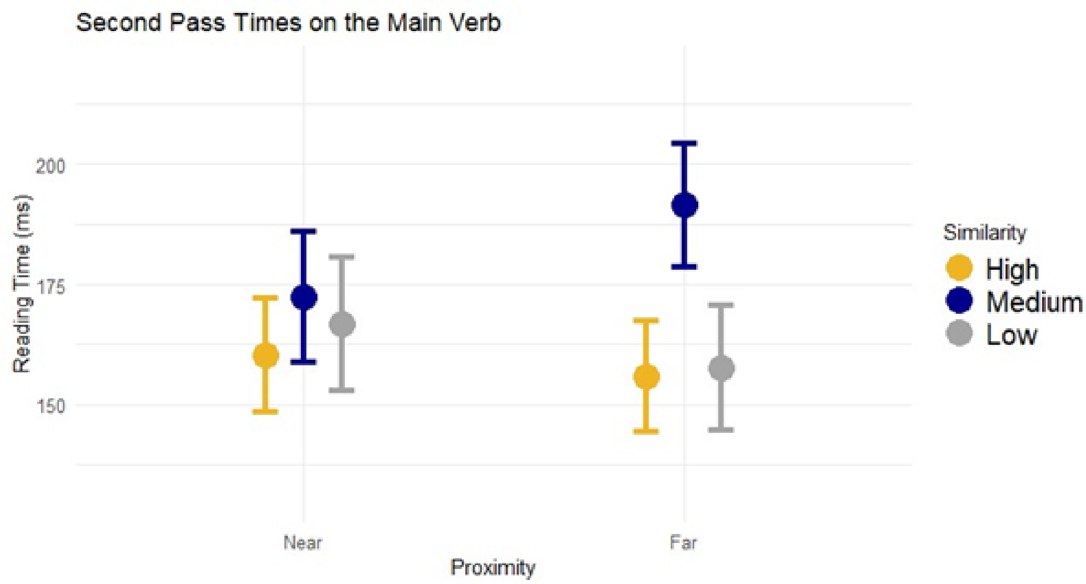

Figure 3.4: Second pass reading times on the main verb region, demonstrating a penalty for Medium similarity over High similarity.

SE = 17 ms). Finally, there was a marginal interaction, $t = 1.77$, $p = .07$, at the adverb region in which Medium conditions were again read faster in the Near condition (difference: 39 ms), with little difference for High conditions (difference: 2 ms).

3.6.4 Discussion

The eye-tracking while reading study examined the timecourse of similarity-based interference effects, before and after semantically selective information is available. We analyzed initial reading measures (first fixation, first pass, go past times) prior to semantically selective information, in which differential effects of semantic similarity cannot be attributed to cue overload during retrieval, and various periods in which semantically selective information was available: initial reading measures at or following the main verb, and rereading measures (second pass times, total times, regressions in/out) throughout the critical regions.

Several important generalizations emerged from the eye movement record. Firstly, we found evidence that differential reading based on level of semantic similarity began early in the eye movement record, marginally at the auxiliary verb region *had been*, and significantly at the adverb. These results demonstrate interference that cannot stem from cue overload, as the retrieval cues available at had been should not differentiate between the semantic properties of the target and intervener. We therefore interpret these findings as downstream effects of encoding interference.

Secondly, a crucial component of encoding theories reviewed in this paper is contiguity during initial presentation. This is represented in the eye movement record by increased rereading times and increased regressions into the RC head in the Near (or more contiguous) condition than the Far condition. However, this is complicated by the marginal interaction in first fixation durations at the had been region, in which the penalty for highly similar interveners emerges in the Far condition as opposed to the Near condition. This will be addressed further in the overall discussion for this chapter.

Finally, there was a sustained penalty for the Medium similarity condition compared to the High similarity condition in the reading and rereading of the main verb, in line with the results from Experiment 1, where semantically selective information is available. There are several interpretations compatible with this result. One interpretation is that for the interveners most likely to cause similarity-based interference (High similarity condition), any difficulty is resolved earlier in processing. In that case, the Medium similarity penalty reflects a penalty for cases where semantic similarity was not apparent during initial encoding, or during initial retrieval at *had been*. This interpretation is highly compatible with the initial design of the stimuli, in which the Medium similarity interveners were intended to be similar to the targets only along some dimensions, or only in a particular context. Difficulty for Medium similarity interveners at the main verb, where semantically selective information is available, can therefore be interpreted as cue-based retrieval interference: interference in the presence of cue overload. These results are compatible with theories of encoding in which features are associated not just to the constituent they belong to, but to one another as well, as is the case for conflicting bindings interference (Logačev & Vasishth, 2011), a possibility that will be returned to in the following discussion section.

Overall, the impacts of these findings are twofold: they demonstrate (i) evidence for semantic effects in the absence of cue overload, and (ii) differential downstream effects for interveners predicted to provoke different levels of encoding interference.

3.7 Discussion

Chapter 3 began with the goal of isolating the effects of semantically-based encoding interference from retrieval interference during sentence processing. This was operationalized across a set of experiments by testing the effect of semantic similarity at a retrieval site which did not initially provide any selective semantic information (the auxiliary region *had been*). A secondary goal of this chapter was to identify the role of contiguity in establishing an encoding context in sentence processing, extending findings from the list memory literature. The prediction was that similarity-based encoding interference, unlike retrieval interference, should be dependent on close temporal proximity during initial encoding.

In a series of reading studies, the chapter found converging evidence for the presence of semantic similarity-based interference in the absence of cue overload. We observed critical main effects of Similarity or interactions between Similarity and Proximity prior to the main verb region, a point at which semantic properties could not have been recruited as retrieval cues on the basis of bottom-up input. These findings demonstrate interference during retrieval that originates from an earlier process, which we have argued is encoding. Interactions with or main effects of Proximity provide preliminary evidence for the role of contiguity in encoding during sentence processing, building on theories that suggest a key aspect of encoding interference is an encoding context that includes co-activation in memory.

3.7.1 Similarity

Several notions of similarity were considered for the purpose of this study. The experimenter-defined categories referenced (High, Medium, Low) contrasted high featural overlap across a wide range of characteristics (High) to similarity along limited dimensions or only in particular contexts (Medium), with a baseline condition that was was similar to the target only in grammatical properties (Low). Lab-collected similarity ratings allowed for analyses using continuous measures of similarity, in which participants were not constrained by how they conceptualized similarity. Within each item, grammatical features, such as number and animacy, were held constant.

Previous literature on similarity-based interference suggests that higher degrees of semantic featural overlap should correspond to higher degrees of interference, seen in reading studies as longer reading times at the corresponding retrieval site. Evidence for this pattern appeared either numerically or marginally, suggesting that more power may have been needed to detect this effect. The more robust pattern we observed was one in which Medium similarity interveners incurred the greatest cost, which emerges both prior to and after the availability of semantically selective information across both Experiments 1 and 2. When observed after semantically selective information becomes available, this finding is compatible with cue overload on the basis of contextually-appropriate features determined by the semantic properties of the main verb. This understanding of the pattern suggests that such features are in fact available to be used as retrieval cues. However, given that we also observe this effect prior to the main verb, it seems that at least part of the contributing factor to difficulty for Medium similarity interveners may stem from interference driven by encoding. In particular, this pattern is compatible with two sources of interference.

Under the superposition account (Oberauer et al., 2012), two items with highly overlapping features may in fact result in facilitatory interference, with the feature bindings strengthening one another. In turn, this could then result in a pattern that looked like a penalty for Medium interveners.

Another account that supports a penalty specifically for Medium similarity interveners, compared to High similarity interveners, is the conflicting bindings account (Logačev & Vasishth, 2011), in which interference can occur not just in feature-item bindings but also in feature-feature bindings, resulting in greater levels of interference for partial matches than maximally similar items. This has been explored before in terms of grammatical features, where it could be possible to isolate very clearly the difference between, for example, number and animacy features (e.g. $\text{Item1}_{SG,-anim}$, $\text{Item2}_{PL,-anim}$ as partial matches). Scaling up to include featural representations of conceptual properties, including physical characteristics or contextual uses, it becomes difficult to say what 'maximally' similar items would be, beyond two synonymous lexical items. However, building on the insight that conflict comes from instances of mutually incompatible feature pairings, we can examine how this might play out using the example stimulus of the experiments in this chapter. For the purposes of this example, we might assume that each relevant property is represented in fact by a number of individual features that, when activated together, give rise to a higher-level property such as a type of material or ability to be used in a certain way. In the case of the example stimulus (knife, sword, stick, shirt), the High similarity intervener may share with the target word all of the features that give rise to the understanding of a physical material (metal), whereas the Medium similarity item would features representing a mutually incompatible material. In the encoding of each word, according to the conflicting bindings account, the features indicating the type of material (metal or wood in this case) would also each be individually bound to features that represent an item's ability to be sharpened. In the case of the High similarity intervener, these feature-feature associations would strengthen one another, potentially facilitating the speed of later retrieval (although potentially not accuracy). In contrast, when ensuring the retrieved subject is sharpenable in the Medium similarity condition, the association between sharpenability and the target's intrinsic properties would be weaker, potentially leading to longer latencies. As a result, we might expect to see the penalties that were observed for Medium similarity interveners here.

Here, we have demonstrated categorically different performance on the basis of separating out intrinsic characteristics such as physical characteristics. In Experiment 7, we will explore differences in similarity on the basis of category membership. Regardless, revisiting the concept of partial matching in such high-dimensional feature space, it is possible that features could be weighted differently, such as category membership being weighted more highly than surface-level physical characteristics. Or it is possible that this is more of a gradient effect, and we could think of it as simply the proportion of overlapping features. In support of this latter view, it is likely that category membership could in fact be determined by the overlapping of a number of defining characteristics. Notions of similarity will continue to be explored throughout the dissertation, and summarized in Chapter 6.

3.7.2 Proximity: Open questions

The studies in this chapter explored the concept of contiguity during initial encoding as close temporal proximity, which was achieved by placing the semantically similar intervener in the subject position of a RC. If close temporal proximity increases the chances of interference between semantically similar competitors, Near conditions should yield larger interference effects than the Far conditions. Evidence in support of this was found, for example, as a larger difference between High and Medium conditions with Near proximity, and for increased rereading times on the RC head in the Near condition.

However, effects in the opposite direction, slower reading times for the Far conditions, were also observed. In Experiment 2, a marginal penalty for higher similarity interveners at the auxiliary *had been* region emerged in the Far condition. This suggests that close temporal proximity during initial encoding isn't always necessary to produce effects of semantic similarity-based interference that cannot be explained by cue overload. These findings pose an interesting question about the process of reactivation at gap sites. The Far conditions were constructed by creating an association between the RC head and a prepositional object gap (e.g. the knife that someone placed the sword near [the knife]). A consequence of the Far condition, therefore, is the reactivation of the RC head in close proximity to the initial encoding of the intervener within the RC.

Interference found in Far conditions then suggests that interference may arise between the initial encoding of a lexical item and the reactivated mental representation of another. Experiment 3 will explore manipulations of temporal proximity in which the corresponding Far condition doesn't also manipulate the position of reactivation for the RC head. Experiment 3 will also address the potential confound that this Proximity manipulation introduces: the fact that the target and intervener in the Near condition are structurally more similar, both occupying subject positions, which could lead to increased retrieval interference during a search for a subject at the main verb.

There is another concern introduced by the Near condition: as the intervener occurs in the subject position of the RC, if any attempt had been made to establish a subject gap, this would result in an attempt to bind both the target and the intervener to the same position, and would likely result in a moment of overlapping activation. Any interference seen as a result of the Near condition could therefore be explained either as an effect of proximal initial encodings, *or* overlapping activation at the RC subject position. It is possible both would contribute to encoding interference, and could be occurring here, but our experimental design is not able to distinguish between these possibilities. One possibility would be to explicitly manipulate animacy, with the expectation that animate RC heads would result in a posited subject gap with the RC, whereas inanimate RC heads would not. Therefore, while holding structure constant, we might expect to see differential levels of encoding interference reflected in downstream processing. Another option would be to explore instances of dependencies in comparison to a baseline where there was no reactivation. This will be addressed in Experiment 7 in Chapter 5.

These questions also motivate the establishment of linguistic domains that matter for contiguity, for example, the role of prosodic or syntactic boundaries in creating encoding contexts. It may be that, despite close temporal proximity, two nouns may be less susceptible to interference if they are on either side of a clause boundary, compared to a case when they are within the same clause. The experiments in Chapter 4 explicitly address notions of contiguity that go beyond temporal contiguity and take into account the specific properties afforded by linguistic structure.

Chapter 4

Encoding contexts

As noted at the end of Chapter 3, the materials for Experiments 1 and 2 manipulated temporal proximity, but in doing so, also manipulated the structural similarity of the retrieval target and intervener. Not only was the Near intervener temporally close to the RC head target, it was also the subject of the RC.

The experiments in this chapter delve deeper into the notion of an encoding context and shared activation. Experiment 3 follows-up on the experiments in Chapter 3 using similar stimuli, but without the confound of structural similarity, to again probe whether temporally proximal encoding may yield a greater likelihood of encoding interference. Experiments 1-3 examine the role of temporal proximity primarily in terms of feature disruption as a result of co-activation, examining the ease of later retrieval, under the linking hypothesis that difficulty in retrieval in cases of overlapping features indicates a disrupted representation of those features or of the binding of those features to one another. Experiments 4 and 5 examine the role of temporal proximity during initial encoding in sentential contexts more directly, under the framework of the Temporal Context Model (Howard & Kahana, 2002). In those experiments (co-authored with Lalitha Balachandran), we examine the degree to which linguistic structure in the form of syntactic or prosodic boundaries may delineate encoding contexts in memory.

4.1 Experiment 3: Manipulating contiguity and not structural similarity

The experiments in Chapter 3 relied on the notion of temporal contiguity in defining a notion of a shared encoding context. However, the Early and Late conditions differed in syntactic structure as well as linear distance between the target and intervener. The Early conditions are not simply the temporally proximal cases, but also the conditions in which both the target and the intervener are in subject positions and carry nominative case. Therefore, the results from Experiment 1 cannot definitively claim that the increased interference is due to temporal proximity, especially when subjects generally are more accessible and retrieved faster than other elements in a sentence (Franck & Wagers, 2018). The explanations of temporal proximity and structural similarity or subjecthood are not mutually exclusive: both structural similarity and temporal contiguity may have been contributing to the effects seen in previous experiments. The interveners were all equally strong competitors in the Early conditions, structurally speaking, so the differential effect of similarity still remains. What is left to be addressed is the notion of temporal contiguity, which we relied on in Experiment 1 as the hypothesized encoding context. Structural similarity has already been shown to be instrumental during the retrieval process. The question that remains is to what degree, if any, temporal contiguity also played a role in creating interference.

To address the potential confound, we devised a follow-up study to introduce temporal distance between the target and the intervener while holding the structure of the relative clause constant. To do this, we added a parenthetical between the target and intervener which did not introduce any discourse referents and were as semantically bleached as possible. The similarity condition was collapsed to high similarity and low similarity, resulting in a 2x2 crossing Proximity (Early, Late) and Similarity (High, Low). The final item set included 40 items from the 42 items used in Experiment 1. No changes were made to the remaining items, other than the addition of the parenthetical in the Late conditions.

(22) a. EARLY-HIGH It seems that the knife that the sword was placed near...
b. LATE-HIGH It seems that the knife that, as you know, the sword was placed near...
c. EARLY-LOW It seems that the knife that the stick was placed near...
d. LATE-LOW It seems that the knife that, as you know, the stick was placed near...
...had been recently sharpened in the kitchen the other day.

As with Experiment 1, this experiment was run on two separate populations with the same stimuli, and will therefore be treated as two subexperiments, Experiments 3a and 3b.

4.1.1 Experiment 3a

Participants

Forty-four participants were recruited via the online experimentation service Prolific. The same participant restrictions from Experiment 1 were used. Additionally, participants who had participated in Experiment 1 were barred from this experiment. Participants were compensated $7 for their participation.

	Early	*Late*
High	0.83 (0.004)	0.77 (0.004)
Low	0.76 (0.004)	0.78 (0.004)

Table 4.1: Accuracy data per condition for Experiment 3a, M (SE).

Procedure

The same experimental online, the remote experimental procedure used in Experiment 1 was used in Experiment 3a. The same data analysis pipeline was used as well.

Question Accuracy

Accuracy data were analyzed using generalized linear models with random intercepts for subjects and items. There was a main effect of contiguity, where Early conditions received higher accuracy than Late conditions, $p < 0.001$, a main effect of Similarity where High similarity intervener conditions were more accurate than Low similarity intervener conditions, $p < 0.001$, and a significant interaction between Similarity and Contiguity, $p < 0.001$. Means and standard errors can be found in Table 4.1.

Results: Relative clause region

In the relative clause regions, there was a main effect of Contiguity in the regions 2 words and 3 words down from the intervener itself. In the region int+2, the Late condition (M = 313 ms, SE = 6 ms) was read faster than the Early condition (M = 321 ms, SE = 6 ms), $t = 2.00$, $p < .05$. The same pattern persisted into the following region int+3 (difference: 25 ms), $t = 3.88$, $p <.001$.

Results: Extended retrieval site

Once reaching the extended retrieval zone, but before reaching the main verb, a similar pattern emerged as the regions within the relative clause. The Late conditions was read more quickly than the Early conditions at had (difference: 10 ms), t = 1.99, $p < .05$, and marginally at been (difference: 8 ms), t = 1.79, $p <$ .1.

Results: Regions at and following the main verb

Once reaching the main verb, there is a main effect of Similarity such that the High conditions resulting in faster reading times than the Low conditions (difference: 14.5 ms), t = -2.39, $p < .05$. This pattern continued as a significant main effect in regions verb+3, t = -2.4, $p < .05$, and verb+4, t = -2.52, $p < .05$. In the verb+3 region, there was also a main effect of Contiguity where the Late conditions (M = 322 ms. SE = 5) were read faster than the Early conditions (M = 330ms, SE = 6), t = 3.79, $p < .001$.

Interim discussion

Experiment 3a found slower reading times for the Early conditions throughout the RC region persisting well into the auxiliary verb region. At the onset, this region follows different material across conditions, so it is possible that some difference in reading times is due to differences in preceding material. No effect of Similarity is found until the main verb, at which point semantically selective information is available. At that point, there is a facilitatory effect of increased similarity. This pattern could be interpreted as evidence of an earlier misretrieval; if that were the case, we might expect that only semantically incompatible nouns would result in an error signal, which could translate to increased reading times.

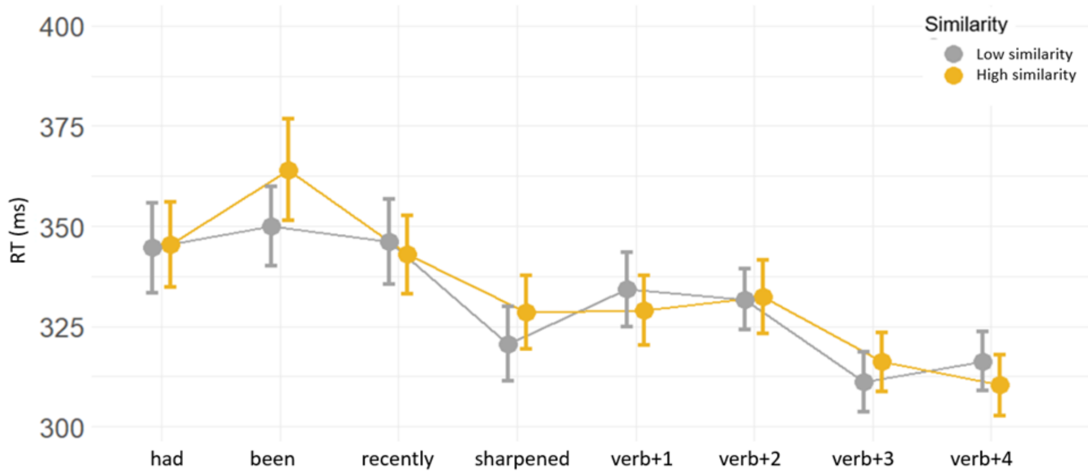

Figure 4.1: Mean reading times in the Early conditions in Experiment 3a in the extended retrieval site regions and spillover regions.

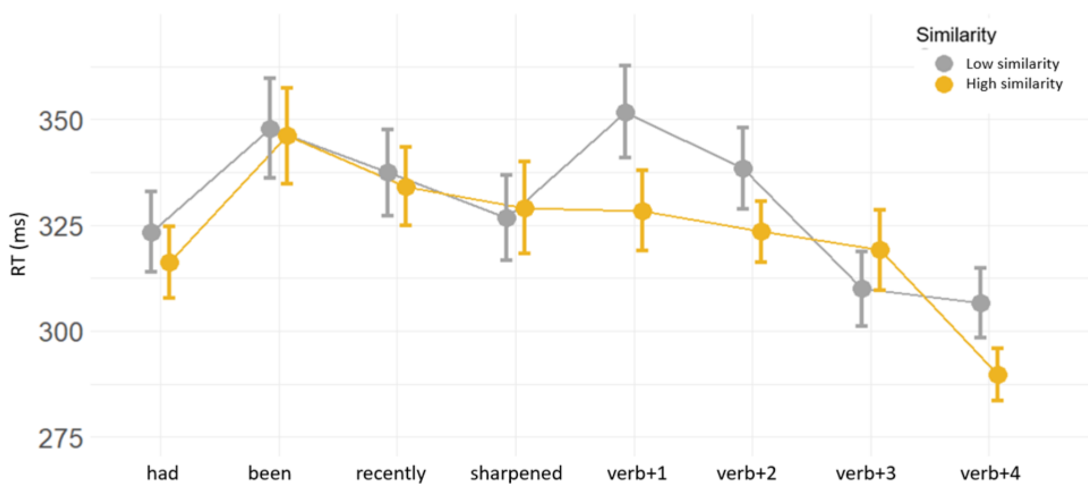

Figure 4.2: Mean reading times in the Late conditions in Experiment 2a in the extended retrieval site regions and spillover regions.

4.1.2 Experiment 3b

Just as with Experiment 1, we aimed to replicate Experiment 3a in the laboratory. However, due to restrictions placed on in-person human subjects testing as a result of COVID-19, we were unable to run the replication in a laboratory setting. Instead, we replicated Experiment 3a using the same materials using the UC Santa Cruz SONA Subject Pool online.

Participants

Sixty-four undergraduate students at UC Santa Cruz were recruited via the UCSC SONA Subject Pool and were granted one course credit for their participation. Participants who indicated that English was not their native language, and who didn't learn English in an immersive setting before the age of eight, were excluded from the final data set, as were participants who disclosed a reading or attention disorder. Participants who scored below 70% on question accuracy were excluded. An additional eight participants were excluded to reach a final fully counterbalanced set of 44 participants.

Procedure

Participants received verbal experiment instructions directly from a researcher over Zoom (https://zoom.us/) and were then provided a subject code and link to the study hosted on IbexFarm. Once they began the experiment, the procedure did not differ from Experiment 1 or Experiment 3a. When they reached the end of the experiment, they were instructed to return to the Zoom meeting for verbal debriefing.

	Early	*Late*
High	0.78 (0.004)	0.77 (0.004)
Low	0.80 (0.004)	0.76 (0.004)

Table 4.2: Accuracy data per condition for Experiment 3b, M (SE).

Question accuracy

Accuracy data were analyzed using the same procedure as in the previous experiment. There was a main effect of Contiguity, such that participants were more accurate in Early conditions than late conditions, $p < 0.001$, and a significant interaction between Similarity and Contiguity, $p < 0.001$. Means and standard errors can be found in Table 4.2.

Results: Relative clause region

Beginning at the intervener, there was a main effect of Contiguity where the Late conditions (M = 365, SE = 8) were read faster than the Early conditions (M = 383 ms, SE = 8 ms), $t = 2.41$, $p < .05$. This pattern persisted in the regions following the intervener: int+1 (difference: 25ms), $t = 3.83$, $p < .001$, int+2 (difference: 40 ms), $t = 5.13$, $p < .001$, and int+3 (difference: 34 ms), $t = 4.01$, $p < .001$.

Only one interaction approached significance: we found a marginal interaction between Similarity and Contiguity in the analysis of the region int+1, such that the High conditions differed significantly between the Early (M = 390 ms, SE = 11) and Late condition (M = 348 ms, SE = 9 ms), whereas the Low similarity conditions did not differ, $t = 1.66$, $p < .1$.

Results: Extended retrieval site

There was a continued significant main effect of Contiguity at the auxiliary *had* such that Late conditions (M = 407 ms, SE = 10 ms) were read faster than Early conditions (M = 450 ms, SE = 11ms), t = 4.54, $p < .001$. There was a marginal main effect of Similarity at *been* where the High conditions (M = 387 ms, SE = 9 ms) were read faster than the Low conditions (M = 397 ms, SE = 8 ms), t = -1.77, $p < .1$. This effect also reached significance at the adverb region (difference: 22 ms), t = -2.49, $p < .05$.

Results: Regions at and following the main verb

No significant effects were found at the main verb, but significant effects were found in the spillover regions, reported below.

There was a significant main effect of Similarity in the verb+2 region such that the High conditions (M = 336 ms, SE = 5 ms) were read slower than the Low conditions (M = 350 ms, SE = 5 ms), t = -2.88, $p < .01$. The same effect was found to be marginal in the verb+4 region (difference: 10 ms), t = -1.91, $p < .01$. These main effects are qualified by the interactions described below.

It should first be noted that there was a significant main effect of Contiguity in the verb+1 region such that the Late conditions (M = 358 ms, SE = 6 ms) were read faster than the Early conditions (M = 374 ms, SE = 6 ms), t = 2.71, $p < .01$, and the same pattern of results carried through the verb+2 (difference: 15 ms), t = 2.59, $p < .01$, and verb+4 (difference: 10 ms), t = 2.09, $p < .05$ regions. These, too, are qualified by the interactions described below.

There were significant interactions between Similarity and Contiguity in the following regions: verb+1, t = 3.46, $p < .001$, verb+2, t = 2.02, $p < .05$, verb+3, t = 2.03, $p < .05$, and verb+4, t = 2.82, $p < .05$. The pattern of data is qualitatively

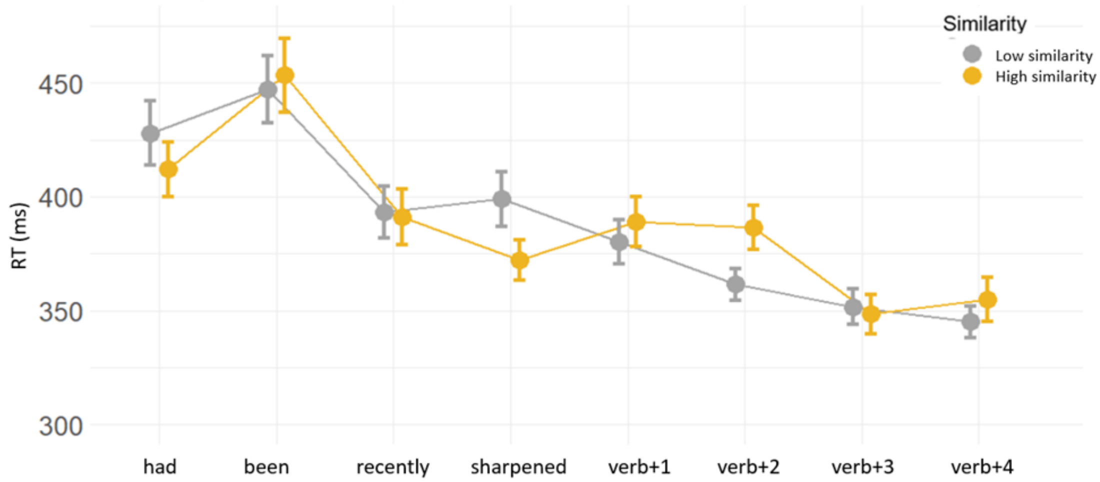

Figure 4.3: Mean reading times in the Early conditions in Experiment 3b in the extended retrieval site regions and spillover regions.

similar across these regions: there the difference between Early and Late for Low similarity is much smaller (difference at verb+1 = 8 ms, difference at verb+2 = 5 ms, difference at verb+4 = 6 ms) or not significant (difference at verb+3 = 11 ms) compared to the difference for High similarity conditions (difference at verb+1 = 39 ms, difference at verb+2 = 25 ms, difference at verb+3 = 19 ms, difference at verb+4 = 27 ms), where the Late condition was read faster than the Early condition.

Interim discussion

In Experiment 3b we found influences of Contiguity throughout the RC region; however, these may be a result of differences in processing following the parenthetical. Crucially, these effects do not carry throughout the sentence, meaning that whatever spillover effects may have resulted from the Late conditions containing additional material subsided before the interactions following the retrieval site. In the extended retrieval zone prior to the main verb, we found an effect of similarity, replicating our previous findings that the degree of semantic similarity between

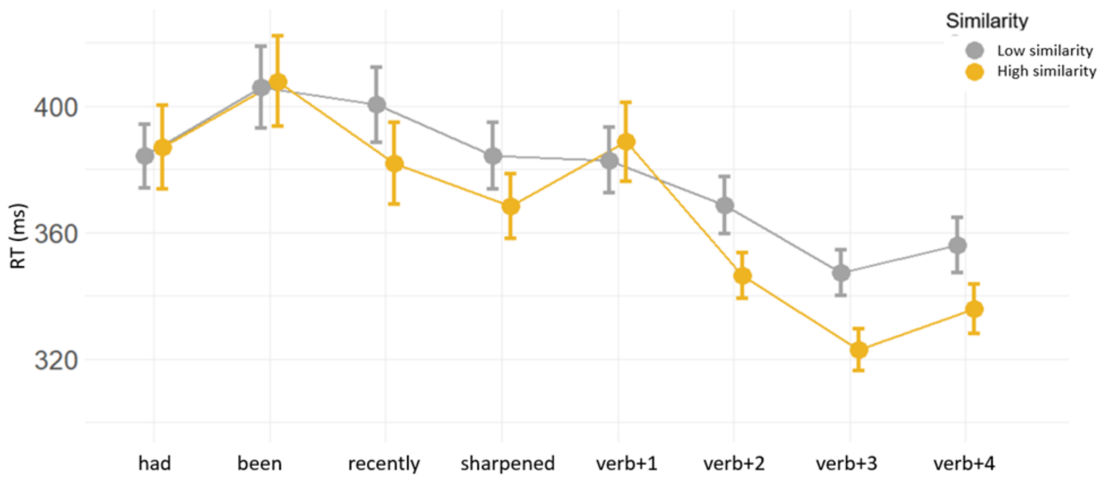

Figure 4.4: Mean reading times in the Late conditions in Experiment 3b in the extended retrieval site regions and spillover regions.

the retrieval target and an intervener can affect retrieval before semantically selective information is reached. Finally, we find a series of interactions following the main verb that demonstrate a larger effect of Contiguity for the highly similar interveners as compared to the dissimilar interveners.

The interactions in the spillover regions, found consistently regardless of the treatment of similarity, are in line with predictions that temporal contiguity at encoding matters for later retrieval: in the High similarity conditions, or for Highly rated pairs, there was a penalty for temporally contiguous encoding. This then did not matter as much for the Low similarity conditions. We interpret this as heightened encoding interference between a target and semantically similar intervener in cases where they are encoded in close temporal proximity, i.e. share an encoding context. This is not to say that structural similarity does not also play a role, but that the interference seen in Experiment 1 may also have been influenced by temporal contiguity.

We find these effects later than predicted. An effect of Similarity does arise before semantically selective cues are available (marginally at been in both analyses;

significantly at the adverb in the categorical analysis), replicating our previous findings that semantic similarity can affect later retrieval even when not used as retrieval cue. However, the critical interactions arise only after the main verb has been encountered. This may simply be a side effect of self-paced reading studies: we can confidently claim that an effect seen prior to a disambiguating region could not be affected by the disambiguating region, but we cannot necessarily claim that an affect seen afterwards is not in fact a result of processing that occurred earlier in the sentence. We assume that the interactions seen following the main verb are a result of processes occurring at that verb or prior, but it is difficult to pinpoint whether the effect hinges on having already encountered semantically selective information. One possible interpretation of this pattern of results is in line with the idea that feature overwriting may lead to increased chances of misretrievals, which might then only be recognized later on. The key findings of this experiment still hold: we do see evidence for semantic similarity based interference compatible with encoding interference as well as an effect of temporal proximity during encoding, holding the structural role of the intervener constant.

4.1.3 Discussion

Experiment 3 sought to expand on the experiments presented in Chapter 3, seeking to understand whether we could detect a difference on the basis of temporal proximity during initial encoding, without manipulating the RC structure that the intervener is encountered in. This was achieved using a parenthetical phrase, such as *as you know*. We found both instances of difficulty, and instances of facilitation, for similar interveners on the basis of temporal proximity. As noted in the previous interim discussion section, this may be indicative of misretrievals, which could then be recognized later on. However, on the whole, the results of this experiment are somewhat inconclusive in terms of the underlying mechanism driving interference.

The design of this experiment begs some questions as to our notions of encoding context. While the manipulation of an added parenthetical doesn't alter the structural similarity of the RC head and the intervener, the interpretation of the material within the parenthetical may prompt additional processing, and activation for neighboring material. For example, Potts (2002) treats *As*-parentheticals as structural sisters to the phrase that must be interpreted within the parenthetical (Figure 4.5). It should be understood then that while these parentheticals may introduce a clausal boundary, they are not fully separate from the following material. However, despite being dependent on the clause in which they're found, previous work on the processing of parentheticals has demonstrated that they do not create the same burden on processing and memory as at-issue content (Dillon, Clifton Jr, & Frazier, 2014).

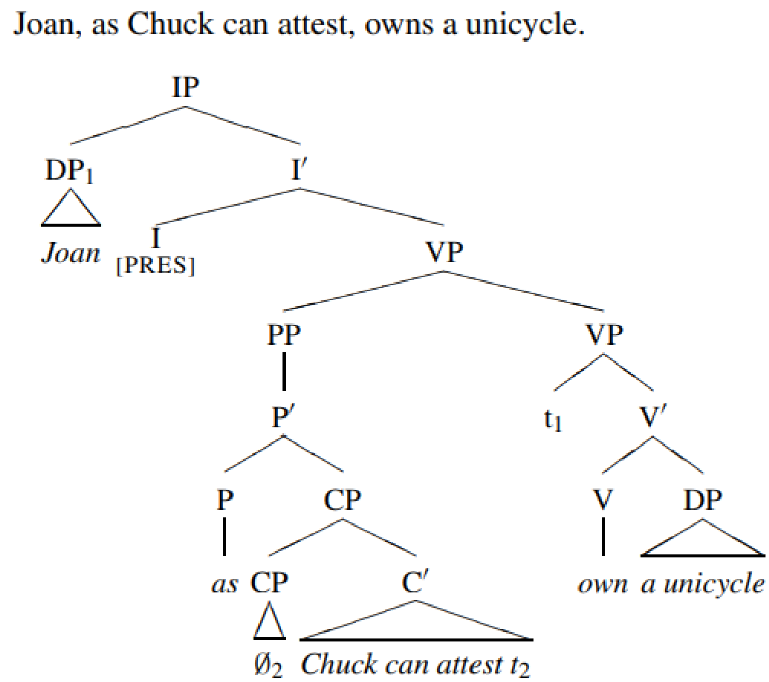

Figure 4.5: Example (63) from Potts (2002) demonstrating the attachment of an *as*-parenthetical.

Returning to the broader question of understanding an encoding context in memory, Experiments 1-3 leave many open questions in terms of pinpointing contiguity in sentential contexts. In the next two experiments, we turn to the specific framework of the Temporal Context Model to understand how contiguity could be understood in linguistic contexts.

4.2 A closer look at the Temporal Context Model

Thus far, we have examined the effect of contiguity on the encoding of syntactic and semantic features. The remainder of the chapter will focus on how contiguity is represented in the actual featural encoding of each item, following insights from the Temporal Context Model.

4.2.1 Encoding contexts in list-memory research

As reviewed briefly in Chapter 2, The Temporal Context Model (Howard & Kahana, 2002) supposes that each item is encoded in memory not just with features describing the item's properties, but also with a feature that represents the context in which it is encoded. This theory depends on the Distributed Memory Hypothesis, namely that the representation of an item in memory is distributed across various attributes, features, and dimensions. We can understand the attributes of an item to be the properties that could be used to specify the identity of an item. For a boat, for example, its attributes would reflect any dimension along which boats may typically vary (e.g. size). A vector representation of an item can be formed by creating an n-dimensional vector for n attributes.

Information about the encoding context can be included in the distributed representation of the item as one of the attributes (a contextual attribute). A simplified way to understand the encoding of the contextual attribute, in which the context encoding doesn't depend on the items being processed, is given in Chapter 3 of Kahana (2012). In this simplified understanding, the context (t) at time i can be understood as the sum of t_{i-1} and some amount of noise, modeled as a random number generated from a normal distribution. This creates a context vector which continually evolves at each time of encoding, resulting in contextual drift between items by some random amount. Items encoded far apart therefore have more dissimilar contextual encodings.

More complex notions of context are introduced by Howard and Kahana (2002). In these conceptions of context, the context vector is also influenced by the identity of the items in the context, creating a contextual representation that contains not just temporal information, but information about the item currently being processed, as well as information that directly came before it. Shifts in the encoding environment (ranging from presentation list in a list-memory experiment, to a person's location) result in a more drastic shift in the context vector.

The continually evolving context vector, and the greater shifts in response to shifts in external context, can explain many findings in list-memory research. One prominent finding has been termed the contiguity effect. Across many different experimental set-ups, a unified pattern emerges: in free recall studies, if a participant has correctly recalled item i, they are most likely to continue on to recall either item i+1 or i-1. This effect is asymmetrical in the forward direction: participants are much more likely to recall item i+1 (Healey et al., 2019). The contiguity effect can be understood as the use of a context vector as a recall cue. This can also explain the robustly found recency effect in list-memory research, the finding that recall accuracy is improved for end-of-list items compared to items presented in the middle of a list. Howard and Kahana (2002) attribute this to the use of the context vector at the time of testing as a recall cue. Naturally, the last item in the presentation list would be closest contextually to the time of testing.

The Temporal Context Model was first used to model results in list-memory research, but there is reason to believe these concepts are applicable outside of this narrow context. Healey et al. (2019) review several cases of contiguity effects in the recall of events. When asked to recall events in their own lives, upon recalling one event, participants are most likely to continue on to recall events within a similar time frame (Moreton & Ward, 2010). Similarly, when asked to recall news stories from a particular time period that was months or years long, participants were most likely to, after the recall of one event, transition to an event that took place within a few days of the first event (Uitvlugt & Healey, 2019).

As it pertains to boundaries that might prompt large shifts in encoded context, a number of influences may result in a shifted context, such as a shift in physical location, a shift in goal, a shift in perceptual input, and narrative shifts, as reviewed by Clewett, DuBrow, and Davachi (2019). These hypothesized shifts in context are evidenced by consistent behavioral findings, such as an overestimation of the time between events occurring across a boundary compared to within, and impaired temporal order memory for items spanning a boundary. To clarify the latter finding, there is a consistent finding that if participants are exposed to a list of items containing some kind of context boundary, and then asked to recall the order of two contiguous items from that list, they will consistently perform worse if the two items spanned a boundary than if they did not (Clewett et al., 2019). However, Wen and Egner (2022) complicate this finding by demonstrating that if items are tied to their context in a meaningful way, and information linked to that connection is available at retrieval, an opposite pattern emerges: participants show decreased accuracy in temporal order memory for items within a context than across a boundary. The authors point out that this speaks to anecdotal experiences of recalling autobiographical memories. They provide the example of remembering events that comprise getting reading in the morning where, anecdotally, it seems more difficult to remember whether you took a sip of coffee before or after having a first bite of a bagel, compared to remembering whether you ate breakfast before or after washing dishes[1]. They argue that this stems from a connection between the events to the context (bagel-breakfast, coffee-breakfast), and that this contextual connection is often present during retrieval (e.g. asking yourself *what did I eat for breakfast this morning?*).

[1] I also may add, anecdotally, that it becomes difficult to remember whether you had microwave mac n cheese for dinner on Wednesday or Thursday, for example, if you spent all evenings that week in the same location (e.g. Room 232 of the Stevenson Academic Building at UC Santa Cruz) doing the same activity (e.g. working on your dissertation).

4.2.2 Connections to language

It has already been supposed that narrative shifts might provide a shift in events (e.g. Zwaan & Radvansky, 1998), but what about at the level of a clause? It has been argued, for example, that syntactic or prosodic structure could provide structure for memory. Therefore, boundaries in syntactic or prosodic structure could serve as boundaries in memory.

In general, grouping material benefits later memory for the grouped items. A common example of this is the practice of writing telephone numbers in groups of three or four, which has been argued to help provide structure for those items in memory and facilitate recall (Bower & Winzenz, 1969). On the prosodic level, Frankish (1995) expanded the study of memory for grouping to include aspects of natural prosodic structure, finding, for example, that abrupt changes in pitch at group boundaries facilitated later recall. On a syntactic level, it has been proposed that structure lessens demands on working memory during sentence parsing (Frazier & Fodor, 1978).

Wagers (2008) hypothesized that, during syntactic processing, clausal membership information could be captured by means of encoded contextual vectors, supposing that the context shift would be greater upon entering a new clausal domain. This would have the result of items within one clause being more featurally similar to one another than to items in another clause. The following two experiments will build on this idea specifically.

4.3 Motivating Experiments 4 and 5

Experiments 1-3 have examined instances of interference during encoding and retrieval have done so using reading methodologies, following a linking hypothesis that reading times correspond to processing difficulty. Following this linking hypothesis, when looking at the reading profile of a retrieval site in a condition with two similar candidates compared to a baseline, longer reading times can be understood as inhibitory interference, and faster reading times as facilitatory interference. However, in order to make direct connections between the field of language processing, and research on domain-general memory, we aimed to use explicit tests of memory. Experiments 4 and 5 explicitly test the effects of clausal membership and linguistic boundaries on recall and recognition, seeking to replicate findings surrounding contiguity in recall and temporal order memory during recognition in sentential contexts.

The highest level research question of the following two experiments surrounds the concept of an *encoding context*, and the effect linguistic boundaries have on context encodings.

In list-memory studies, a context can be understood as one list. Within that context, the contextual vectors of each item in the list are similar to one another, with neighboring items sharing the most similar contextual vector. When instead thinking of a full sentence, in which words are grouped with hierarchical structure, there are several possibilities for how these structural properties might interact with the gradually evolving contextual vector. Experiments 4 and 5 question whether syntactic boundaries (such as clause-level boundaries) or prosodic boundaries delineate these contexts in memory.

In both experiments, we entertain two hypotheses: Linguistic Sensitivity, in which linguistic cues to grouping (e.g. syntactic or prosodic boundaries) influence the incremental change of the contextual vector such that neighboring items within a boundary have closer contextual vectors than neighboring items across a boundary; and Temporal Contiguity, in which temporal contiguity outweighs any additional cues that linguistic structure may provide, resulting in a roughly equivalent difference in contextual vectors for items within or across boundaries. The specific predictions as they pertain to each experiment will be explored in each experiment subsection respectively.

4.4 Experiment 4: Sensitivity to linguistic boundaries during recall

Experiment 4 used a free sentence recall methodology to examine the influence of linguistic boundaries on contextual reactivation. The hallmark pattern of contextual reactivation, coming from the list memory literature, is the forward asymmetrical contiguity effect, in which correct recall of word w_i increases the likelihood of subsequent correct recall of w_{i+1} and w_{i-1}, with the greatest likelihood being the subsequent correct recall of w_{i+1}. Experiment 4 tested whether this pattern of results could be observed in sentential stimuli that contained linguistic cues to grouping, specifically clausal boundaries.

As it pertains to recall, the Temporal Context Model predicts that not just the recalled item is reactivated, but also other items with similar contextual features, resulting in the asymmetrical contiguity effect. I will refer to this as contextual reactivation. The two hypotheses outlined above, Linguistic Sensitivity and Temporal Contiguity, make different predictions about the role of linguistic cues to grouping on the phenomenon of contextual reactivation.

The Linguistic Sensitivity hypothesis, in which the encoding of contextual vectors is influenced by linguistic cues to grouping, would predict that contextual reactivation would be bounded by those linguistic groupings. Mechanistically, this would mean that at a clause boundary, the contextual drift between the words on either side of the boundary would be much larger than the drift between two neighboring words within the same clause. During recall, if one word is reactivated, a within-clause neighbor should receive a boost in recall as a result of activation of similar contextual vectors, whereas an across-clause neighbor should receive a slight boost, if at all. In other words, there should be no recall advantage for $w_{\mathrm{i+1}}$ following reactivation of w_{i} if w_{i} and $w_{\mathrm{i+1}}$ are separated by a linguistic boundary.

The Temporal Contiguity hypothesis, in which the role of temporal contiguity in the encoding of contextual vectors overshadows any influence of linguistic cues to grouping, would predict that contextual reactivation would *not* be bounded by those linguistic groupings. Mechanistically, this would mean that contextual drift between neighboring words either does not differ depending on the presence of a clause boundary between them, or the difference is so small as to not have any effect on the pattern of contextual reactivation. Under this view, contextual reactivation is determined by contiguity, regardless of linguistic boundaries, predicting contiguity effects cross linguistic boundaries, or, in other words, a recall advantage for $w_{\mathrm{i+1}}$ following reactivation of w_{i} even if w_{i} and $w_{\mathrm{i+1}}$ are separated by a linguistic boundary.

The recall study presented here is designed specifically to test for levels of contextual reactivation within and across linguistic boundaries to demonstrate whether the presence of a linguistic boundary delineates the boundaries of contextual reactivation and to provide insight on how contextual information is encoded incrementally in sentential contexts.

4.4.1 Methods

Materials

The stimuli for this experiment consisted of 48 sentences, each containing four comma-separated simple SVO clauses. Each clause contained definite, animate subjects and objects which were distinct across all clauses such that the same noun was never seen twice within one sentence. A number of factors were controlled for to ensure differences in recall could not be attributed to unintended coincidences in noun placement or lexical characteristics. While these were clauses, to be agnostic as to how participants were representing these stimuli, we will refer to them as chunks.

It was the case that some nouns were used more than once across the entire experiment, for a maximum of three times. To avoid potential interference between items creating differences in recall, care was taken to ensure that any potential repetition of nouns avoided systematicity that could create a confound. For example, we wanted to avoid an instance where recall for a particular noun was boosted because it had been encountered in the same position in a previous item. Therefore, repetitions were controlled such that a repeated noun was never found in the same syntactic category across items (e.g. if a noun was a subject in one item, and then repeated in another item, it would be an object in the other item). Whenever possible, there was an effort to ensure the repeated nouns did not occur in neighboring chunks across items (e.g. if a noun was in Chunk 1 in one item, it would not appear in Chunk 2 in another item), in order to avoid any potential cross-item effects of contiguity.

To avoid potential confounds as a result of lexical characteristics, e.g. a recall boost for a certain position due to an unintended likelihood for high frequency words in that position, several lexical characteristics were held consistent across the nouns in any given item. Nouns were length-matched as much as possible (within 1-2 characters of one another) as well as frequency-matched. Frequency was determined using a word's LgSUBTWLF frequency as given by the English Lexicon Project (Balota et al., 2007), a metric of frequency that represents the log frequency of a word in the SUBTLEX corpus. Within a given item, nouns were checked so that they did not exceed a +/- 1 difference in LgSUTBTWLF frequency to the other nouns. Furthermore, the full range of lexical frequency was considered across the experiment, such that the distribution of frequency was even across the different noun positions. Finally, similarity was controlled for using the word2vec word embedding metric of similarity (Mikolov, Chen, Corrado, & Dean, 2013), via the University of Colorado Boulder word embedding analysis website (http://wordvec.colorado.edu/). Within an item, pairwise similarity between nouns was kept below a threshold of .35 in order to avoid instances of facilitation in recall due solely to a benefit from a present semantic associate.

As we were interested in investigating the shape of contextual reactivation in sentential contexts, we prompted participants to recall a specific noun from the sentence prior to recall, to assess whether targeted reactivation of a particular noun would lead to recall benefits for the following noun, even across a clause boundary. I will refer to this as the reactivation question, which was manipulated across three levels in a within-subjects design. The reactivation question targeted either the Chunk 2 object (Ch2Q) or the Chunk 3 object (Ch3Q). This was compared to a third baseline in which there was no baseline question (NoQ). An example of an item and the possible reactivation questions can be seen in (23).

(23) The guest$_{\text{Ch1-Subj}}$ loved the voter$_{\text{Ch1-Obj}}$, the miner$_{\text{Ch2-Subj}}$ loved the guide$_{\text{Cl2-Obj}}$, the enemy$_{\text{Ch3-Subj}}$ loved the groom$_{\text{Ch3-Obj}}$, and the boxer$_{\text{Ch4-Subj}}$ loved the artist$_{\text{Ch4-Obj}}$.

NoQ: —–

Ch2Q: Who did the miner$_{\text{Ch2-Subj}}$ love?

Ch3Q: Who did the enemy$_{\text{Ch3-Subj}}$ love?

Participants

Forty-eight participants were recruited via Prolific and compensated with $12 for their participation. From the large subject pool available through Prolific, participants were screened to include only those participants who had marked English as a first language, had completed at least a high school level of education, and did not report any literacy difficulties (including dyslexia and ADHD). One participant had to be excluded, and will therefore not be included in the final, analyzed data set.

Procedure

Participants accessed the experiment link through Prolific, and were instructed to take the experiment at a time when they could complete the experiment in one sitting and in a location with minimal distractions. The experiment was expected to take 60 minutes, but the actual time varied across participants (min: 32 minutes; max: 2 hours, 41 minutes). As this experiment was run online, we could not anticipate how long participants spent actively engaged in the task, or whether they may have taken a long break.

The experiment itself was built and hosted on the platform PCIbex. Each trial consisted of three phrases: an exposure phase, in which the participant read through the sentence in clause-by-clause cumulative self-paced reading; a reactivation phase, in which participants were prompted to respond to one of the reactivation questions (or no question); and finally a recall phase, in which par-

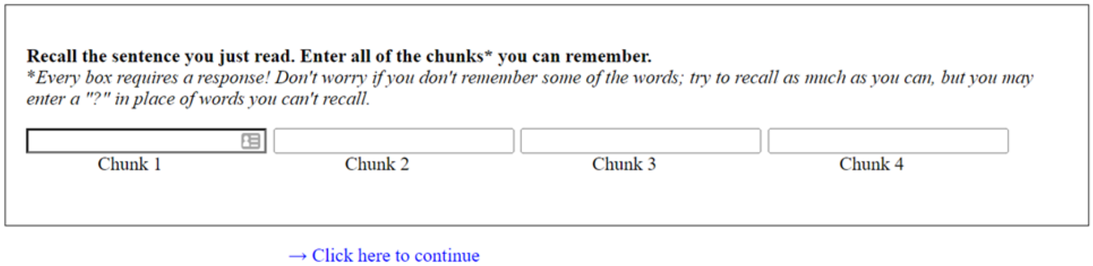

Figure 4.6: Participant view of the recall phase.

ticipants were instructed to type the full sentence verbatim into four text-entry boxes. The participant view of the recall phase can be seen in Figure 4.6.

Before beginning the experimental trials, participants completed a series of guided practice items which gradually introduced them to the phases of the experiment. They were instructed to do their best to remember the sentence, but that if they couldn't remember any component of the sentence, they could insert a '?' to represent the word they had forgotten.

Data processing and analysis

In order to analyze these data, two approaches were taken. First an accuracy measure was determined automatically on the basis of whether the nouns were recalled in the correct chunk. To understand accuracy by position, the data were fit to Bayesian mixed effects models with random intercepts and slopes for participants and items using the BRMS package (Bürkner, 2017) in R with default priors and a Bernoulli family. Contrasts were coded such that Ch2Q and Ch3Q were separately compared to the baseline condition, NoQ.

To examine intrusions, or responses that were not '?' but were not the intended noun, incorrect responses for Chunks 2 and 3 were hand coded. The following properties of the intrusion in relation to the correct noun were marked: whether it was a (morphologically) related word, a phonological competitor, an orthographic competitor, and/or a semantic competitor. The following relationships between the intrusion and item were recorded: whether the intrusion was an intra-item intrusion, extra-item intrusion but occurred elsewhere in the experiment, or extra-experiment intrusion. Finally, if the intrusion was an intra-item intrusion, the positional relationship between the intrusion and the correct noun were recorded: whether the intrusion came from a contiguous clause or a contiguous position (forwards or backwards). We also noted instances of the subject and object swapping within a chunk, and instances of subjects intruding on a subject position or objects intruding in an object position.

4.4.2 Predictions

Regardless of which hypothesis is correct, we predicted that the presence of a reactivation question would boost recall for at least one noun. After all, the reactivation question does present one of the nouns to the participant (either the Ch2 subject or the Ch3 subject). The hypotheses differ in terms of how much recall accuracy would increase at other positions. Under the Temporal Contiguity hypothesis, we would expect to see a boost in recall following a reactivation question following the forward asymmetrical pattern established in previous literature, disregarding the presence of a clause boundary. Therefore, a reactivation question such as the CH2Q question *Who did the miner love?* should boost recall not just for *miner* and *guide*, but also *enemy*. In contrast, under the Linguistic Sensitivity hypothesis, we would not predict the reactivation benefit to extend into the next clause, e.g. to *enemy*.

Q-Corr	β	95% CrI	β	95% CrI
		Ch2Q		Ch3Q
Ch2-Subj	**1.8**	(1.2,2.5)	0.1	(-0.3,0.5)
Ch2-Obj	**2.3**	(1.6,3.2)	0.1	(-0.3,0.5)
Ch3-Subj	-0.09	(-0.4,0.2)	**1.6**	(1.1,2.2)
Ch3-Obj	-0.2	(-0.5,0.08)	**1.6**	(1.2,2.1)

Table 4.3: Output for brms mixed effect model for the correct trial analysis.

4.4.3 Results

Accuracy: Trials with correct reactivation Q responses

Compared to the NoQ baseline, there were credible differences between the conditions. The differences followed the predictions of Linguistic Sensitivity, such that following a correct response to the Ch2 reactivation question (compared to the NoQ baseline) participants were more accurate in their recall of both the Ch2-Subj and Ch2-Obj, but not in their recall of even the immediately following word, Ch3-Subj. Similarity, following a correct response to the Ch3 reactivation question, compared to the NoQ baseline, participants were more accurate in their recall of the Ch3-Subj and the Ch3-Obj, but not in their recall, for example, for the position immediately preceding, Ch2-Obj. Model outputs can be found in Table 4.3, and the accuracy means by position are presented in Figure 4.7.

Accuracy: Trials with incorrect reactivation Q responses

Unlike recall following correct responses to the reactivation questions, credible differences in recall accuracy in each condition compared to the NoQ baseline were not confined to the reactivated chunk. See Table 4.4 for model output, and Figure 4.8 for accuracy means by position.

Following an incorrect response to Ch2 reactivation question, participants were less accurate in recalling Ch2-Obj. This is not so striking: any correct recalls of

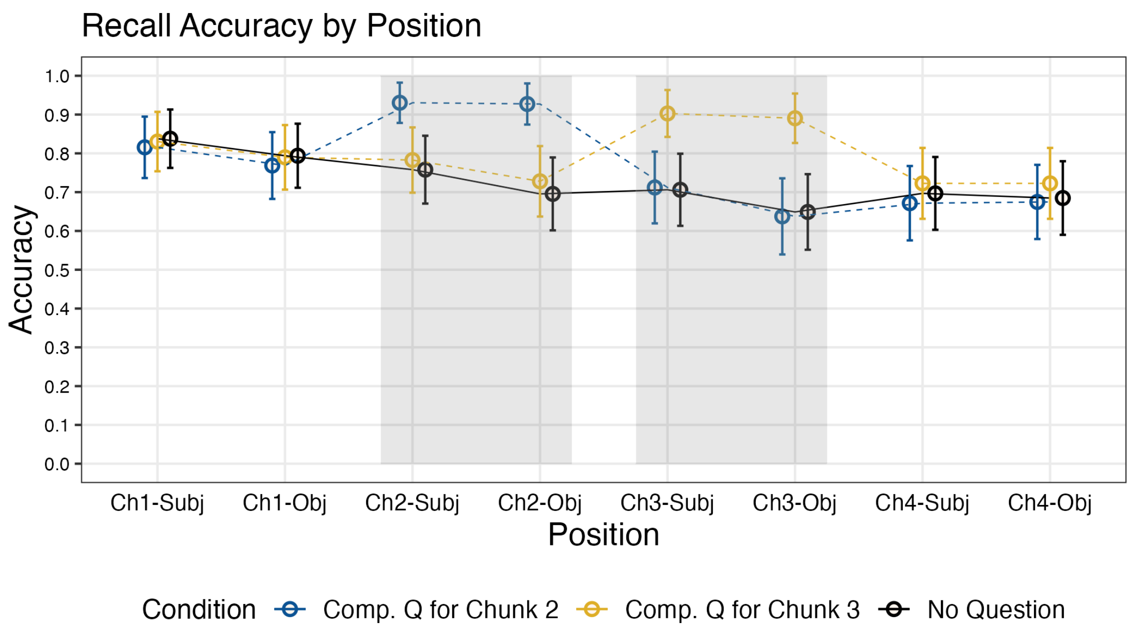

Figure 4.7: Recall accuracy following a correct response for the reactivation question.

Ch2-Obj here in fact reflect an initial instance of forgetting Ch2-Obj and then remembering upon viewing the full recall prompt. However, this decrease in accuracy is not limited to Ch2: Ch3-Subj and Ch3-Obj are both less accurately recalled following an incorrect recall of Ch2-Obj, compared to the NoQ baseline.

Similarly, following an incorrect response to the Ch3 reactivation question, participants were less accurate in recalling not only Ch3-Obj, but also both positions in the preceding chunk, Ch2-Subj and Ch2-Obj.

Q-Incorr	β	95% CrI	β	95% CrI
		Ch2Q		Ch3Q
Ch2-Subj	-0.7	(-1.4,0.02)	**-1.2**	(-1.8,-0.7)
Ch2-Obj	**-3.2**	(-5.8,-1.8)	**-1.0**	(-1.6,-0.4)
Ch3-Subj	**-1.7**	(-2.3,-1.2)	-0.5	(-1.3,0.4)
Ch3-Obj	**-1.1**	(-1.7,-0.5)	**-3.4**	(-5.1,-2.3)

Table 4.4: Recall accuracy following an incorrect response for the reactivation question.

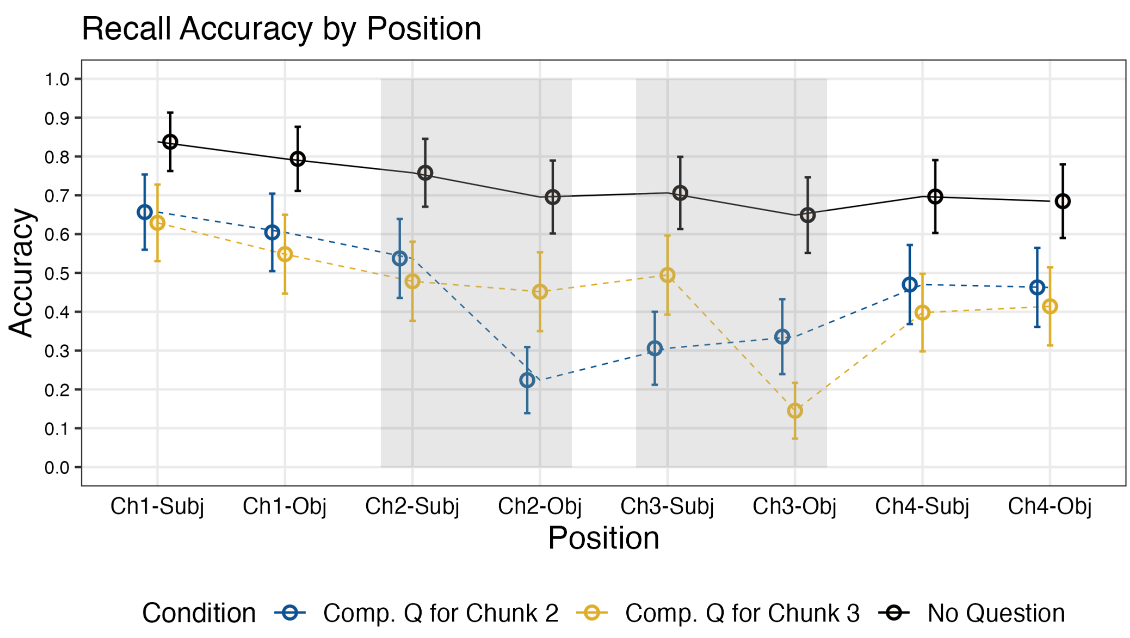

Figure 4.8: Recall accuracy following an incorrect response for the reactivation question.

Intrusions

As participants were given the option to respond '?' if they had forgotten a word, overall rates of intrusions were relatively low, ranging between 7-11% of errors across the positions we analyzed. See Table 4.5 for the rate of intrusion by position, as well as how many of those intrusions were from the experiment but not that particular item (extra-item intrusions) and intrusions from within that trial, i.e. from within that item (intra-item intrusions).

As there were relatively few intrusion errors, we opted to present the overall distribution of responses, with broad generalizations. In future work, we hope to induce more intrusion errors, potentially by discouraging blank or '?' responses.

Looking specifically at the within-item intrusions (see Table 4.6), we observed several trends in the distribution of intrusion type. Many of the intrusions came from contiguous chunks (50-65%), and in those cases, it was most often the case that the intrusions came from the following chunk, particularly in the case of Chunk 2.

Intrusions from contiguous positions represented a relatively small portion of intrusions, either from a position across a chunk boundary (2-15%), or from a contiguous position within the same chunk (2-9%). Interpretively this means that while it was common to see intrusions from neighboring chunks, they were often

	Raw count	% intrusions	Extra-item	Intra-item intrusions
Ch2-Subj	129	9.63%	17.05%	72.87%
Ch2-Obj	122	10.18%	25.41%	70.49%
Ch3-Subj	92	7.26%	27.17%	70.65%
Ch3-Obj	109	8.6%	33.94%	63.3%

Table 4.5: Intrusion data: Percent of errors that were intrusions, by position. In the following two columns, of the intrusions, the breakdown of extra- and intra-item intrusions.

	CC_{all}	CC_{for}	CC_{back}	CP_{AC}	CP_{WC}
Ch2-Subj	64.34%	44.96%	19.38%	1.55%	8.53%
Ch2-Obj	59.02%	40.99%	18.03%	7.38%	3.28%
Ch3-Subj	55.43%	20.65%	34.78%	2.17%	7.61%
Ch3-Obj	50.46%	23.85%	26.61%	15.6%	2.75%

Table 4.6: Percent of intrusions (out of all intrusions in each position) that displayed the following characteristics: CC, an intrusion from a contiguous chunk; CC_{for}, CC intrusions from the following (forward) chunk; CC_{back}, CC intrusions from the preceding (backward) chunk; CP_{AC}, intrusions from a contiguous position across a chunk boundary; CP_{WC}, intrusions from a contiguous position within the same chunk.

from non-contiguous positions, i.e. an intrusion into a subject position was most likely to come from the following subject, not the contiguous objects on either side. While the number of observations is too small to determine if this is a consistent effect, we did also note that within a single chunk, objects were more likely to intrude into the subject position (7-9%) compared to subjects intruding into the object position (2-4%).

4.4.4 Discussion

The results of Experiment 4 overall support the hypothesis of Linguistic Sensitivity, that linguistic boundaries can delineate encoding contexts in memory. This is seen in the lack of recall facilitation in the positions surrounding a reactivated chunk.

An interesting pattern emerges in the accuracy following incorrect responses. Lower accuracy is confined not just to the chunk that was probed during the reactivation question phase, but also in the surrounding chunks. One could argue that this is simply a matter of poor attention across the entire trial, but these decreases in accuracy are not as severe in the first chunk (although we would expect the first chunk to receive some benefit from primacy).

To return to the origins of the contiguity effect in recall generally, it is predicted that the use of a contextual feature as a retrieval cue during recall would prompt recallers to, once establishing one remembered token, continue following that cue until it no longer yields any further tokens. The lack of recall boost in neighboring chunks in the correct response trials suggests that, for example, Ch3-Subj might be feautrally distinct enough (in terms of the encoded context feature) that a boost in activation for Ch2-Obj does not facilitate the use of a contextual retrieval cue to correctly recall the next item. On the other hand, the lowered accuracy around a locus of error in the recall following incorrect responses to the reactivation question may be an indication of disruption in contextual encoding. To give a concrete example, this could look like, during an attempted reactivation of Ch2-Obj, Ch3-Obj is reactivated instead, and upon retrieval during the reactivation question, re-encoded with a contextual feature that now includes information about the Ch2-Subj. This has a consequence of disrupting sequence information: the contextual encoding of Ch3-Obj is no longer a valid retrieval cue for the next positions, and could be more likely to result in further intrusions, or blank responses.

Intrusion data illustrates a role for linguistic information and structure over temporal contiguity, with subjects often intruding into other subject positions (and objects into object positions), despite the fact that these are not temporally contiguous positions. It does seem, though, that there is an effect of contiguity on the level of a chunk, as these intruding subjects often come from a neighboring chunk.

We have thus far remained agnostic about what kind of 'chunk' these small clausal units are treated as by participants, and what kinds of boundaries the participants might be sensitive to in this experiment. At least three signals to boundary can be found in this experiment: presentational chunking, where the sentences are uncovered chunk-by-chunk; syntactic chunking, in which each chunk corresponds to a full clause; and prosodic chunking, in which the implicit prosodic structure that participants might assign to these stimuli would likely reflect some level of boundary at the chunk boundaries (e.g. a pause). Future work following this experiment and line of research would aspire to confirm what types of boundaries participants are sensitive to, and confirm that it is indeed the linguistic and not presentational cues to boundaries they attend to. Sensitivity to structural position in the intrusion data provides preliminary evidence that participants are sensitive to the linguistic structure, rather than simply viewing these stimuli as grouped (but otherwise unstructured, i.e. no hierarchical structure beyond the groupings) word lists. One possible future study to expand on this might attempt to manipulate the ways that participants treat these stimuli: either as contentful, meaningful sentences, or grouped word lists, by manipulating the types of questions asked (*Who did the miner love?* compared to *What noun came after miner?*). We hypothesize that questions that treat the stimuli as an unstructured word list may diminish the clause-boundedness of the recall boost for correct reactivation question responses[2].

One limitation in the analysis of intrusions was the low number of intrusion errors given. In future iterations, we may instruct participants to guess a contentful word, even if they had forgotten. Another design change that would be beneficial to explore in the future is in the collection of recall responses. The traditionally established contiguity effect is specifically in reference to the order of recall, but the set up of this experiment did not allow for us to analyze the typed order of responses. Instead, here, we reason about effects of contiguity on the basis of accuracy patterns. In future iterations, during the recall phase we

[2]Thanks to Mandy Cartner for this suggestion.

could prompt participants to respond word by word, or collect responses verbally. Lastly, another limitation of this study is that this experimental setup may not be fully sensitive to errors within a clause, as the reactivation question always provided the subject position in Chunk 2 or Chunk 3. This will be addressed in Experiment 5.

4.5 Experiment 5: Sensitivity to linguistic boundaries during recognition

Many of the experiments in this dissertation examine interference on the basis of semantic similarity. Many of the theories referenced emphasize a role for co-activation, temporal contiguity, and/or shared encoding context. This chapter focuses on this notion of encoding context, and entertains whether linguistic boundaries delineate encoding contexts in memory. Experiment 4 examined recall for nouns, not on the basis of semantic similarity, finding that reactivation for one noun did not boost recall for the following noun across a clause boundary, providing preliminary evidence for the clause boundary increasing the contextual difference between two contiguous nouns. Experiment 5 sought to extend this line of research to examine whether two highly similar nouns are more susceptible to interference within a clause as compared to across a clause boundary. This interference is operationalized in Experiment 5 through a recognition memory task. We hypothesized that a larger degree of interference, or greater likelihood of interference, would lead to overall less accurate recognition memory. The two hypotheses considered for Experiment 4 can be considered here again.

The Linguistic Sensitivity hypothesis, to restate again, suggests that the encoding of contextual vectors is influenced by linguistic cues to grouping, resulting in two contiguous nouns having a much larger difference in contextual vector if they occur on opposite sides of a linguistic boundary compared to two contiguous nouns within boundaries. Now adding the layer of semantic similarity, the Linguistic Sensitivity Hypothesis suggests that two neighboring semantically similar nouns are differentially similar depending on the context in which they are encoded: there would be an overall higher degree of featural overlap if they were encountered and encoded in the same clause. We would therefore predict to see lower accuracy rates in recognition memory for within-clause semantically similar competitors, compared to across-boundary similar competitors. Given previous findings about temporal order memory, wherein temporal order memory is worse within a context specifically when the items are meaningfully associated to the context, we are also hypothesizing that sentential contexts are able to do just that: provide a meaningful connection between the items and the context itself.

The Temporal Contiguity hypothesis, on the other hand, suggests that any difference stemming from linguistic boundaries is overshadowed by temporal contiguity. Therefore, the overall featural overlap between two neighboring semantically similar nouns would not depend on whether a clause boundary came between them. Consequently, we would not expect to see recognition memory performance depend on the clauses where each noun was first encountered and encoded.

A third hypothesis emerges, one in which by virtue of presentation, the clause boundary is interpreted as a boundary, but the linguistic context does not provide a meaningful connection between the items and the context itself. If this were the case, we would predict that participants would in fact have worsened temporal order memory for items across a clause boundary rather than within.

4.5.1 Methods

Materials

The stimuli for this experiment consisted of 48 bi-clausal sentences. Each clause had a conjoined DP subject (e.g. *the butcher and the landlord*). In each sentence, two of the nouns were semantically similar to one another. We manipulated whether the semantically similar items appeared in the same clause (Within clause), or across a clause boundary (Across clause boundary). This distinction can be seen in (24).

An elided verb phrase was always used for the first clause, for several reasons. Firstly, this decreases the temporal distance between encountering the second noun and the third noun, minimizing a potential confound that a decrease in interference across a clause boundary actually stems from greater intervening material (and time). Secondly, this prevents the argument structure and lexical semantics of the verb from having an effect on how the nouns are initially encoded. For example, any verb that results in a reciprocal interpretation (e.g. *dated*, *hugged*, *met*) may introduce a tighter relationship between the two nouns than a verb that does not (e.g. *swam*, *skipped*). The cataphoric structure used in the items avoid any potential effects of this during the initial encoding, but does not rule out any effect of this later on, when the full verb phrase is reached in the second clause.

Each item had a corresponding set of probe sentences, used during the probe recognition phase of each trial, seen in (25). The probe sentence was either exactly the same as the exposure sentence (Match), or different (Mismatch). All mismatches for experimental stimuli switched the position of the semantically similar words, consequently either within a clause boundary or across.

(24) Exposure sentence

a. Within: Before the butcher and the landlord could, the *scientist* and the *researcher* solved the problem.

b. Across: Before the butcher and the *scientist* could, the *researcher* and the landlord solved the problem.

(25) Probe sentence (Mismatch conditions)

a. Within: Before the butcher and the landlord could, the *researcher* and the *scientist* solved the problem.

b. Across: Before the butcher and the *researcher* could, the *scientist* and the landlord solved the problem.

This resulted in a 2x2 design, crossing Similarity (Within, Across) and Probe (Match, Mismatch). As with Experiment 4, careful attention was paid to the lexical characteristics of the words used. Word length was controlled within each item such that the difference in length between any two words was not greater than two. The similarity between the unrelated nouns was also controlled for, using the same methods as in Experiment 4.

Participants

As in Experiment 4, 48 participants were recruited to take this experiment. The same process, including screening criteria, was used in this experiment as for Experiment 4. Three data files were corrupted, so data from 45 participants is represented in this analysis.

Procedure

The same experiment hosting process used in Experiment 4 was used in Experiment 5.

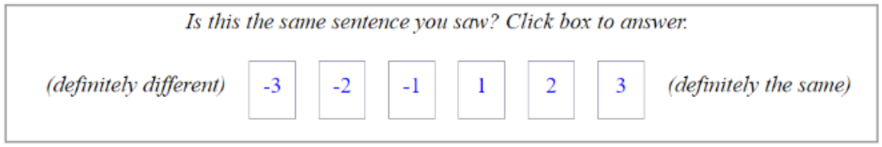

Figure 4.9: Recognition decision judgment scale.

During the experiment itself, each trial consisted of three phrases: an exposure phase, during which the participant read the sentence one word at a time in moving window display self paced reading; a math distractor phase, during which the participant was prompted to respond to a simple, randomly generated math problem (addition or subtraction of values 0-9); and finally a recognition memory task, during which participants were shown a target sentence in its entirety and asked to judge if it was the same as the exposure. Participants were instructed to indicate their judgments on a 6-point scale from -3 to 3. Negative numbers corresponded to a response of *definitely different* and positive numbers corresponded to a response of *definitely the same*. Confidence was reflected in the distance from 0, with -3 and 3 corresponding to more confident responses, and -1 and 1 responses to less confident responses. The decision prompt as seen by participants can be found in Figure 4.9.

The items were displayed alongside 60 fillers, which were constructed to obscure the central design of the experiment. As the experimental items always probed a difference either between N3 and N4 (Within clause condition) or N2 and N3 (Across clause condition), the fillers included mismatches between other noun positions, as well as mismatches between anaphoric and cataphoric verb phrase ellipsis. The filler exposure sentences included both structures that were identical to the experimental sentence structures, as well as those that were different. The balance of Match and Mismatch responses was kept evenly split across the experiment.

Data processing and analysis

Participant math accuracy was intended to be used to exclude participants who were not attending to the experiment. All recruited participants scored above 75%, and therefore no participants were excluded.

Results were analyzed using Unequal Variance Signal Detection Theory (Hautus, Macmillan, & Creelman, 2021; Dillon & Wagers, 2021) using the pROC package (Robin et al., 2011) in R. We measured accuracy in each condition (d_a) as well as sensitivity (area under the receiver operating characteristic curve; AUC). To compare sensitivity across conditions, D_{boot} was computed using stratified bootstrapping (2000 bootstrap replicates) using the roc.test function.

4.5.2 Predictions

If the encoded contextual features for the nominal items in this experiment were dependent on temporal contiguity, we would not expect to see differential sensitivity to noun-swaps within a clause compared to across a clause boundary.

If, however, clause boundaries do serve to delineate contexts in memory, we would expect nouns encoded in the same clause to be more similar to one another in terms of their featural encoding of context. Under one version of Linguistic Sensitivity, this would result in decreased sensitivity to noun-swaps within a clause compared to across clause boundaries, suggesting decreased temporal order memory for items within the same encoding context, which is predicted for contexts in which the individual items have a meaningful relationship to the context itself. Therefore, this pattern of results secondarily confirms that sentential contexts, unlike unstructured word lists, give rise to stronger ties between specific items and their encoding context.

Under another version of Linguistic Sensitivity, clause boundaries are used to

Condition	d_a	AUC	95 % CI
Within	1.1	0.78	[0.75, 0.81]
Across	1.4	0.84	[0.82, 0.87]

Table 4.7: Results of pROC Signal Detection Theory model.

delineate encoding contexts in memory, but do not provide the connection between item and context as proposed by Wen and Egner (2022). This would result in an opposite pattern of results, where temporal order memory would be better within clauses rather than across. Consequently, participants would be more sensitive to swaps within clauses.

4.5.3 Results

We found greater sensitivity (AUC) to changes in the order of related nouns in the Across conditions, where the related nouns occurred across a linguistic boundary, compared to Within conditions, where the related nouns occurred within the same linguistic grouping (D_{boot} = -3.3, p = .001). The results of the model can be found in Table 4.7, and ROCs in Figure 4.10.

4.5.4 Discussion

In sum, Experiment 5 tested the sensitivity of readers to swaps between highly similar nouns. We found that participants were sensitive to these swaps, more so when nouns were exchanged across a clause boundary compared to when these swaps happened within a clause. This pattern of results provides further support for the broader hypothesis of Linguistic Sensitivity entertained in this chapter: linguistic boundaries appear to be used to delineate contexts in memory.

In the case of this experiment, we see that temporal order memory decreases within a clause, which provides evidence for the version of Linguistic Sensitivity

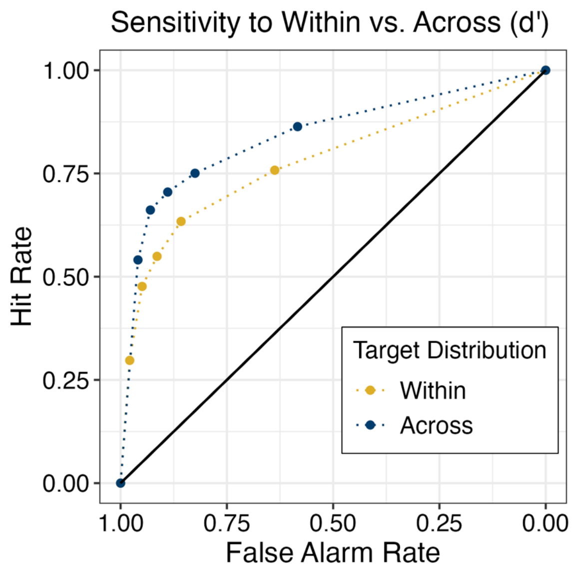

Figure 4.10: Receiver operating characteristic curves for Within and Across conditions, scaled against Mismatch conditions.

that suggests that not only do clause boundaries delineate boundaries of encoding contexts, but that they also provide a meaningful relationship between each noun and its context. One could ask, what specifically drives this connection? What is sufficient to provide such a relationship? These are empirical questions that we hope to test in future experiments. One possibility is that argument structure may help provide this meaningful link, by tying nouns to the broader event within the clause. We could then test differential memory for arguments and adjuncts. In a recall task, Chromỳ and Vojvodić (2024) found better recall for arguments than adjuncts, but it would be informative to expand this to a recognition memory task, explicitly testing temporal order memory.

Further testing would also be beneficial to generalize across a few of the idiosyncrasies of these materials in particular. For example, these stimuli always introduced the within-clause nouns as conjoined subjects. It would be useful to test temporal order memory of nouns within and outside of conjoined phrases (but still within the same clause) to understand what level of structural grouping may be driving the effects seen in this experiment. Furthermore, we elected to use cataphoric verb phrase ellipsis in the first phrase to avoid early effects of verb-specific details, and to ensure that the clauses would not differ drastically in terms of the events they are depicting. One remaining point, however, is understanding the events that we introduced via these stimuli, namely whether they were interpreted collectively or distributively. Similar to the role of syntactic structure, we might imagine that within a collective event, temporal order memory would be worse than within a distributed event. Across our stimuli, some events could plausibly read as collective events (e.g. *x and y raised some money*, where x and y could jointly raise the money or individually do so), whereas others seem more biased to distributed events (e.g. *x and y read the book*, where x and y each individually read the book). Future iterations could explicitly manipulate the types of events included to see if this influenced the temporal order memory of the event participants.

One final potential confound to consider is that the Within/Across manipulation also alters whether the Mismatch condition changes the truth conditions of the sentence. This is especially exaggerated with a subordinator like *before* in the particular sentence frames used: the sentence in (26a) is true only in cases wherein the researcher did solve the problem, but the scientist did not, while the opposite is the case in the corresponding Mismatch probe sentence (26b). This change in truth conditions does not occur in the case of the Within conditions for this item.

(26) Sample Across condition

a. Exposure: Before the butcher and *the scientist* could, *the researcher* and the landlord solved the problem.

b. Mismatch: Before the butcher and *the researcher* could, *the scientist* and the landlord solved the problem.

This experiment used a number of subordinators that differed in how large the resulting shift in truth conditions was between the exposure sentence and Mismatched probe sentence (*just like, just as, whenever, right after*). The most drastic shift in truth conditions such that a person did participate in an action in the exposure sentence and did not in the Mismatch probe was limited to stimuli that used the subordinator *before* (10 items). Changes in temporal ordering of events between the exposure and probe sentence were limited to stimuli that used the subordinator *right after* (9 items). For the other three subordinators (29 items), the Across Mismatch conditions did not result in such drastic changes in participation in events or temporal ordering of events. Therefore, our results are likely not only a result of sensitivity to such drastic changes in truth conditions, but future analyses or future iterations of this study could more carefully ensure that the truth conditions across the exposure and probe sentences in the Mismatch conditions were held constant.

4.6 Discussion

Chapter 4 sought to concretize the notion of contiguity and encoding context within linguistic contexts. Drawing insight from the Temporal Context Model (Howard & Kahana, 2002), we hypothesized that linguistic structure (whether syntactic or prosodic) could influence the encoding of a contextual feature. We entertained two hypotheses, a hypothesis of Linguistic Sensitivity, in which this was the case, and a hypothesis of Temporal Contiguity, where any influence of linguistic boundaries is overshadowed by the role of temporal proximity. Across a recall study and recognition study, we found that participants were sensitive to linguistic structure: reactivation of one item only benefited the recall of other nouns in that clause, intrusions during recall respected syntactic position rather than contiguity (e.g. subjects intruding into other subject positions, despite being further away), and there was greater sensitivity to changes made across boundaries rather than within. Overall, while there is more work to be done to expand on these particular studies, we argue that this is evidence that linguistic structure does play a role in the encoding of contexts in memory.

Mechanistically, we argue that effects of encoding contexts come about as a result of contextual drift between items. Some degree of contextual drift occurs between any two item encodings. However, we argue that the transition between syntactic or prosodic groupings triggers a greater shift in context, such that two items encoded within the same clause will be more featurally similar than if those same two items had been encoded in different clauses, even beyond what would be predicted on the basis of typical contextual drift between items.

There are many avenues to explore in terms of contextual features during language processing. For example, we might wonder whether context encodings could facilitate the resolution of anaphora, or help guide the parser back to a point of misanalysis.

One very pertinent line of future research is to replicate these findings with different forms of boundaries in mind. We hope to explore the role of prosodic boundaries by using auditory stimuli. This could allow us to address, for example, the question of coordinated subjects in Experiment 5, by prosodically grouping mentioned nouns differently. Furthermore, the concept of a prosodic boundary is not one single acoustic cue, often indicated by a combination of a pause, phrase final lengthening, downstepping, and/or declination. While the length of a prosodic break is not necessarily consistent across boundaries (Calhoun, 2010), some evidence suggests that listeners can distinguish between different types of subordinate clauses based on the quality of a prosodic break (Lelandais & Ferré, 2018). Therefore, careful follow-up studies could manipulate the acoustic properties of a break, and observe the degree to which these influenced the patterns of recall and recognition observed here.

Chapter 5

Availability of features

At this point, we've seen evidence for interference on the basis of semantic similarity, and that this is enhanced by contextual similarity and overlapping activation. We've also seen preliminary evidence for how linguistic boundaries may shape the encoding of context.

This chapter will address the filled gap effect as an instance of overlapping activation. I'll repeat the example of a filled gap site from (10b) from Stowe (1986) here in (27). At *who*, an active search for a gap position begins, and a penalty emerges when another item is found in an early, grammatically plausible gap position (in this case, *us*).

(27) My brother wanted to know who_i Ruth will bring **us** home to $\_{i}$ at Christmas.

What processes comprise the filled gap effect? Evidence for penalties on the basis of plausibility of the wh-filler as an object of the verb suggest that some degree of semantic interpretation is attempted as part of the gap postulation process. For example, using stimuli such as the sentences in (28), Traxler and Pickering (1996) found that readers slowed down at the verb (*wrote*) if the filler was an implausible object (*city*) compared to conditions where the filler was a plausible object (*book*).

(28) We like the city/book that the author wrote unceasingly and with great dedication about _ while waiting for a contract.

The plausibility effect in filler-gap dependency processing suggests that some form of semantic or thematic information must be available at the postulated gap site, either because it has been maintained for the duration of the dependency, or because it has been reactivated at the postulated gap site. As discussed in Chapter 2, Ness and Meltzer-Asscher (2019) put forth an argument that the features that are maintained are those which the parser uses to generate expectations about the position of upcoming gap sites. The experiments in this chapter contribute to this conversation in two ways. Experiment 6 manipulates semantic similarity between a wh-filler and the item found in a projected gap site, and then looks at later reactivation of the wh-filler to examine whether semantic similarity-based interference occurred during the gap filling process. Experiments 7-9 explore whether encoded thematic roles can be used to predictively mitigate a filled gap effect. At the highest level, this chapter seeks to further understand what features are maintained, and used, to predictively postulate gap sites; and to what degree semantic features may create interference, if they are maintained.

5.1 Experiment 6: Filler gap encoding

One aspect of dependency resolution that was not explicitly explored in Experiments 1-3 is the fact that encountering another word in a projected gap site can be considered an instance of contiguous lexical activation. The Near conditions of those experiments treated the fact that the two nouns were encountered in close proximity as contiguity, but they were also instances in which the parser may have attempted to associate the subject position within the RC with the RC head, particularly in the case of animate head nouns. It's not possible, therefore, to disentangle the nearby presentation of these nouns from potential co-activation at the early subject position. One of the purposes of this experiment was to provide an explicit comparison between instances of dependency resolution which may involve co-activation at the same position, and instances without such a dependency.

In thinking further about co-activation at filled gap site, mechanistically, we can understand that several things are occurring: i) a moment of overlapping activation, in which the features of the filler word are active while lexical access of the encountered word occurs, ii) fast, subsequent attempts to bind different words (bundles of features) to the same position. Across the theories of memories discussed in this dissertation, this pattern of activation would be predicted to create interference. Here I will revisit three of the theories addressed in Section 2, which I believe make the clearest predictions about simultaneous activation: feature overwriting, superposition, and the Temporal Context Model.

In terms of Feature Change accounts, in which the featural representation itself is disrupted or changed by the presence of a similar intervener, different patterns may be expected. Feature overwriting, in which contiguity in encoding can result in a shared feature only being represented in one of the two items, would predict that in the case of more similar nouns, there would be a larger degree of feature loss. Increased feature loss would then hinder later retrieval, and potentially later interpretation during comprehension questions. On the other hand, the process of superposition, in which similar feature bindings strengthen one another, would predict facilitatory interference for highly similar interveners, which might result in faster reading times for conditions with highly similar interveners. However, this speed-up in reading would likely not be accompanied by higher accuracy for highly similar interveners, given the fact that superposition is predicted to result in greater intrusion errors. In other words, facilitatory interference by superposition may serve to facilitate the retrieval process later on by virtue of strengthening the availability of relevant features, but may be more likely to result in unnoticed misretrievals, which would then be reflected in lower question accuracy.

In terms of Feature Interaction accounts, and specifically thinking about the encoding of contextual features as per the Temporal Context Model, it is worth considering what the final state of the contextual feature for a reactivated filler must be. The Multitrace Distributed Memory Model (Kahana, 2012) suggests that each encoded presentation of an item would leave its own memory trace, and that during subsequent retrievals of an item, it is re-encoded in memory.

To assess whether semantic similarity based interference can be found as a consequence of a filled gap, we employed a design similar to Kim et al. (2020), repeated in (29), looking for a number-mismatch penalty at a second, later gap site, following the completion of the dependency. Kim et al. (2020) found that this effect was weaker in instances of reactivation, compared to the first gap site that resolves the dependency, but that it is still detectable.

(29) Ungrammatical-SG conditions of Experiment 3 in Kim et al. (2020)

a. Which mistake$_{SG}$ in the program that will be disastrous for the company $\_{SG}$ are$_{PL}$ harmful for everyone involved.

b. Which mistake$_{SG}$ in the program $\_{SG}$ will be disastrous for the company and $\_{SG}$ are$_{PL}$ harmful for everyone involved.

Kim et al. (2020) looked for interference on the basis of number, which would be explicitly used as a retrieval cue at the verb in order to establish subject-verb number agreement. To understand the role of encoding interference in wh-filler gap dependencies, we took inspiration from the findings in Smith et al. (2021), which demonstrate a greater penalty for number-mismatches in instances of higher semantic similarity, e.g. (30d) compared to (30b).

(30) a. The canoe by the cabin likely was damaged in the heavy storm.
b. The canoe by the cabins likely was damaged in the heavy storm.
c. The canoe by the kayak likely was damaged in the heavy storm.
d. The canoe by the kayaks likely was damaged in the heavy storm.

Here, semantic similarity increases the chances of erroneously attempting to agree with N2 (cabin, kayak), despite the fact that the retrieval process at the later verb (*was*) would not reference any semantic features. The authors describe this finding in terms of the Self-Organized Sentence Processing model (Smith et al., 2018; Smith, 2018).

It is also potentially possible to interpret these results within the framework of conflicting bindings (Logačev & Vasishth, 2011), where partial matches can give rise to greater interference than non-matching items, or fully matching items, as a result of interference from different item-item bindings. In this case, this would look like interference between the two bindings BOAT-SG and BOAT-PL, where we could understand BOAT as a stand-in for a high dimensional vector that represents boat-like properties.

The study described below expands on this, looking at the same binding of semantic category information and grammatical number, to see if this interference is heightened by co-activation during dependency resolution.

Similarity	**word2vec**	**Length difference**
High	0.57 (0.01)	1.18 (0.11)
Low	0.18 (0.01)	1.08 (0.13)

Table 5.1: Characteristics of similarity and length of the interveners, in comparison to the filler, Mean (SE).

5.1.1 Methods

Materials

To examine the effects of similarity during this reactivation process, sentences with and without d-linked wh-fillers were constructed. The experimental stimuli consisted of 60 items that crossed the position of the Gap in the first clause (pre-intervener, post-intervener, none) as well as the Similarity between the two nouns (High, Low). The No gap condition instead contained a *whether*-clause that contained the two nouns as clause mates. An example itemset can be found in Table 5.2.

The nouns used in this study were adopted from the materials used in Smith et al. (2021). Modifications were made to control for length and word2vec similarity. The maximum length difference between the filler and intervener was 3 characters; the average difference was 1.15 characters. Smith et al. (2021) included 40 items, and so 20 additional noun pairs were created as well. In addition to checking word2vec similarity, the high similarity nouns always shared a category membership (e.g. *fork* and *spoon* as silverware) that the low similarity noun did not, even if the low similarity noun may be found in similar contexts (e.g. *pot* as an item in a kitchen that is not silverware). All nouns were inanimate. Item characteristics (similarity and length) can be found in Table 5.1.

Similarly to Experiments 1-3, the other referents, especially in subject position, were of meaningfully different categories to avoid the possibility for interference

(Gordon et al., 2001). The matrix subject was always a name, and the embedded subject was always a local pronoun or indefinite expression (*I*, *you*, *someone*, *somebody*).

There were two regions of critical interest in this study. Firstly, the filled gap region (*the kayaks/carts*), in conditions (a-b) and (e-f), would allow us to establish a baseline check that we were successfully able to replicate a well-known penalty in the literature. While we could test for a difference in semantic similarity, there is a concern that in the (a-b) conditions, readers would have already encountered the other noun, whereas in the (e-f) conditions, they would not have. Secondly, the second VP conjunct *was too old* would necessitate retrieval at a gap site. We therefore treat this as a critical region to find effects of similarity and/or overlapping activation during retrieval, with the assumption that difficulty during retrieval here could reflect earlier interference during the initial dependency resolution process. In all items, this region consisted of the singular *was* followed by an intensifier, adverb, or negation, and an adjective.

To maintain acceptability in the No gap conditions, the material leading up to the critical region had to be changed. In the No gap conditions, an explicit anaphor is given and there is no coordination (see Table 5.2).

	Preceding context
	Nora knew...
Gap	**First conjunct**
Pre-intervener	which canoe you placed _ next to the {kayaks | carts }
Post-intervener	which canoe you placed the {kayaks | carts } next to _
No gap	whether you placed the {kayaks | carts } next to the canoe
	Continuation
Pre/Post-int.	and then decided _ was too old for the camp by the lake.
No gap	after deciding it was too old for the camp by the lake.

Table 5.2: Sample item set illustrating the Gap x Similarity manipulation. The Similarity manipulation is demonstrated in-line in brackets, with the High similarity noun first followed by the Low similarity noun.

Participants

A total of 66 participants were recruited on Prolific. Participants scoring under 70% accuracy on a set of simple fillers were excluded from analysis and replaced according to counterbalancing list. The final dataset contains 60 counterbalanced participants. All participants reported having English as a first language, having completed at least a high school level of education, and not having experienced literacy difficulties such as dyslexia or ADHD. Participants were compensated $12 for their participation.

Procedure

Experiment 6 following the same experiment hosting procedure as used in previous online experiments.

During the experiment itself, participants first read a sentence by pressing a space bar in a moving-window self-paced reading display. Following the sentence, participants were prompted to respond to a question about the sentence. For the experimental stimuli, this question always probed the reactivated subject of the second conjunct embedded verb (*was too old*). For each question, two options were given, the filler (e.g. *canoe*) and the intervener (e.g. *carts/kayaks*). The presentation order of responses was randomized. The correct response should always be the filler, and this is made unambiguous by number agreement. As an example, the question in (31) would correspond to the High similarity conditions in the example item in Table 5.2.

(31) Q: What was too old?
A: a canoe the kayaks

To obscure the patterns of experimental stimuli questions referring to the second conjunct and answers always corresponding to the singular option, the experimental stimuli were presented with 60 filler sentences created to intentionally obscure the design. All responses to filler sentences were plural, so that the grammatical number of the correct response was evenly distributed across the whole experiment. Furthermore, roughly half of the filler questions probed the first half of the sentence. Roughly one third of the filler sentences closely replicated the structure of the experimental stimuli, so that just upon seeing the structure of the experimental stimuli, a participant would not be able to predict that the question would probe the second half of the sentence or guess that the answer would be singular in number.

5.1.2 Data processing

As mentioned in the description of participants, there was a set of structurally simpler filler items. These were treated as attention checks, and correspondingly errors in these items were treated as lapses in attention and not theoretically interesting. They were therefore used to set an exclusion threshold (70% accuracy on these simpler fillers). For the final analysis reading times over 2000 ms or below

100 ms were excluded. The same modeling procedure was used as in the previous self-paced reading studies in this dissertation.

As with the other self-paced reading studies in this dissertation, a reading time analysis was conducted on Log-transformed reading times, using correct trials. However, unlike the other experiments, we reasoned that trials with incorrect responses might not reflect uninformative lapses in attention, but rather a consequence of interference that we predicted could occur. We therefore performed a secondary analysis on all trials, testing a three-way interaction between Gap, Similarity, and Accuracy. The contrasts were sum-coded, such that for Similarity, High was .5 and Low was -.5. For the filled gap effect analysis, the Pre-intervener conditions were excluded, and therefore the Gap factor was sum-coded as well, with Post-intervener .5 and None -.5. Finally, during the critical region at and following the reactivated gap site, the Gap factor was Helmert coded, first comparing the Pre-intervener condition to the No gap condition, and then comparing the Post-intervener condition to the No gap baseline.

Two regions of interest were examined in the analysis. First, we looked to establish a filled-gap effect, looking for reading time differences between the Post-intervener conditions and the None conditions beginning at the intervener *the kayaks/carts* and continuing into the prepositional phrase *next to*. For this analysis, the data were subset to include only Post-intervener and None conditions. Secondly, to investigate downstream effects of overlapping activation during a filled gap effect, we looked at reading times in the second conjunct, beginning with the position directly following the second gap site (Pre-intervener and Post-intervener conditions) or anaphor (No gap condition), at *was* and the following two words.

5.1.3 Predictions

The purpose of the Gap manipulation was to manipulate the pattern of activation for the filler and the intervener. In the Pre-intervener condition, the gap is encountered prior to the intervener *kayaks/carts*. We therefore predict that the representation for the filler *canoe* would begin to decay by the time the intervener is reached. In contrast, the gap in the Post-intervener condition is encountered following the intervener. We therefore predict two sources of overlapping activation: the moment of overlapping activation in the direct object position following the verb *placed*, and the continued activation for the filler following this before finally encountering the true gap site later on. In contrast, the No gap condition includes both the filler word and intervener from the other two conditions, but no active dependency. Following the role of contiguity in interference as given by the models of encoding interference and memory outlined previously, we predict that any detectable interference at a later gap site should be greatest in the Post-intervener condition, resulting in longer reading times at the critical region.

The verbal structures used in the first conjunct always included secondary predication. We considered whether this may influence the rate of decay of the first noun. For example, it could be possible that the representation of *canoe* would be maintained longer in (32a) than in (32b).

(32) a. I saw the canoe on the dock.

b. I saw the canoe on Tuesday.

However, we do not believe this disrupts the general shape of our predictions, as even if the verbal structures used in this experiment do encourage longer activation for their arguments, we believe the relative activation levels would still be different depending on the presence of an active search to resolve a dependency.

With respect to the Similarity manipulation, in this study, similar to Smith et al. (2021), in the High similarity conditions there must be a binding of BOAT-like features to different number features. The Low similarity items do still involve some partial matching, e.g. they are both inanimate and yet differ in number, but will overall have less similar featural representations than the High similarity interveners. We therefore predict the most interference in the High similarity conditions, when there is a greater degree of overlapping activation during the initial dependency resolution process.

5.1.4 Question accuracy

The question accuracy means can be found in Table 5.3. All main effects and interactions reached significance. There was a main effect of Similarity, such that responses to High Similarity trials were less accurate than Low Similarity trials, z = -22.11, $p < .001$. There was a main effect of Dependency, such that participants were most accurate in conditions that contained a filler-gap dependency, z = 78.70, $p < .001$, as well as a main effect of Overlap, such that the Pre-intervener (non-overlapping) conditions were more accurate that the Post-intervener (overlapping) conditions, z = -27.02, p <.001. Finally, both interactions were significant. There was a larger difference in accuracy between the Similarity conditions if there wasn't a Dependency, compared to the conditions where there was, z = 4.98, $p < .001$. Between the two conditions that did contain a dependency, the difference in accuracy was greater in the Post-intervener (Overlapping) conditions than in the Pre-intervener conditions, z = -6.51, $p < .001$.

Gap	Similarity	Mean	SE
Pre-intervener	*High*	0.79	0.004
Pre-intervener	*Low*	0.81	0.003
Post-intervener	*High*	0.66	0.004
Post-intervener	*Low*	0.74	0.004
No gap	*High*	0.43	0.004
No gap	*Low*	0.53	0.004

Table 5.3: Question accuracy for Experiment 6, means (M) and standard error (SE).

5.1.5 Question response time

Response times to questions were also analyzed, and can be seen in Figure 5.1. Analyses both of correct trials only and all trials with accuracy included in the model are presented.

Correct trials only

There was a significant main effect of Similarity, such that High similarity conditions (M = 3769 ms, SE = 28 ms) resulted in longer response times than Low similarity conditions (M = 3472 ms, SE = 27 ms), t = 2.13, $p < .05$. There was a main effect of Dependency, such that conditions with a dependency (M = 3518 ms, SE = 23 ms) were overall responded to faster than those without (M = 3897, SE =36), t = -4.27, $p < .001$. Between the two types of dependencies, overlapping dependency conditions (Post-intervener; M = 3762 ms, SE = 36 ms) were read slower than non-overlapping dependency conditions (Pre-intervener; M = 3305 ms, SE = 29 ms), t = 2.38, $p < .05$. There was a marginal interaction between Similarity and Dependency, t = -1.87, p = .06, and a significant interaction between Similarity and Overlapping dependency, where there was a larger difference between Pre- and Post- intervener conditions for High similarity conditions (difference: 568 ms) than Low similarity conditions (difference: 367 ms), t

= 2.14, p < .05.

Analysis with incorrect trails

There was a main effect of Dependency, such that participants were overall faster when there was a dependency (Pre- or Post-intervener gap; M = 3637 ms, SE = 27 ms) versus No Gap (M = 3831 ms, SE = 21 ms), t = -12.60, p < .001. There was also a main effect of Overlapping activation, such that the overlapping dependency conditions (Post-intervener gap; M = 3925 ms, SE = 32 ms) yielded longer response times than the non-overlapping dependency conditions (Pre-intervener gap; M = 3348 ms, SE = 22 ms), t = 14.88, p < .001. Furthermore, there was an effect of question accuracy, such that ultimately incorrect responses (M = 3907 ms, SE = 25 ms) were slower than correct responses (M = 3593 ms, SE = 18 ms), t = 28.48, p < .001.

These main effects were qualified by a number of interactions. There was an interaction between Similarity and Dependency, where the difference between the Low similarity conditions (difference: 230 ms) was greater than the High similarity conditions (difference: 159 ms), t = -6.12, p < .001. There was also an interaction between Similarity and presence of an Overlapping dependency, t = 2.79, p < .01. There was an interaction between Similarity and accuracy, where inaccurate responses showed a larger effect of similarity (difference: 593 ms) than correct responses (difference: 49 ms), t = -16.91, p < .001. Finally there were interactions between accuracy and presence of a Dependency, where conditions with a dependency were faster than the no dependency condition when the response was ultimately correct (difference: 314 ms), but slower when the answer was ultimately incorrect (difference: 146 ms), t = 25.24, p < .001, as well as between accuracy and whether a dependency was Overlapping, where there

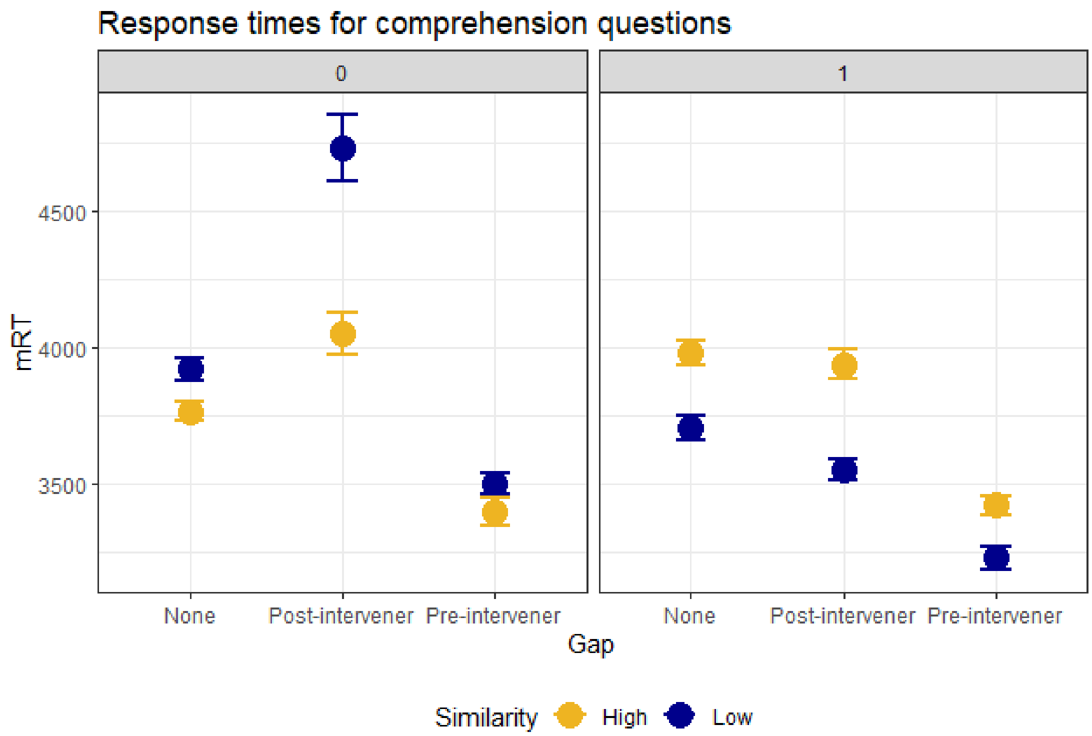

Figure 5.1: Response times to comprehension questions targeting the retrieved subject of the continuation (_ *was too old*).

was a larger response time penalty for incorrect responses in the Post-intervener conditions (difference: 609 ms) than the Pre-intervener conditions (difference: 124 ms), t = -2.13, $p < .05$.

5.1.6 Reading time results: Correct trials only

The filled gap effect

Reading times across the filled gap regions were summed (see Figure 5.2). There was a main effect of dependency, such that the dependency condition (M = 1615 ms, SE = 22 ms) was read slower than the no-dependency condition (M = 1508 ms, SE = 23 ms), t = 2.68, $p < .05$. While there was a numerical difference in the size of the filled gap for High similarity interveners (105 ms) compared to

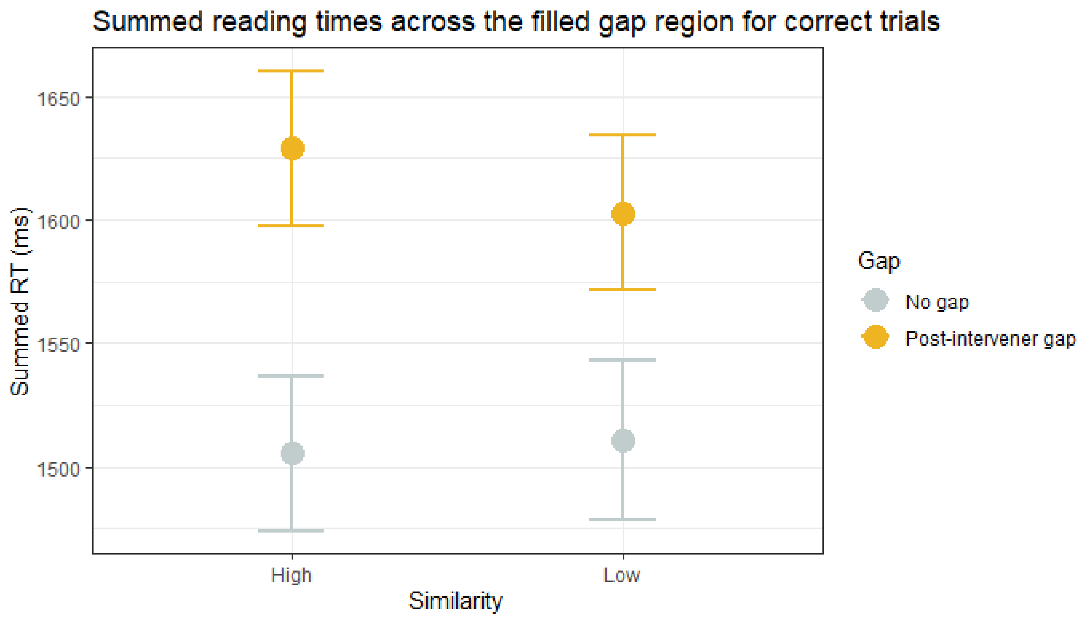

Figure 5.2: Summed reading times across the filled gap region *the kayaks/carts next to*, comparing the Post-intervener conditions to the No gap conditions.

Low similarity interveners (79 ms), no effect of Similarity reached significance.

Reactivation at a later gap

Reading times following the second gap site (see Figure 5.3) were analyzed.

At the verb (*was*) following the second gap (in the Pre- and Post-intervener conditions) or the anaphor (in the No gap conditions), there was a significant penalty for Post-intervener condition (M = 372 ms, SE = 7 ms) compared to the No gap baseline (M = 351 ms, SE = 7 ms), $t = 2.06$, $p < .05$, and no such penalty for the Pre-intervener condition compared to the baseline.

In the next region (*too*), the reverse was true: there was a penalty for the Pre-intervener condition (M = 359 ms, SE = 6 ms) compared to the baseline (M = 334 ms, SE = 6 ms), $t = 3.83$, $p < .001$, while the difference between the baseline and the Post-intervener conditions did not reach significance.

In the following region (*old*), there was a marginal interaction between the

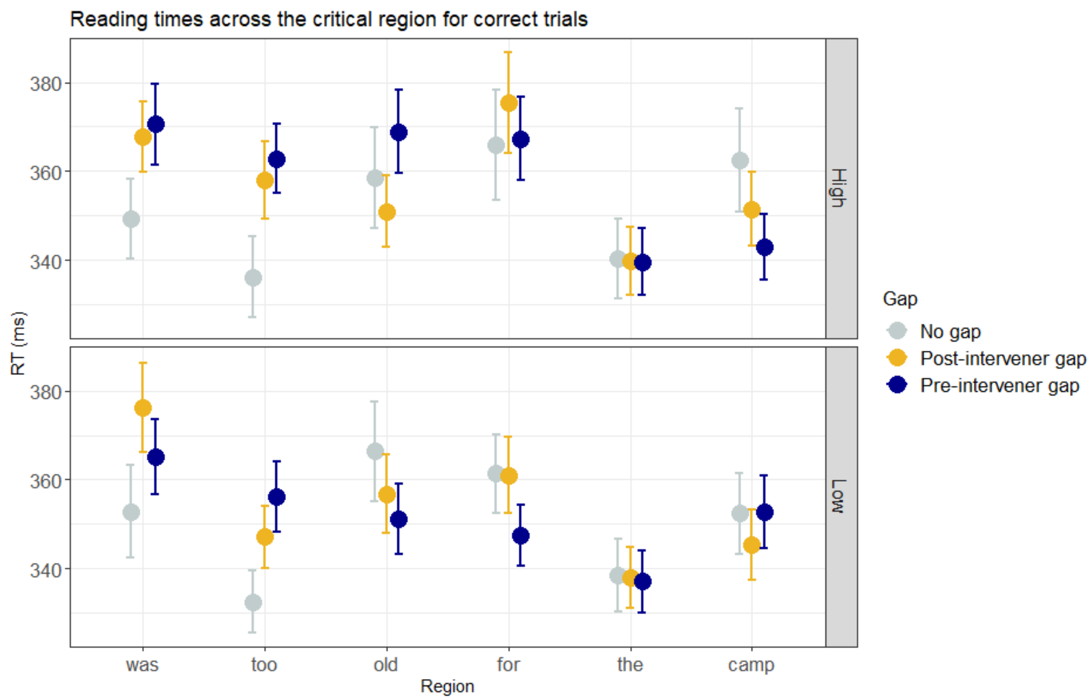

Figure 5.3: Reading times (ms) for the critical region following later reactivation, *was too old* and spillover *for the camp*, correct trials only.

presence of a Pre-intervener gap, and Similarity, $t = 1.80$, $p = .073$, such that High Similarity conditions were read slower than Low Similarity conditions with a non-overlapping dependency (Pre-intervener; difference: 18 ms), while the difference was smaller and in the opposite direction in the No gap conditions (difference: 8 ms).

5.1.7 Analysis with incorrect trials

Unlike the experiments in Chapter 3, all of the comprehension questions for the experimental stimuli in this experiment probed the interpretation at the second gap site. Incorrect responses are therefore especially theoretically interesting, as they indicate potential errors in the proper encoding of the plural feature.

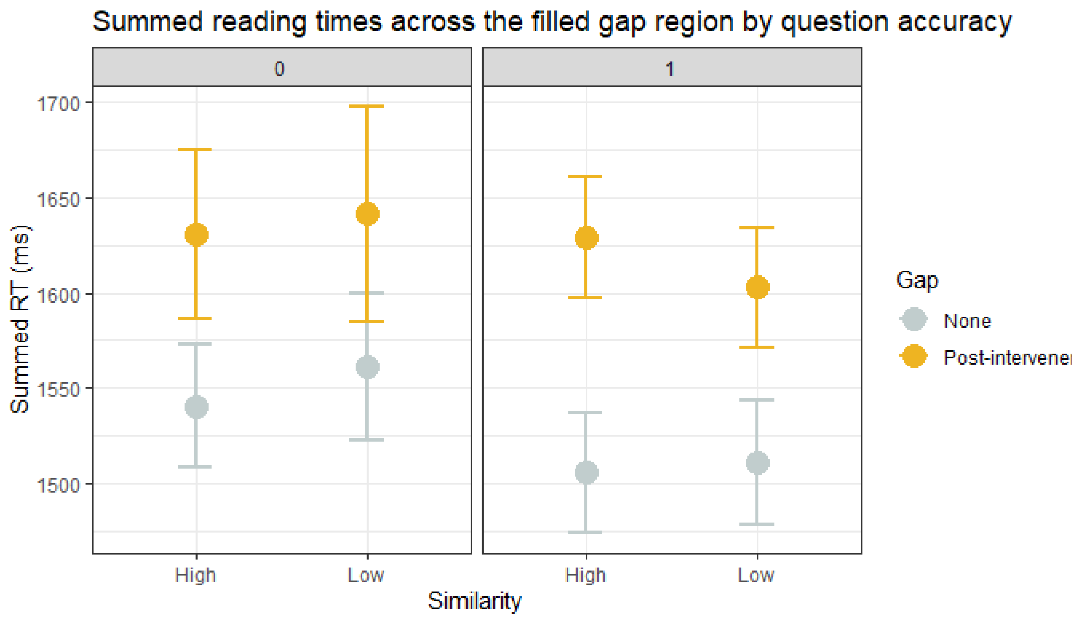

Figure 5.4: Summed reading times across the filled gap region *the kayaks/carts next to*, comparing the Post-intervener conditions to the No gap conditions, faceted by incorrect (0) vs correct (1) responses.

The filled gap effect

In the analysis including incorrect trials, the basic filled gap penalty remained, where there was a main effect of dependency, $t = 5.14$, $p < .001$. There was also a main effect of question accuracy, such that participants who were ultimately incorrect in identifying the subject of the later verb incurred a greater penalty than those who were ultimately correct, $t = 2.6$, $p < .05$. See Figure 5.4.

Reactivation at a later gap

At the verb (*was*), the same penalty for the Post-intervener condition over the No gap baseline emerged, $t = 3.83$, $p < .001$. There was also a main effect of accuracy, such that incorrect trials were read slower, $t = 2.27$, $p < .05$.

Following the verb (*too*), just as with the correct trials only analysis, there was a penalty for the Pre-intervener conditions compared to the baseline, $t =$

3.31, $p < .001$. Including the incorrect trials, a penalty for the Post-intervener conditions (M = 359 ms, SE = 5 ms) compared to the baseline (M = 340 ms, SE = 4 ms) also reached significance, $t = 2.25$, $p < .05$. As with previous regions, there was also a main effect of accuracy, such that inaccurate trials were read slower (difference: 11 ms), $t = 3.97$, $p < .001$. Here, we also found interactions with Similarity. There was a marginal interaction between Similarity and cases of overlapping activation (Post-intervener gaps), such that there was a greater penalty for the High Similarity conditions in the case of overlapping activation (Post-intervener; 8 ms difference) than for the baseline (1 ms difference), $t = 1.67$, $p = 0.96$. There was also a significant interaction between Similarity and accuracy, such that High Similarity conditions were read slower in correct trials (8 ms difference) and faster in the incorrect trials (15 ms difference), $t = -2.19$, $p < .05$. This interaction between Similarity and accuracy can be seen in Figure 5.5. And finally, there was a significant three-way interaction between Similarity, non-overlapping activation, and accuracy, $t = -1.94$, $p < .05$.

At the next region (*old*), there was a marginal penalty for incorrect trials, $t = 1.90$, $p = .057$. There was also a marginal three-way interaction between Similarity, non-overlapping interaction, and accuracy, $t = -1.72$, $p = .086$.

5.1.8 Discussion

Several generalizations can be made about the results in Experiment 6. Overall, we found several instances where Similarity influenced either the accuracy of the representation of the filler, or reading times when reaching the site of reactivation. We found that High similarity interveners resulted in lower accuracy and longer latencies in recalling the identity of the wh-filler during post-sentence comprehension questions, and furthermore that between the two dependency con-

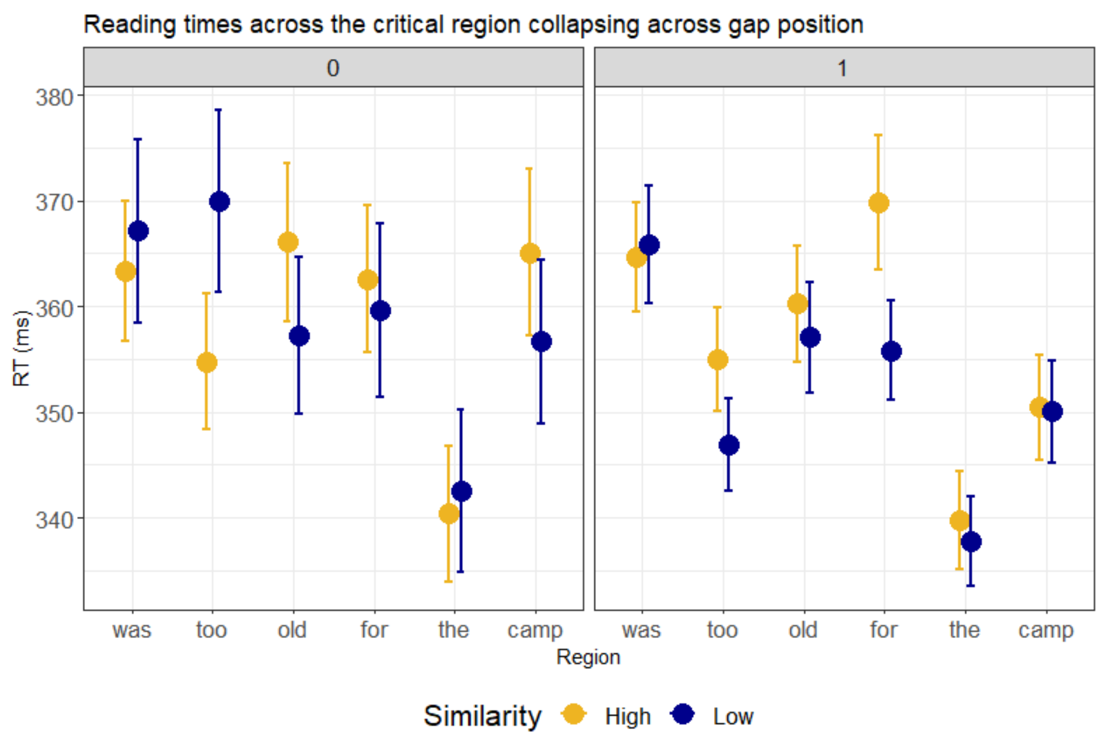

Figure 5.5: Reading times (ms) for the critical region following later reactivation, collapsing across gap position, faceted by incorrect (0) vs correct (1) responses.

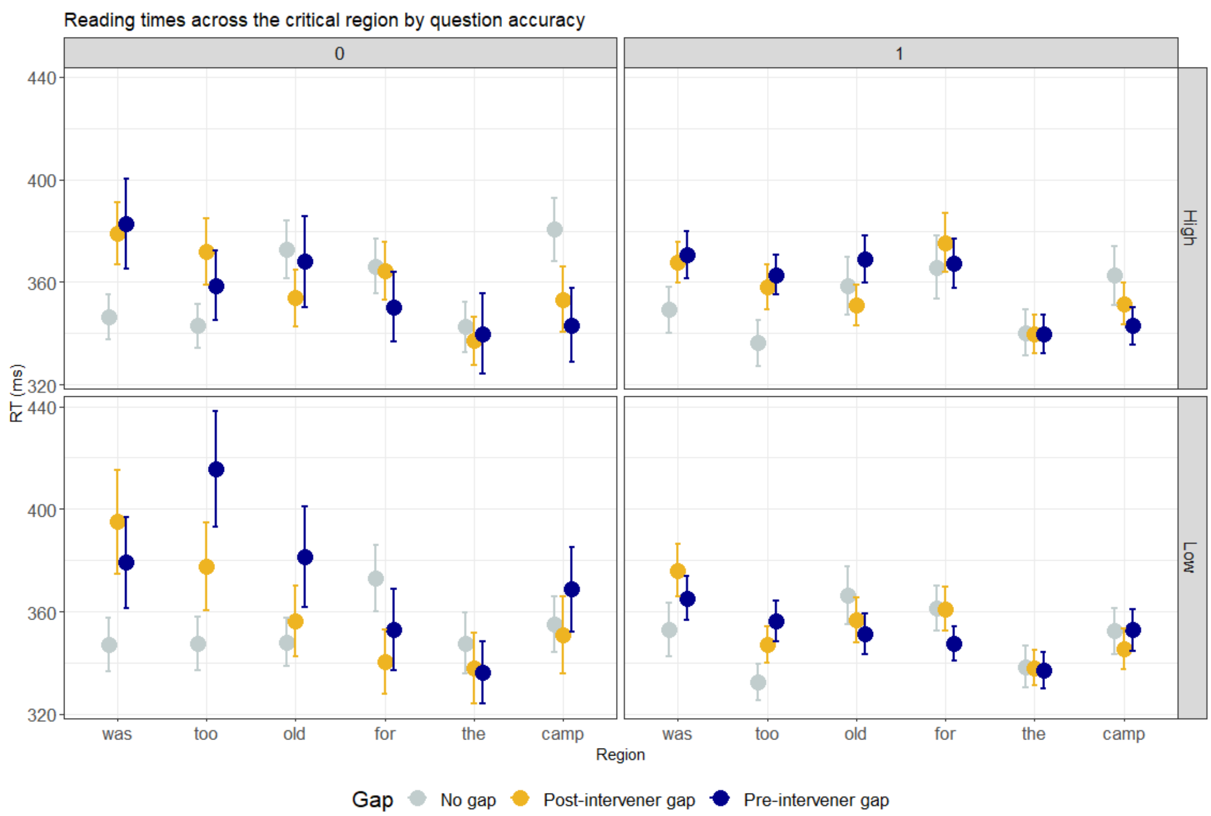

Figure 5.6: Reading times (ms) for the critical region following later reactivation, *was too old* and spillover *for the camp*, faceted by incorrect (0) vs correct (1) responses.

ditions, accuracy was most impacted by Similarity in a case of overlapping activation, or the Post-intervener conditions. During reading, we found that ultimately correct interpretations of the sentence corresponded to instances of slower reading times for High similarity interveners, for example in the region immediately following the verb *was*.

A few other patterns are worth noting. Firstly, there is one instance of a penalty for the Pre-intervener gap position compared to the No gap conditions, seen in correct trials in the region following the verb *was*, and more exaggeratedly in incorrect trials. This seems in line with classic findings in agreement attraction literature, where an N2 that does not agree with the verb may interfere. The Pre-intervener condition is precisely a configuration (SG...PL...) that we may expect to produce agreement attraction errors in production.

Secondly, question accuracy for the No gap conditions was very low, and also reflected an increased penalty for High similarity interveners. This again could be seen as evidence for similarity-based interference on the basis of presence of semantically similar competitors. It is possible that participants were more likely to refer back to a more prominent position, one of the necessary arguments of the verb. It is not clear whether in this case participants are simply choosing to ignore a correctly encoded number mismatch, or whether some degree of number attraction occurred.

One open question is that of the active maintenance of semantic features, such as the features that would correspond to category membership, e.g. attributes of BOAT-hood. No effects of Similarity reached significance in the filled gap region, despite numerical trends of increased reading times for the High similarity conditions. It may be that more power was needed to detect this effect. An early effect may have suggested that these features were in fact maintained. As it stands, the presence of later interference on the basis of semantic similarity, especially in cases of overlapping activation, suggests that there was some enhanced interference on the basis of this overlap. Even if these features are not actively maintained, and are simply reactivated at the site of a postulated gap, here we see evidence that the overlapping activation does in fact trigger later interference.

In considering the underlying mechanism of this interference, it is possible that this pattern of results could differentiate between the different Feature Change accounts. Of the Feature Change accounts, feature overwriting and superposition, disruptions to the featural encodings may be expected to influence the activation levels of the affected items in different ways. In the case of two highly similar items undergoing feature overwriting, any feature shared between the two items may only be encoded on one or the other item, resulting in lower overall activation for affected items. In contrast, in an event of superposition, the association of the shared features to that position would be strengthened. We therefore may expect faster recall in instances of superposition. However, in thinking of accuracy, it may be the case the superposition would result in a greater likelihood of misretrievals. We did not see that overall incorrect responses corresponded to faster reading times, and nor did High similarity interveners. This suggests that the type of interference experienced during this form of overlap is not in line with the account of superposition, which would predict facilitatory interference in the case of higher proportions of shared features, potentially coupled with increased errors in recall. Therefore, if this interference is a result of interference in the actual encoding of features, it is most likely due to feature overwriting, resulting in an overall degraded featural representation of the two nouns.

There is one pattern of findings from the comprehension question responses that may warrant further examination: in incorrect responses, particularly in the case of an overlapping dependency, we see that the Low similarity condition results in longer response times. One potential interpretation of this is that an increased overlap of features results in faster errors, a pattern that is compatible with superposition. Future work looking more closely at post-sentence judgments may be beneficial in teasing apart possible explanations for encoding. For example, in future work, we could more explicitly probe participants' judgments about words they had seen, e.g. a probe recognition study with confidence ratings.

And in what contexts should we look for superposition? In the case of this experimental context, we imagined the filled gap site as a moment in naturalistic sentence processing where theories of superposition may become relevant, as two different items are attempted to be bound to a particular position in rapid succession. However, it should be noted that encountering another word in the filled gap site is an immediate cue that the filler must be integrated elsewhere. By the end of the sentence (or at least before the later retrieval site in the case of the experimental stimuli in Experiment 6), the filler has been integrated into a different position in the sentence. If we do not find effects that look like superposition in this experimental format, it does not necessarily mean that superposition cannot be observed in sentential processing. It may require a different experimental design, or it may only occur in certain configurations or instances of simultaneous activation.

This interference could also be explained by interactions between accurately encoded features. It's possible that this instance of overlapping activation would result in contextual features that would be more similar than items encoded even in quick succession. Or it's possible that, under an account such as conflicting bindings, that inhibitory interference would arise from the pairwise bindings between each BOAT attribute and each of the levels of grammatical number in

English.

Future work would be necessary to tease apart these different possibilities. To explicitly test how contextual features are assigned during moments of simultaneous activation, as in instances such as the filled gap effect, different methodologies could be used to probe the representation of these items in memory, by using a recognition or recall tasks, as in Chapter 4. This would allow us to see how overlapping activation, or at the very least an unpronounced site of reactivation, might influence well-established patterns in memory, such as the forward asymmetrical contiguity effect (Howard & Kahana, 2002; Healey et al., 2019). For example, could we see instances of backward asymmetrical contiguity effects? Considering only pronounced items, we might expect the recall of (33) to follow the pattern in (33a), as activating *figurines* would serve as a good cue for the following item, given similarity in contextual features. We would most likely expect, however, a re-encoding of the filler at the gap site, with a new contextual feature. This process could inspire a recall pattern more similar to (33b). Further manipulations would be needed to tease apart differences in recall on the basis of reactivation with a re-encoded contextual feature, and disruption in the contextual feature encoding as a result of overlapping activation.

(33) I saw which *magazine*$_1$ you placed the *figurines*$_2$ next to $\_{}_1$ on the *shelf*$_3$.
Recall the nouns in order:

a. After recalling *figurines*$_2$, most likely to recall *shelf*$_3$

b. After recalling *figurines*$_2$, most likely to recall *magazine*$_1$

The form of overlapping activation considered here is specifically in the attempted resolution of a dependency in the presence of another item. One property of this type of overlapping activation is that the representation of the filler is still mandatorily needed following the first erroneous dependency resolution attempt.

This means that the representation of the filler cannot be fully inhibited or suppressed, and may therefore exaggerate effects of overlapping activation. In the conclusion in Chapter 6, I will consider another form of overlapping activation: overlapping activation between a highly predictable word and an unexpected encountered word. In the case of prediction, we must also then discuss the role that inhibition may play.

To conclude, this experiment demonstrates interference on the basis of semantic similarity during active dependency resolution. The next experiments, too, look at what features may be active during wh-filler gap dependency processing, and whether they can be used predictively to avoid a filled gap effect.

5.2 Parallelism

Experiment 6 investigated the activation of semantic features during active dependency formation to test whether similarity along those dimensions could result in later interference. These features were not examined as possible information that could be used to direct the postulation of gaps, however.

In the following three experiments, we will turn to think about what encoded information, beyond features such as number and animacy, is reactivated when making structural predictions about a potential gap site. In particular, Experiments 7-9 leverage a parsing preference for parallel structures in coordination to investigate whether encoded thematic information may interact with predictions about a second conjunct.

The parsing preference for parallel structures in coordination has been found in a wide range of studies (e.g. Branigan, Pickering, Liversedge, Stewart, & Urbach, 1995; Frazier, Munn, & Clifton, 2000; Pickering & Ferreira, 2008; Sturt, Keller, & Dubey, 2010; Bock, 1986), exemplified in reading times as facilitation for sentences like (34a) compared to (34b).

(34) a. Hilda noticed a strange man and a tall woman ...

b. Hilda noticed a man and a tall woman ...

As it pertains to filler-gap dependency processing, this preference for parallelism is grammaticized in the Coordinate Structure Constraint, which observes that a filler cannot be extracted from only one of two coordinated phrases (Ross, 1967), such that (35) is grammatical while (36) is not. This kind of extraction from multiple phrases has been termed Across the Board (ATB) extraction, and studies of incremental language processing have found that comprehenders are sensitive to this grammatical constraint (Wagers & Phillips, 2009; Sturt & Martin, 2016; Parker, 2017).

(35) They dislike the poetry$_i$ that the New York Times reviews $\__i$ or publishes $\__i$.

(36) * They dislike the poetry$_i$ that the New York Times reviews $\__i$ or publishes interviews.

Parker (2017) has demonstrated that the preference for parallelism across conjuncts allows readers to avoid a filled gap effect at a filled direction object position (*the sterile beakers*), if the first conjunct contains prepositional object (PO) gap, as in (37b), compared to a baseline PP filler (37a). The typical filled gap effect emerges in the case where the first conjunct contains a direct object (DO) gap, and the second a PO gap.

(37) a. The chemicals [with which]$_{PP}$ the technician sprayed the equipment $\__{PP}$ and prepared the sterile beakers $\__{PP}$...

b. The chemicals which$_{DP}$ the technician sprayed the equipment with $\__{PO}$ and prepared the sterile beakers with $\__{PO}$...

c. The chemicals which$_{DP}$ the technician sprayed $\_{DO}$ and prepared the sterile beakers with $\_{PO}$...

(38) a. The chemicals which$_{DP}$ the technician sprayed the equipment with $\_{PO}$ and he prepared the beakers with $\_{PO}$...

b. The chemicals which$_{DP}$ the technician sprayed the equipment with $\_{PO}$ and which$_{DP}$ he prepared the beakers with $\_{PO}$...

This might be explained not as preference for parallelism, but as a strict adherence to a grammatical constraint during parsing. However, in a follow up study, they show that parallelism still facilitates PO gap processing even in non ATB configurations such as (38b), suggesting that parallelism on its own does play some role in easing the processing of (37b) compared to (37c).

The effects of parallelism in Parker (2017) could be explained mechanistically as a matter of structural priming, where a benefit for repeated structure can be understood as the structural representation being activated, and then reused. However, there is also the matter of the encoding of the initial filler, specifically in the case of (38b). Is the filler encoded with the same set of features as the first filler, which was encoded with very little prior context? Does a benefit from parallelism stem entirely from pre-activated structure (in this case the structural representation of a prepositional object gap), or in part from a change in how the second filler is encoded?

To present an example, this dissertation has already discussed a preference for animate RC heads to appear in subject positions and receive agentive roles. If prior context were able to suggest that the upcoming animate entity were the patient of the upcoming action, would that decrease the likelihood of a posited subject gap? And, conversely, if a preference for parallelism indicated that an upcoming RC structure would contain an object gap, and yet had an animate head, would the RC head preemptively be encoded as a patient?

The following three studies aim to explore the encoding of a filler in the presence of a biasing context, specifically investigating whether a preference for parallelism across conjuncts can diminish the filled-gap effect in the case of PO gaps, even if the first conjunct doesn't contain a filler-gap dependency. These will all be cases where a grammatical constraint (the Coordinate Structure Constraint) will not be applicable. In such cases, the parser cannot rely on an exact match in the structure, and therefore an effect of parallelism wouldn't be solely a case of structural priming, but also the matter of the initial encoding of a filler.

To understand this, I consider three potential influences on how a second-conjunct filler might be encoded: prior syntactic structure, prior highlighted thematic role, and prior information structure. Across the three studies, we also leverage a common association between various thematic roles and structural positions. In this case, a tendency for instrument roles to be found in prepositional phrases. However, while it may be common for instrument roles to be expressed in prepositional phrases, there is variety in the structural positions instruments can occupy. Specifically, in English, the common positions for instruments include as a prepositional object following *with* (39a) and as the object of *use* (39b), following an example from Rissman and Rawlins (2017). Therefore, it should be clear, these studies are not an example of whether participants predict an upcoming grammatically necessary structure, but rather whether they make use of frequently associated roles and structural positions.

(39) a. Brutus stabbed Caesar <u>with a knife</u>.

b. Brutus used <u>a knife</u> to stab Caesar.

Rissman and Rawlins (2017) in an examination of instruments following both *with* and *use* in English conclude that the instrument role is not one primitive feature, in line with theories that treat thematic roles as clusters of event properties. In the case of the instrument role, they define an instrumental event as a subevent of a larger event, in which the agent of the larger event acts upon the instrument. A distributed model of memory could accommodate either understanding of thematic information: as a category that, once satisfied, would be attached to the lexical representation, or as a cluster of features that, when activated, give rise to the properties we understand to fall under various thematic interpretations.

The following three studies will examine whether thematic role information in a first clause may influence predictive processes upon encountering a second clause. I entertain three hypotheses about the initial encoding of a filler in prior context that highlights an instrument role. First, a Structure Only hypothesis, which suggests that parallelism in filler-gap dependency resolution is driven by the pre-activation of syntactic representations, and not the activation of lexically encoded features. Next, an Encoded Features hypothesis, which suggests that given sufficient preceding context, a wh-filler may be predictively encoded with features on the basis of that preceding context, which may then be used to direct the postulation of gap positions, potentially avoiding a filled gap effect. Finally, an Activated Features hypothesis, which suggests that some kind of featural pre-activation can be used, as in priming, which would facilitate recognition and reanalysis, but would not be used predictively.

5.3 Experiment 7: Parallelism in syntactic structure

In the first experiment examining how a preference for parallelism influences filler encoding, parallelism was introduced on the level of syntactic structure, as closely as possible without including a dislocated prepositional object. This was achieved by including a prepositional phrase in the first conjunct (*with a spoon*), as seen in the following simplified version of the Experiment 7 stimuli (40).

(40) I ate the dessert with a spoon but I don't know what$_i$ everyone else ate the dessert with $\_{}_i$.

5.3.1 Methods

Materials

The experimental materials consisted of 24 items crossing the presence of a mentioned instrument in the first clause (+Instr, -Instr) and whether or not the second clause contained a filler gap dependency (+Gap, -Gap). The +Gap conditions always had a gap in the prepositional object position, with the same object from the first embedded clause repeated in the direct object position in the second embedded clause (e.g. *the dessert*). The instrument mentioned in the first clause was always in the form of a prepositional phrase, meaning that in the +Instr condition, the two clauses were parallel both in argument structure and in syntactic structure.

An example itemset can be found in Table 5.5.

	Preceding context
	Ben said that...
Instrument	**First clause**
-Instr	Carla ate the dessert
+Instr	Carla ate the dessert with a spoon
Gap	**Second clause**
+Gap	but he wasn't sure what Karen ate **the dessert** with at the party.
-Gap	but he wasn't sure if Karen ate **the dessert** with a fork at the party.

Table 5.4: Sample item set illustrating the Instrument x Gap manipulation. The Instrument manipulation here refers to whether instrument PPs are found in both clauses. The critical region (*the dessert*) in the second clause is bolded.

In English these PP attachments (*with a spoon*) are often ambiguous between NP attachment and VP attachment, corresponding respectively to an NP-modification interpretation and a modification of the event, in these cases describing an instrument used to participate in an action. We were interested in the latter, and therefore took effort to select PPs that would be more plausibly understood as an instrument. However, it is possible that the items varied in this way, as no separate norming study was conducted to judge the likelihood of the two different interpretations. A future norming study could be completed. This concern will also be addressed in Experiment 8, where this ambiguity will not be present.

Participants

This experiment was run alongside Experiment 1, therefore using the same participants. A total of 84 fully counterbalanced participants were included, 32 from Prolific, and 52 recruited from the UC Santa Cruz SONA Subject Pool.

Procedure

The materials for this experiment were included as fillers for Experiment 1, and therefore the same procedure was followed.

5.3.2 Data processing and analysis

Reading times above 2000 ms and below 100 ms were excluded from the final analysis. As the comprehension questions for this experiment were not relevant to the critical manipulation, trials with incorrect responses were treated as uninformative lapses in attention and therefore excluded. The reading time analysis was conducted following the same analysis procedures as in Experiment 1 and in Experiment 3. Sum contrast coding was used for both the Gap and Instrument factors. Analysis focused on the filled gap region (*the dessert*). We also analyzed several spillover regions following the critical regions.

5.3.3 Predictions

Following a general preference for parallelism, the hypotheses considered all predict some level of mitigation of the filled gap effect, either through syntactic priming, or encoded thematic information. However, the timecourse of mitigation will be informative with respect to how parallelism plays a role in the active search for a gap. If this information is used predictively, we would expect to see no filled gap effect in the +Instr condition, compared to the -Instr condition. If, instead, syntactic or thematic information is available to facilitate reanalysis, we would predict a filled gap effect regardless of the first clause, with a difference in recovery following the initial filled gap effect, such that the penalty for the -Instr condition would either be larger or would persist further into the following regions.

5.3.4 Results

At the determiner of the filled gap region (*the*), there was a main effect of Gap, such that +Gap conditions (M = 356 ms, SE = 5 ms) were read slower than -Gap conditions (M= 329 ms, SE = 5ms), $t = 3.68$, $p < .001$.

At the noun in the filled gap region (*dessert*), the main effect penalty for +Gap conditions continued. Here there was also a main effect of Instrument, such that the -Instr conditions (M = 339 ms, SE = 6 ms) were read slower than the +Instr conditions (M = 323 ms, SE = 5 ms), $t = 2.18$, $p < .05$. There was also a marginal interaction between Instrument and Gap, such that there was a bigger filled gap effect in the -Instr conditions (difference: 33 ms) than the +Instr conditions (difference: 12 ms), $t = 1.87$, $p = .06$.

The -Instr penalty continued into the following region, with -Instr conditions (M = 345 ms, SE = 6 ms) being read slower than +Instr conditions (M = 325 ms, SE = 5 ms), $t = 3.70$, $p < .001$. There was no effect of the Gap manipulation at this point.

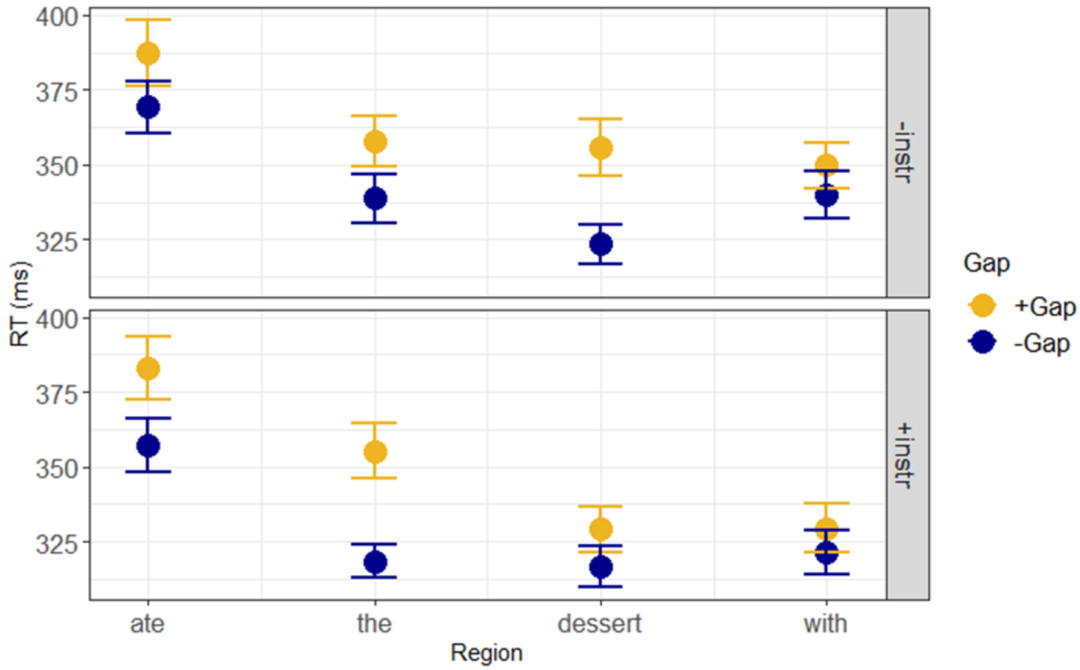

Figure 5.7: Reading times for Experiment 7.

5.3.5 Discussion

In Experiment 7, we found evidence for a filled gap effect in a direct object position, following a clause that contained a prepositional phrase. This suggests that a preceding prepositional phrase specifying an instrument role does not, upon entering the next clause and encountering a wh-filler, prompt an initial interpretation of the wh-filler as an instrument, and the gap position to be within a later prepositional phrase. We do find, however, that the filled gap effect persists longer in the case of non-parallel syntactic structure (the -Instr) condition. We interpret this as evidence that recent activation for either the preceding syntactic structure or thematic structure, or both, facilitates reanalysis at the filled direct object position.

It is reasonable that a prior prepositional phrase, without any filler-gap dependency, would not be sufficient to predictively postulate a prepositional object gap. After all, it is possible to satisfy a preference for parallelism in prepositional phrase structure, while still having a direct object gap (41).

(41) I used a spoon to eat the dessert, but I don't know what$_i$ Jonah ate $\__i$ with a spoon.

This experiment alone indicates at least the availability of information from a preceding clause to facilitate the recovery from a filled gap penalty, but is not designed to distinguish between the role of syntactic structure and thematic interpretation. Therefore, in the next experiment, thematic role specifically is examined, without parallelism in syntactic structure.

5.4 Experiment 8: Parallelism in thematic roles

In Experiment 7, the parallel conditions (+Instr) were parallel both in terms of syntactic structure and the thematic roles assigned. In Experiment 8, we used the construction *used a _* in order to highlight an instrument role, without introducing syntactic structure that might structurally prime prepositional phrase structure in the second conjunct. A simplified example of this can be found in (42).

(42) I used a spoon to each the dessert, but I don't know what$_i$ everyone else ate the dessert with $\__i$.

5.4.1 Methods

Materials

Just as with Experiment 7, the experimental materials for Experiment 8 consisted of 24 items crossing the presence of a mentioned instrument in the first clause (+Instr, -Instr) and whether or not the second clause contained a filler gap

	Preceding context Ben said that...
Instrument	**First clause**
-Instr	Carla ate the dessert
+Instr	Carla used a spoon to eat the dessert
Gap	**Second clause**
+Gap	but he wasn't sure what Karen ate **the dessert** with at the party.
-Gap	but he wasn't sure if Karen ate **the dessert** with a fork at the party.

Table 5.5: Sample item set illustrating the Instrument x Gap manipulation. The Instrument manipulation here refers to whether an instrument is mentioned in the first clause. The critical region (*the dessert*) in the second clause is bolded.

dependency (+Gap, -Gap). Again, just as in the previous experiment, the +Gap conditions always had a gap in the prepositional object position, with the same object from the first embedded clause repeated in the direct object position in the second embedded clause (e.g. *the dessert*). The central change for this experiment was the structure used to introduce the instrument role in the first embedded clause. In this experiment, the instrument always followed the construction *used a(n) [instrument] to VP*, meaning that in the +Instr condition, the two clauses were parallel in argument structure but, unlike the previous experiment, not in syntactic structure.

An example itemset can be found in Table 5.5.

Participants

This experiment was run alongside Experiment 3, therefore using the same participants. A total of 88 fully counterbalanced participants were included, 44 from Prolific, and 44 recruited from the UC Santa Cruz SONA Subject Pool.

Procedure

The materials for this experiment were included as fillers for Experiment 3, and therefore the same procedure was followed.

5.4.2 Data processing

Data were processed and analyzed in the same method as in Experiment 7.

5.4.3 Predictions

Without any amount of parallelism in the syntactic structure, we aimed to explore the difference between the Encoded Features and Activated Features hypotheses, to test whether a prior instrument role could be used to predictively avoid a filled gap penalty at the direct object position, or to facilitate the reanalysis following a filled gap effect.

Mechanistically, the Encoded Features hypothesis would suggest that a highlighted instrument role in the first clause might result in preemptively interpreting the underspecified wh-filler in the second clause as an instrument as well. Once encoded with an instrument role, which, in English, typically occurs in prepositional object position, the postulation of gap structure might predictively assume a prepositional object gap structure. This would lead to a lack of filled gap effect for the +Instr condition. The Activated Features hypothesis, on the other hand, would predict that the encoded of an instrument role would facilitate future instrument role encoding, i.e. boosting activation for the correct analysis during reanalysis following a filled gap penalty.

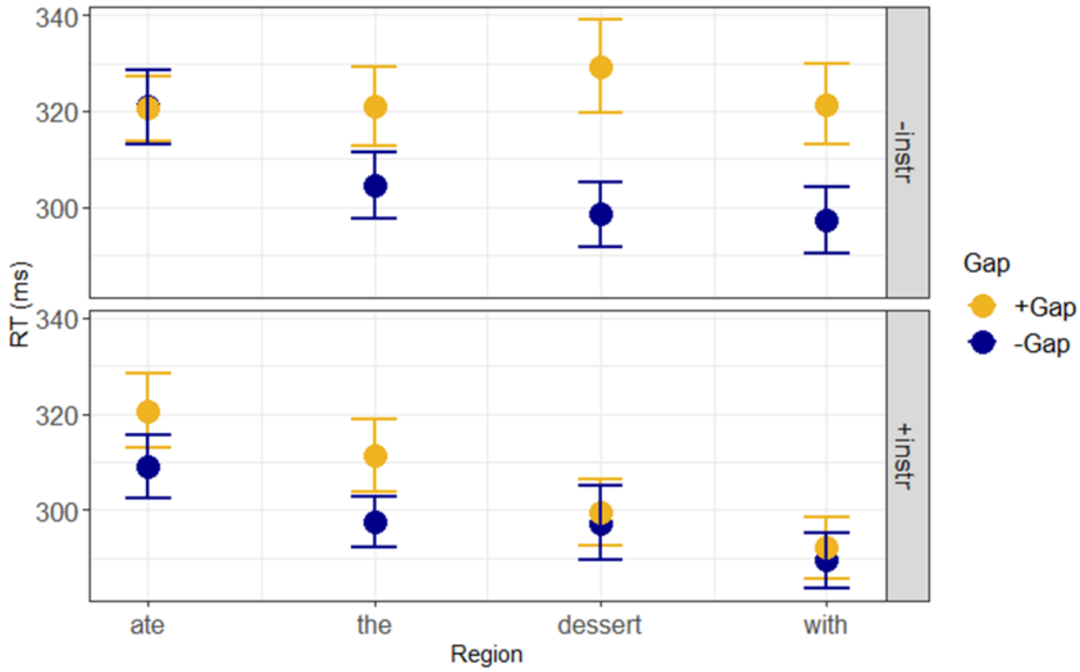

Figure 5.8: Reading times for Experiment 8.

5.4.4 Results

At the determiner of the filled gap region (*the*), there was a marginal effect of Gap, such that +Gap conditions (M = 316 ms, SE = 6 ms) were read slower than -Gap conditions (M = 301 ms, SE = 4 ms), t = 1.94, p = .06.

At the noun in the filled gap region (*dessert*), the main effect penalty for +Gap conditions continued, t = 2.49, $p < .05$, along with a main effect of Instrument, such that the -Instr conditions (M = 314 ms, SE = 6 ms) were read slower than the +Instr conditions (M = 298 ms, SE = 5 ms), t = 2.13, $p < .05$.

In the following region, both the -Instr penalty (difference: 19 ms), t = 2.27, $p < .05$, and the +Gap penalty (difference: 13 ms), t = 3.17, $p < .01$, continued.

5.4.5 Discussion

In Experiment 8, we again found evidence for a filled gap effect in a direct object position, regardless of the preceding clause, at the direct object determiner. As with Experiment 7, we interpret this as evidence that the postulation of a direct object gap was not fully avoided, and therefore, that thematic role information was not used to predictively postulate a prepositional object gap. At the following region, we found an effect of Instrument, with a significant advantage in reading times for the +Instr conditions. We argue that this is evidence for facilitation during reanalysis following a filled gap penalty.

Unlike Experiment 7, an advantage for a prior thematic role cannot be attributed to syntactic priming for a prepositional phrase. Therefore, here we see evidence that a prior instrument role can provide facilitation for a syntactic structure that had not been encountered before in that sentence, but is a likely associated syntactic structure with that role.

In regards to the initial encoding of the filler, it seems likely that the filler was not preemptively encoded with an instrument role, and that instead, the facilitation stems from prior activation of a particular role. Further discussion of what this activation might look like will be given in Section 5.6.

5.5 Experiment 9: Parallelism in information structure

Finally, we consider the role of broader information structure as a potential influence on the initial encoding of a filler.

There is reason to believe that information structure could interact with parallelism in this way. Deane (1991, 1992) suggests that parallelism in coordination arises not only from a syntactic constraint, but also from a preference for conceptually parallel conjuncts. Conceptual parallelism here may be understood to be parallelism in thematic roles, or in information structure. Deane (1992) appeals to stress assignment in coordinate structures, demonstrating parallel focus across conjuncts (43).

(43) a. The nurse polished her TROMBONE and the plumber computed my TAX.

b. The NURSE polished her trombone and the PLUMBER computed my tax.

Discourse prominence has also been shown to influence dependency resolution processes, for example, in preferences to refer to indefinites over definites in sluicing constructions and in a preference to resolve dependencies to a discourse-prominent position (Kuno, 1976; Harris, 2015).

Topicalized expressions have also been shown to trigger active dependency completion, despite the fact that would not be grammatically necessary. For example, Keshev and Meltzer-Asscher (2020) demonstrate a penalty reminiscent of a filled gap effect in the spillover region a direct object position (*the new security guard*) despite the fact that there is no wh-filler, in sentences such as (44, translated from Hebrew).

(44) The staff asked regarding the cashier if the tall manager forced the new security guard to throw him out last week.

The authors argue that not only syntactic dependencies, but also pragmatic motivations such as information structure considerations can trigger predictive dependency formation.

In Experiment 9, as a beginning look into the role of information structure in parallelism during filler gap dependency processing, we manipulated the definiteness of the direction object and the instrument prepositional object. The highlighted instrument role corresponded to conditions where the instrument in the first clause was preceded by an indefinite determiner. There are a number of motivations for this, including the tendency for discourse-new entities to be introduced as indefinite, and the fact that indefinites may trigger the generation of alternatives more so than definites.

5.5.1 Methods

Materials

Just as in Experiment 7 and 8, Experiment 9 consisted of 24 items crossing the presence of a highlighted instrument in the first clause (+Instr, -Instr) and whether or not the second clause contained a filler gap dependency (+Gap, -Gap). The material in the second clause remained unchanged from the previous two experiments. The first clause in this instance always contained a PP with an instrument. The +Instr condition in this corresponded to conditions where the instrument was an indefinite DP, for the reasons outlined above. The -Instr condition corresponded to definite DP instruments.

An example itemset can be found in Table 5.6.

	Preceding context Ben said that...
Instrument	**First clause**
-Instr	Carla ate **a** dessert with **the** spoon
+Instr	Carla ate **the** dessert with **a** spoon
Gap	**Second clause**
+Gap	but he wasn't sure what Karen ate **the dessert** with at the party.
-Gap	but he wasn't sure if Karen ate **the dessert** with a fork at the party.

Table 5.6: Sample item set illustrating the Instrument x Gap manipulation. The Instrument manipulation here refers to whether the instrument has a prominent role in terms of information structure (manipulated via definiteness). The critical region (*the dessert*) in the second clause is bolded.

The items were adapted directly from Experiment 7, and only after completion of the experiment was it noticed that the definiteness manipulation differentially affected the plausibility of the events depicted in the stimuli. To generalize, instances of implausibility arose when a definite direct object led to an improbable repeated event. For example, in the example item given in Table 5.6, it is possible to achieve a kind reading for *the dessert*, such that each party-goer could eat one token of the dessert. In contrast, an item such as *kill the bug* seems implausible - a kind interpretation seems unlikely. Instead, it seems most likely that the first clause is describing the killing of a specific bug. A generalization of these improbable events is that they depicted instances of the object being consumed or destroyed by the event, therefore rendering it implausible to refer to it again in the second clause.

Several examples of these differences can be found in Table 5.7. To acknowledge these potentially unacceptable items, a post-hoc analysis was done, splitting the items in this fashion. This will be described further in Section 5.5.1.

Possible related events (14 items)	Improbable repeated events (10 items)
eat the dessert	kill the bug
cross the lake	dig the hole
watch the match	break the vase

Table 5.7: A set of examples of items for Experiment 9 that conveyed possible and improbable repeated events.

Participants

68 participants were included in the final data set, recruited from Prolific.

Procedure

The procedure was consistent with the previous two experiments, but this iteration was run as fillers for a different set of experiments, including subject-verb number agreement manipulations and reflexive antecedent dependencies.

Data processing and analysis

The initial analysis procedure was the same as in the previous two experiments.

As mentioned above, following the initial analysis we noticed an issue with a subset of the items that made them implausible in the -Instr conditions, which could have introduced additional difficulty, obscuring the effect we were interested in. We therefore conducted a post-hoc analysis of items that did not have this problem, leaving only 14 items. The results will be presented here, but with an acknowledgment that the study may be under-powered, and need replication.

5.5.2 Results: All items

At the determiner (*the*) in the filled gap position, there was a main effect of Gap, such that + Gap conditions (M = 379 ms, SE = 6 ms) were read slower than -Gap conditions (M = 361 ms, SE = 5 ms), $t = 2.66$, $p < .05$. There was a marginal effect of highlighted Instrument, such that there was an advantage for +Instr conditions (M = 364 ms, SE = 5 ms) over -Instr (M = 376 ms, SE = 6 ms) conditions, $t = 1.73$, $p = .086$. While the interaction did not come out to be significant in the model, the facilitation for +Instr appears to be driven by an advantage specifically for the +Instr -Gap condition.

At the noun in the filled gap region (*dessert*), the penalty for +Gap conditions (M = 382 ms, SE = 6 ms) over -Gap conditions (M = 361 ms, SE = 5 ms) continued, $t = 3.69$, $p < .001$. Despite a numerically larger difference in means between the -Instr conditions in the +Gap conditions (30 ms) than in the -Gap conditions (12 ms), neither a main effect of Instrument nor an interaction reached significance.

At the following region, at the preposition (*with*), the penalty for +Gap conditions (M = 383 ms, SE = 6 ms) over -Gap conditions (M = 361 ms, SE = 5 ms) continued, $t = 4.15$, $p < .001$.

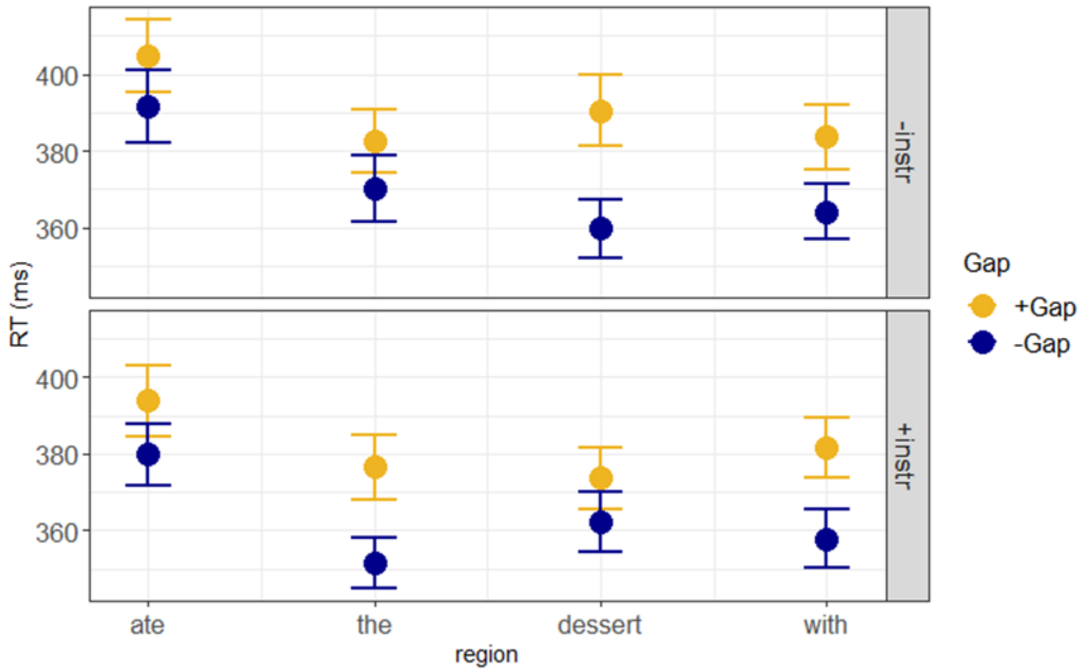

Figure 5.9: Reading times for all items in Experiment 9.

5.5.3 Results: Probable repeated events

At the determiner (*the*) in the filled gap position, there was a significant penalty for the +Gap conditions (M = 385 ms, SE = 7 ms) over the -Gap positions (M = 358 ms, SE = 7 ms), $t = 3.01$, $p < .01$. There was also a marginal penalty for -Instr conditions (M = 380 ms, SE = 8 ms) compared to +Instr conditions (M = 363 ms, SE = 7ms), $t = 1.89$, $p = .068$.

At the noun in the filled gap region (*dessert*), again, the penalty for +Gap conditions (M = 389 ms, SE = 8 ms) over -Gap conditions (M = 354 ms, SE = 7 ms) continued, $t = 4.2$, $p < .001$. In the analysis with only probable repeated events, a marginal effect of Instrument emerged, such that there was an advantage for +Instr conditions (M = 365 ms, SE = 7 ms) compared to -Instr conditions (M = 379 ms, SE = 8 ms), $t = 1.88$, $p = .06$. A significant interaction also emerged between Gap and Instrument, $t = 2.2$, $p < .05$, in which there was a larger filled

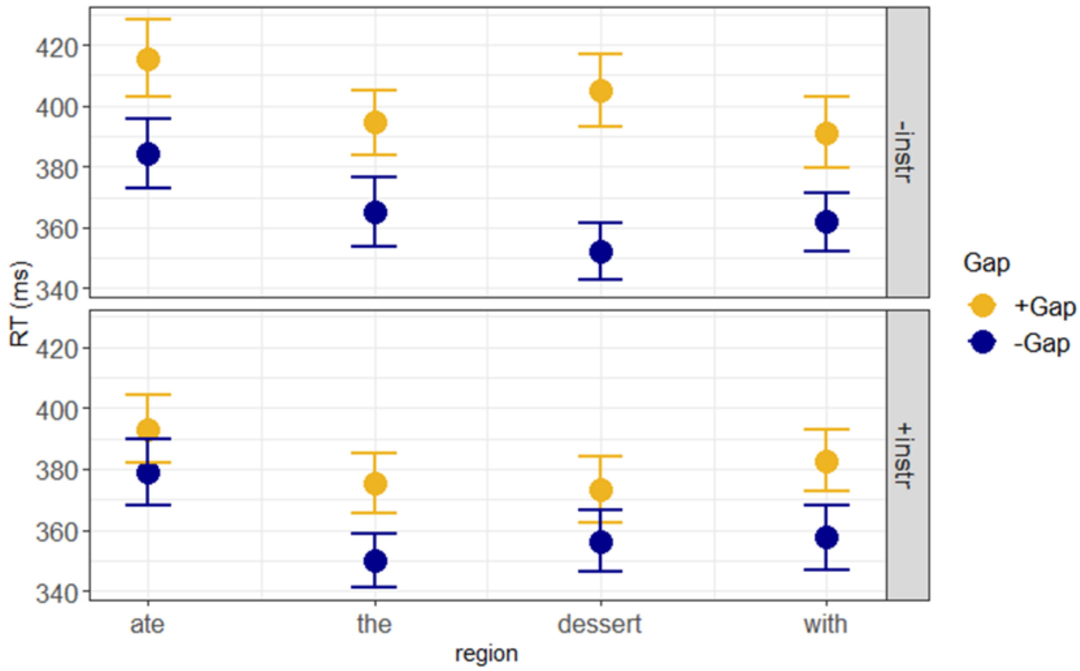

Figure 5.10: Reading times for probable repeated events in Experiment 9.

gap penalty for the -Instr conditions (53 ms) compared to the +Instr conditions (17 ms).

By the following preposition, only the filled gap penalty remains: a penalty for the +Gap conditions (M = 387 ms, SE = 8 ms) over the -Gap conditions (M = 360 ms, SE = 7 ms), $t = 3.27$, $p < .01$.

5.5.4 Discussion

Experiment 9 aimed to manipulate how highlighted an instrument role was, while still maintaining an instrument in each condition. This was done via manipulation of definiteness, following the typicality of definites to be given, and backgrounded, in discourse. Furthermore, indefinites are more likely to trigger a generation of alternatives, which could then be targeted by a future wh-filler construction.

Experiment 9 found an overall similar pattern to Experiments 7 and 8, of a more widespread penalty for the -Instr conditions (in this case, definite instruments) compared to the +Instr conditions, but a clear and present filled gap effect regardless. Again, this suggests that discourse-prominence of an instrument role was not enough to predictively postulate a gap structure consistent with a common structural position for instruments, but did facilitate the reanalysis of the gap position at the filled gap region.

In the case of a definiteness manipulation, the underlying mechanisms behind the +Instr advantage could be similar to the activation account as discussed for Experiment 8. It could also be more indirect: if participants are more likely to generate alternatives for the indefinite noun, this would result in the activation of potential alternative instruments. In other words, given the specifics of the example in Table 5.6, participants might be more likely to generate potential alternatives for the condition represented in (45b) than (45a).

(45) a. Carla ate a dessert with the spoon –> { cake, soup, pastry, ...}

b. Carla ate the dessert with a spoon –> { fork, knife, her hands, ...}

Encountering a filled direct object position, when a direct object gap had been postulated, involves not just the reanalysis of the syntactic structure, but also the information structure as well. Therefore, having already generated alternatives for an instrument role may facilitate reanalysis not at the syntactic level, but at a more conceptual level.

Future work would be needed to tease apart these different sources of facilitation. To begin, it would be beneficial to replicate this design using an appropriate set of events, to avoid the issues encountered here in the +Instr conditions.

5.6 Discussion

Overall, Chapter 5 examined the ways in which encoded features influence wh-filler gap dependency processing, looking either at the consequences of overlapping activation spanning the dependency, or the use of features to (re)postulate the position of a gap site.

5.6.1 Experiment 6 and overlapping activation

Experiment 6 examined the status of semantic information, such as attributes that would determine category membership, during active dependency resolution in wh-filler gap dependencies. We observed evidence for similarity-based interference on the basis of semantic similarity, and cases where this emerged in particular following overlapping activation with a similar intervener.

We concluded that the mechanisms of interference consistent with our results are, in the case of disrupted features, a feature overwriting account, or in the case of interference stemming from interaction between features, interference on the basis of similar contextual features or interference from conflicting feature-feature associations. Future work would be necessary to distinguish between these different sources of interference.

5.6.2 Activated features in the reanalysis of gap positions

Experiments 7-9 test whether a prior thematic role might influence a dependency resolution process, allowing the parser to predictively avoid the filled gap penalty. Results from all three studies provide evidence against the ability to proactively and predictively use this information: in all three studies, a filled gap effect was detected even with a prior thematic role present.

Instead, converging evidence across the three studies indicates that a prior instrument role can facilitate the recovery from a filled gap effect at a direct object position, at least in part due to the relationship between the instrument thematic role and its common position within a prepositional phrase in English. This effect cannot be solely attributed to structural priming from encountered structures, or a preference for parallelism solely in syntactic structure, as the effect holds even without a prior prepositional phrase (Experiment 8).

At the moment, the hypothesis most in line with the results here is one of Activated Features, where prior activation of a feature can facilitate later processing, even if it is not predictively encoded on a relevant token (the wh-filler). If we consider a weight matrix of potential features, if one feature had been encoded recently, there may be some lingering activation there that could be capitalized on during a moment of error, as the resting activation for that feature would be closer to a threshold of encoding. This could be understood as facilitatory interference, where lingering activation facilitates the construction of a future analysis. In the case of thematic roles, this could be modeled either as a binary category (+/- Instrument), or as a cluster of features that, when activated together, provide the conceptual understanding of the instrument role. The explanation for the facilitatory effect laid out here does not depend on a commitment in either direction for how the instrument feature is represented. Future work could test whether it may be possible to find, instead, inhibitory interference during a reanalysis process, where lingering activation in support of the incorrect parse makes it harder to suppress that first incorrect analysis.

A replication of facilitatory interference as it's described here, or inhibitory interference as it's proposed, could be explored via other means other than creating parallel structures. Parallelism in coordination was used here specifically as a means to motivate some expectation of the second clause (i.e. presence of an instrument role). If other features were examined, expectation setting could be achieved through other means. For example, biasing towards an inanimate or animate continuation on the basis of broader contextual cues or selectional restrictions of a preceding verb.

Another open question to address, specifically regarding the encoding of thematic roles, is the level at which this information is encoded, especially given a view that thematic roles reflect clusters of event properties. It may be that thematic information is not encoded at the lexical level, but at a higher level of representation. Neurolinguistic evidence on the encoding of thematic roles has found distinct neural correlates for agent and patient roles, but has not been able to distinguish between the possibility of attaching the role to the context-independent representation of the concept, or to a higher level representation, for example to the event being depicted (Wang et al., 2016).

Chapter 6

Conclusion and future work

In this section, I will give a brief overview of the central findings of this work, and offer some avenues for future exploration stemming from the themes explored by this dissertation.

6.1 Summary of findings

This dissertation aimed to address two high level concepts as they pertain to encoding, maintenance, and retrieval in memory: similarity in encoded features, and some notion of *closeness* during encoding. I will overview what this dissertation has shown about each, and how we can reach a unified understanding of both.

6.1.1 Similarity

Over the course of the dissertation, many forms of similarity have been looked at. Similarity has been treated as degree of featural overlap, assuming a distributed representation of memory, in which an item's attributes are represented across a multidimensional vector.

Experiments 1-3 classified similarity along three levels: similarity along many dimensions including physical properties, uses in context; similarity on the basis of roles within events (i.e. plausibility as the object of a specific verb); and similarity on the basis of grammatical features (e.g. number, animacy). One consistent finding across these studies was that of a penalty for Medium similarity, or similarity in roles in context, but not entirely overlapping in terms of all physical properties.

Experiments 5 investigated sensitivity to noun-noun swaps, in instances where those nouns were more semantically similar to one another than to the other nouns in the sentence. For this metric of similarity, word2vec was used. As these were all animate nouns, high similarity items were often related occupations: in other words, people who perform similar tasks or frequent similar places. As there was no condition looking at the same phenomenon with unrelated nouns, the role of semantic similarity here cannot be directly addressed.

Experiment 6 investigated semantic similarity-based interference for inanimate nouns, using category membership as a metric of similarity, assuming, for example, that there may be a cluster of features that determines a class of objects (e.g. silverware, boats, electronics, furniture). The low similarity items were not always entirely dissimilar, for example, potentially found in similar contexts but not a member of the same category.

Experiments 7-9 looked at similarity across clauses, finding that a previously activated feature (thematic role) could facilitate later structural reanalysis.

6.1.2 Closeness

Closeness during encoding was explored in a number of ways.

Experiments 1-3 examined closeness as temporal proximity during initial encoding, finding differential interference on the basis of initial proximity, but in inconsistent ways. This is likely due to additional consequences of these proximity manipulations (the changing of relative clause structure, the addition of a parenthetical).

Experiments 4-5 examined closeness as membership within the same linguistic grouping, finding promising evidence in recall and recognition that participants did treat linguistic boundaries also as boundaries in memory.

Experiment 6 examined closeness as overlapping activation during dependency resolution, finding more consistent downstream interference in cases of overlapping activation.

6.1.3 Closeness as similarity

The dissertation began by investigating two ingredients in encoding interference: similarity and proximity. In moving forward, we can understand these to not necessarily be two separate things, as proximity can be seen as similarity on the basis of encoded contextual features. The most natural extension forward from this dissertation is a path of further examining the contextual status of lexical items that are reactivated, and the consequences of different types of linguistic boundaries, such as different cues to prosodic boundaries or different clausal relationships.

6.2 Implications and future work

Why all this care about encoding? Where does this help us in understanding sentence processing and language generally, beyond the areas of sentence processing that are already actively engaging in this topic, such as the study of dependency resolution? In this section I'd like to lay out a few areas of future work that stem from the work in this dissertation and the broader literature cited here.

6.2.1 Broader discourse influences on encoded features

The experiments in this dissertation have focused on instances of single, out-of-context sentences in reading. A question may be posed of how immediately surrounding context influences the set of features encoded for a particular word. As an example, we can examine the feature of animacy, which could be thought of as reflecting a characteristic that is inherent to and unchanging for any particular entity. However, in day-to-day language use, people may anthropomorphize inanimate objects, for example ascribing intention to appliances that are acting up[3]. The phenomenon of anthropomorphizing is just one example of how a context may change our understanding of the properties we ascribe to the specific entity being described compared to a general, default understanding. Nieuwland and Van Berkum (2006) demonstrate that expectations surrounding a typically inanimate object can be changed given sufficient anthropomorphizing context, finding no penalty for *yacht* compared to *sailor* as the object of the verb *advised* in the context in (46), so long as there had been several sentences prior that introduced the *yacht* as an entity that received a thematic role typically assigned to animates (e.g. experiencer role). Similarly, they find that out-of-context expectations reverse in the presence of anthropomorphizing context, such that the continuation *salted* becomes more unexpected than *in love* in (47). In both cases, the presence or change of a penalty was detected via ERP signal at these critical regions.

(46) The psychotherapist advised the { yacht | sailor } to be honest.

[3] A quick Google search of 'printer hates me' returns many blog posts and discussion threads describing printers as entities which may intentionally cause disruptions, or are out to get their users.

(47) The peanut was { salted | in love }.

It's unclear, however, at what level of representation the update from default inanimate to contextual animacy is being made. It is possible that a higher-level, discourse representation is changed, but that at the level of an individual lexical item it is the typical animacy that is encoded. Therefore, it is unclear whether a decrease in penalty for real-word inconsistent continuations (e.g. *in love*) might also lead to changes in expectations about the NPs and likely upcoming structural positions. Rich et al. (2022) investigated this particular question, looking for the well-established ORC penalty for animate RC heads in the case of anthropomorphized typically-inanimate objects, using contexts as in (48), compared to a context which did not anthropomorphize any typically-inanimate objects.

(48) In a recent episode of a television show about struggling restaurants, the host went to the kitchen of a local seafood restaurant and interviewed some of the employees: a chef, a bowl, and a plank of wood. Relationships among the employees were strained after a recent violent episode where the chef insulted the bowl, and the bowl tried to hit him in the face. The bowl was visibly upset when the host asked about what happened, and after a little bit of conversation, fists started flying again while the cameras were rolling. Even the host got involved in the fray.

a. SRC: It was tense when the { host | bowl } that struck the chef in the kitchen fell onto the floor.

b. ORC: It was tense when the { host | bowl } that the chef struck in the kitchen fell onto the floor.

Using the Maze task (Forster et al., 2009), in which participants read through a sentence by making a choice about the next upcoming word at each word, this study found that animacy-based ORC penalty differences emerged in the continuation in (48b), such that *the chef* was read slower following the RC head *host* than *bowl*, but only in the non-anthropomorphizing context. This difference collapses in the anthropomorphizing context.

Taken together, the results of Nieuwland and Van Berkum (2006) and Rich et al. (2022) suggest that when a larger discourse establishes a specific entity that differs from a typical, generic understanding, comprehenders update their expectations not only about likely upcoming events, but also likely upcoming structure. To the extent that launching such expectations relies on the featural encoding of a particular lexical item, these findings suggest that biasing context may change the encoding of, in this case, the feature(s) representing animacy. An alternative explanation for the lack of animacy-based difference in the anthropomorphizing context is that comprehenders stopped attending to animacy cues to generate structural expectations. Future iterations of that work could explicitly test this by also looking at typically-inanimate RC heads that were mentioned in the prior discourse as unambiguously inanimate objects.

More broadly, this line of research encourages future study of how the encoded representation of an entity or concept may change over the course of a discourse, as new information is encountered. The contexts used by Nieuwland and Van Berkum (2006) and Rich et al. (2022) were essentially short stories from a single, narrative voice. This work could further be expanded to look at instances of discourse occurring between two or more interlocutors discussing the same set of characters or objects, examining how negotiations in discourse influence the encoding and re-encoding of lexical items.

6.2.2 Predictive consequences of encoding interference

A large body of research suggests that, beyond the predictive postulation of gaps explored in this dissertation, some degree of prediction occurs during sentence processing. Effects of predictability have been shown to have a facilitatory effect during incremental processing, suggesting preemptive activation for highly predictable words. This is evidenced by findings across methodologies, with highly predictable sentence continuations eliciting shorter fixations in reading studies (Ehrlich & Rayner, 1981; Rayner & Well, 1996) and smaller N400s in ERP studies (Kutas & Hillyard, 1984; Kutas & Federmeier, 2011) than their less predictable counterparts. The predictability of a word is determined by how tightly constraining the preceding context is. Language comprehenders are sensitive to constraint that builds cumulatively over the course of a sentence, as well as constraint that stems from a word or two immediately preceding the target word (Fitzsimmons & Drieghe, 2013). Furthermore, comprehenders can rapidly shift their expectations following cues that contradict previous information, as in the case of a classifier in Mandarin that is incompatible with the previously most likely continuation (Chow & Chen, 2020) or the case of concessive markers in English (such as *even though* or *although*) that rapidly reverse a comprehenders' expectation of what is most likely (Xiang & Kuperberg, 2015; Rich & Harris, 2023).

There has been some disagreement about what occurs during the prediction process itself, in terms of the pre-activated representation(s). Is one lexical item pre-activated, or many? Converging evidence suggests it is likely that there could be spreading activation for a number of related potential continuations (Staub, 2015, for review).

What features of the predicted word(s) are pre-activated? For example, is the phonological form represented? Early research suggests there is sensitivity to phonological form, to the point that encountering a word whose form depends on the phonological form of the predicted continuation (e.g. in English, the indefinite determiner *a(n)*) is sufficient to confirm or disconfirm a prediction (DeLong, Urbach, & Kutas, 2005), although there have been failures to replicate this finding (Nieuwland et al., 2018). In languages with pre-nominal articles that are marked for other features, such as gender or number (e.g. in Spanish, la_{FEM}, el_{MASC}), there is a reliable (but small) effect for encountering an article that does not match the features of the predicted noun (Nicenboim, Vasishth, & Rösler, 2020).

How strongly committed is a parser to a predicted word? Lau, Holcomb, and Kuperberg (2013) suggest that there might be a distinction between pre-activating the representation of a predicted word and pre-updating the sentence context with that representation, i.e. some form of meaning integration. Ness and Meltzer-Asscher (2018) propose that if another word is found in the place of a predicted word (a *disconfirmed prediction*), active inhibition of the predicted word would only be necessary if there had been a strong commitment to that representation. Thinking of this now in terms of the Temporal Context Model (Howard & Kahana, 2002), at what point is a contextual vector included in the distributed representation of a word? I would propose that in the case of a strong commitment, if some amount of integration has occurred with the sentence, that the activation of that word would include some added contextual feature as well. Typically the next context state should include information of the previous context, which, in this case, would include information about the predicted word. If that word had not been encountered, is it possible that the contextual feature encoded on the encountered word would in fact include some information about the disconfirmed prediction as well?

Taken all together, we can summarize that prediction in language comprehension is a process that preemptively spreads activation across likely continuations by activating the feature bundles that comprise those representations. I envision three ways in which the study of encoding can inform our understanding of predictive processing, as it pertains both to generating predictions on the basis of past encoded features, and encoding the features of the predicted words themselves:

1. Earlier encoding interference may affect future predictions: As comprehenders incorporate constraint from the broader discourse into lexical-level predictions, and rapidly update their expectations in the face of new and contradicting information, it stands to reason that efficient prediction relies on accurate memory representations for previously encountered linguistic content. If the featural representation of previous items in the sentence, it is possible that the strength or accuracy of the predictive process could be affected.

2. Disconfirmed prediction as a locus for encoding interference: Instances of disconfirmed lexical predictions introduce an instance of overlapping activation, similar to what is discussed for Experiment 7. It is therefore possible that the encoding of the featural representation of the encountered word could be disrupted by the disconfirmed prediction.

3. Encoded nominal features may influence predictions of upcoming structure, even when disconfirmed: Following potential interference on the encoding of a noun, any structural predictions that could have been made on the basis of the disrupted features (e.g. animacy) may then be affected.

Beyond the encoding of nouns in context, it may be that the features of a subject itself are used to preemptively postulate the features of upcoming verbal structure, as was implemented in the Self-Organized Sentence Processing model discussed in Section 2 (Smith et al., 2021). In that case, predictions about an upcoming verb are represented with features compatible with the subject it must eventually associate with. Interference from a highly similar intervener may strengthen the featural representation of the predicted material, resulting in faster reading times at the verb, while interference from a partially matching intervener could weaken the featural representation of the predicted material, resulting in slower reading times at the verb. Given an underlying mechanism such as this, interference during retrieval that cannot be explained by cue overload may result from interference during the initial encoding of a relevant target and intervener *and/or* from interference during the generation of expectations about upcoming lexical content or structure on the basis of those initial encodings.

Further testing is needed to explore the latter possibility. It is possible that degraded featural representations of preceding material could lead to less specific predictions, or render an otherwise tightly constraining context less so if the features that contribute to the high level of constraint are not properly encoded. Or encoding interference may result in an increased reliance on predictions stemming from the broader discourse at large. These potential downstream consequences could be operationalized and examined in many of the ways predictability is traditionally probed, via facilitation in early reading measures, decreases in the N400 component, or by means of examining Cloze results.

6.2.3 (Potentially) unexpected sources of encoding interference

This dissertation has looked at interference in linguistic stimuli between words encountered in those stimuli themselves. I would like to take a moment to address instances of potential interference that may arise in experimental designs as a consequence of the methodology.

The maze task (Forster et al., 2009) invites participants to read through a sentence word by word by making a decision at each point. There are different instantiations of the maze task. For example, in some versions participants may choose between the correct continuation and a series of X's, while in others they may choose between the correct continuation and a nonse word. Another popular version of the maze task displays two words, with the foil a grammatically implausible continuation that is length-matched to the target continuation. The benefits of the maze task are clear: it necessitates closer attention than a task like self-paced reading, asking participants to make an explicit judgment on the basis of whether a word can be integrated into the sentence context they are reading. Consequently, effects that are at times found across spillover regions in self-paced reading are often found more locally in the maze task. However, what is the status of the foil word, and can it interfere with the processing of the target continuation? Gallant and Libben (2020) suggest that foils can interfere in lexical processing during the maze task, finding that a prior semantically-related foil can facilitate reaction times on a later target. Conversely, they also find that a target that is semantically related to a later foil will have an inhibitory effect at that juncture, making it more difficult to reject that foil. Future work could investigate how the lexical encoding of a target is influenced by the foil it is paired with, further testing the ideas in this dissertation surrounding co-activation, even when it is clear that one of the words will not be incorporated into the sentence context.

Other methodologies could be addressed from this perspective as well, such as visual world paradigms, where participants may be looking at (and consequently accessing their mental conceptual representation of) an image while hearing a non-associated word.

6.2.4 Applications beyond neurotypical populations

The population samples in this dissertation are either from a 4-year research university campus, or sampled from Prolific with restrictions on education and literacy difficulties. The literacy difficulties restriction excluded participants who disclosed a diagnosis of ADHD or dyslexia.

Future work would benefit from expanding on these samples in a number of ways, benefiting our understanding of both the role of memory in language processing and the full scale of variation in human language processing. One potential starting point would be in looking at populations with ADHD: research has demonstrated differential performance in recall memory tasks between those diagnosed and not diagnosed with ADHD. One particular finding of interest to this dissertation is the finding that participants with ADHD show overall lower accuracy, but stronger contiguity effects, during recall (Gibson, Healey, & Gondoli, 2019). They suggest that these findings stem from a difference in contextual drift, compared to neurotypical controls. Seeing as work is still being done in understanding how children with ADHD may struggle with written or verbal instruction in schools (e.g. Wassenberg et al., 2010), it seems like a promising and potentially impactful area of future study to understand how the encoding of linguistic context may differ for people with ADHD.

6.3 Final conclusion

In sum, it is my hope that this dissertation has contributed to the field's understanding of the kinds of features we encode during incremental language comprehension - and not only *what* features are encoded, but how and under what circumstances these encodings may be disrupted, interfere with one another, or be used to benefit future processing.

References

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, *29*, 313–341.

Anderson, J. R., & Lebiere, C. J. (1998). *The atomic components of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Arnett, N., & Wagers, M. (2017). Subject encodings and retrieval interference. *Journal of Memory and Language*, *93*, 22–54.

Avetisyan, S., Lago, S., & Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language*, *112*, 104087.

Badecker, W., & Kuminiak, F. (2007). Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in slovak. *Journal of memory and language*, *56*(1), 65–85.

Badecker, W., & Straub, K. (2002). The processing role of structural constraints on interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 748.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The english lexicon project. *Behavior research methods*, *39*, 445–459.

Barker, J., Nicol, J., & Garrett, M. (2001). Semantic factors in the production of number agreement. *Journal of psycholinguistic research*, *30*, 91–114.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... Bolker, M. B. (2015). Package 'lme4'. *convergence*, *12*(1), 2.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive psychology*, *18*(3), 355–387.

Bower, G. H., & Winzenz, D. (1969). Group structure, coding, and memory for digit series. *Journal of Experimental Psychology*, *80*(2p2), 1.

Branigan, H. P., Pickering, M. J., Liversedge, S. P., Stewart, A. J., & Urbach, T. P. (1995). Syntactic priming: Investigating the mental representation of language. *Journal of Psycholinguistic Research*, *24*, 489–506.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, *80*, 1–28.

Calhoun, S. (2010). How does informativeness affect prosodic prominence? *Language and Cognitive Processes*, *25*(7-9), 1099–1140.

Chen, Z., Jäger, L., & Vasishth, S. (2012). How structure-sensitive is the parser? evidence from mandarin chinese. *Empirical approaches to linguistic theory: Studies of meaning and structure*, 43–62.

Chow, W.-Y., & Chen, D. (2020). Predicting (in) correctly: listeners rapidly use unexpected information to revise their predictions. *Language, cognition and neuroscience*, *35*(9), 1149–1161.

Christensen, R. H. B. (2015). Analysis of ordinal data with cumulative link models—estimation with the r-package ordinal. *R-package version*, *28*, 406.

Chromỳ, J., & Vojvodić, S. (2024). When and where did it happen? systematic differences in recall of core and optional sentence information. *Quarterly Journal of Experimental Psychology*, *77*(1), 111–132.

Clewett, D., DuBrow, S., & Davachi, L. (2019). Transcending time in the brain:

How event memories are constructed from experience. *Hippocampus*, *29*(3), 162–183.

Crain, S., & Fodor, J. D. (1985). How can grammars help parsers. *Natural language parsing: Psychological, computational, and theoretical perspectives*, 94–128.

Deane, P. D. (1992). *Grammar in mind and brain: Explorations in cognitive syntax* (Vol. 2). Walter de Gruyter.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, *8*(8), 1117–1121.

Dillon, B., Clifton Jr, C., & Frazier, L. (2014). Pushed aside: Parentheticals, memory and processing. *Language, Cognition and Neuroscience*, *29*(4), 483–498.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85–103.

Dillon, B., & Wagers, M. (2021). Approaching gradience in acceptability with the tools of signal detection theory. In (chap. The Cambridge Handbook of Experimental Syntax). Cambridge: Cambridge University Press.

Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: number agreement in sentence production. *Psychological review*, *112*(3), 531.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, *20*(6), 641–655.

Fedorenko, E., Babyonyshev, M., & Gibson, E. (2004). The nature of case interference in on-line sentence processing in russian. In *Proceedings-nels* (Vol. 34, pp. 215–226).

Fitzsimmons, G., & Drieghe, D. (2013). How fast can predictability influence word skipping during reading? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1054.

Ford, M. (1983). A method for obtaining measures of local parsing complexity throughout sentences. *Journal of verbal learning and verbal behavior*, *22*(2), 203–218.

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior research methods*, *41*, 163–171.

Franck, J., & Wagers, M. (2018). Hierarchical structure and memory retrieval mechanisms in agreement attraction. *bioRxiv*, 422139.

Frankish, C. (1995). Intonation and auditory grouping in immediate serial recall. *Applied Cognitive Psychology*, *9*(7), S5–S22.

Frazier, L. (1987). Syntactic processing: evidence from dutch. *Natural Language & Linguistic Theory*, *5*(4), 519–559.

Frazier, L., & Clifton Jr, C. (1989). Successive cyclicity in the grammar and the parser. *Language and cognitive processes*, *4*(2), 93–126.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325.

Frazier, L., Munn, A., & Clifton, C. (2000). Processing coordinate structures. *Journal of Psycholinguistic Research*, *29*, 343–370.

Gallant, J., & Libben, G. (2020). Can the maze task be even more amazing? adapting the maze task to advance psycholinguistic experimentation. *The Mental Lexicon*, *15*(2), 366–383.

Gennari, S. P., & MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. *Journal of memory and language*, *58*(2), 161–187.

Gibson, B. S., Healey, M. K., & Gondoli, D. M. (2019). Adhd reflects impaired externally directed and enhanced internally directed attention in the immediate free-recall task. *Journal of Abnormal Psychology*, *128*(2), 173.

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of experimental psychology: learning, memory, and cognition*, *27*(6), 1411.

Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of memory and Language*, *51*(1), 97–114.

Gordon, P. C., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of experimental psychology: Learning, Memory, and Cognition*, *32*(6), 1304.

Gordon, P. C., Hendrick, R., & Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychological science*, *13*(5), 425–430.

Gordon, P. C., & Lowder, M. W. (2012). Complex sentence processing: A review of theoretical perspectives on the comprehension of relative clauses. *Language and Linguistics Compass*, *6*(7), 403–415.

Harris, J. A. (2015). Structure modulates similarity-based interference in sluicing: An eye tracking study. *Frontiers in psychology*, *6*, 155701.

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection theory: A user's guide*. Routledge.

Healey, M. K., Long, N. M., & Kahana, M. J. (2019). Contiguity in episodic memory. *Psychonomic bulletin & review*, *26*(3), 699–720.

Hommel, B. (1998). Event files: Evidence for automatic integration of stimulus-response episodes. *Visual cognition*, *5*(1-2), 183–216.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 923.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of mathematical psychology*, *46*(3), 269–299.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2015). Retrieval interference in reflexive processing: experimental evidence from mandarin, and computational modeling. *Frontiers in psychology*, *6*, 125783.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological review*, *99*(1), 122.

Kahana, M. J. (2012). *Foundations of human memory*. OUP USA.

Keshev, M., & Meltzer-Asscher, A. (2020). The effects of syntactic pressures and pragmatic considerations on predictive dependency formation. *Language, Cognition and Neuroscience*, *35*(2), 256–272.

Kim, N., Brehm, L., Sturt, P., & Yoshida, M. (2020). How long can you hold the filler: maintenance and retrieval. *Language, Cognition and Neuroscience*, *35*(1), 17–42.

Kuno, S. (1976). Gapping: A functional analysis. *Linguistic Inquiry*, *7*(2), 300–318.

Kush, D., Johns, C. L., & Van Dyke, J. A. (2015). Identifying the role of phonology in sentence-level reading. *Journal of memory and language*, *79*, 18–29.

Kush, D., & Phillips, C. (2014). Local anaphor licensing in an sov language: Implications for retrieval strategies. *Frontiers in psychology*, *5*, 97190.

Kush, D. W. (2013). *Respecting relations: Memory access and antecedent retrieval in incremental sentence processing* (Unpublished doctoral dissertation). University of Maryland, College Park.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, *62*, 621–647.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmertest package: tests in linear mixed effects models. *Journal of statistical software*, *82*(13).

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating n400 effects of prediction from association in single-word contexts. *Journal of cognitive neuroscience*, *25*(3), 484–502.

Lawn, A. R. (2020). *The varying roles of morphosyntax in memory and sentence processing: Retrieval and encoding interference in brazilian portuguese* (Unpublished doctoral dissertation). University of California, Los Angeles.

Lee, M.-W. (2004). Another look at the role of empty categories in sentence processing (and grammar). *Journal of Psycholinguistic Research*, *33*, 51–73.

Lelandais, M., & Ferré, G. (2018). Perception of prosodic boundaries by naïve listeners in three different types of subordinate syntactic constructions. In *9th international conference on speech prosody 2018* (pp. 104–108).

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 375–419.

Logačev, P., & Vasishth, S. (2011). Case matching and conflicting bindings interference. In *Case, word order and prominence: Interacting cues in language production and comprehension* (pp. 187–216). Springer.

McElree, B. (2006). Accessing recent events. *Psychology of learning and motivation*, *46*, 155–200.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of memory and language*, *48*(1), 67–91.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moreton, B. J., & Ward, G. (2010). Time scale similarity and long-term memory for autobiographical events. *Psychonomic Bulletin & Review*, *17*, 510–515.

Nairne, J. S. (2006). Modeling distinctiveness: Implications for general memory theory. *Distinctiveness and memory*, 27–46.

Ness, T., & Meltzer-Asscher, A. (2018). Lexical inhibition due to failed prediction: Behavioral evidence and erp correlates. *Journal of experimental psychology: Learning, memory, and cognition*, *44*(8), 1269.

Ness, T., & Meltzer-Asscher, A. (2019). When is the verb a potential gap site? the influence of filler maintenance on the active search for a gap. *Language, Cognition and Neuroscience*, *34*(7), 936–948.

Nicenboim, B., Vasishth, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? evidence from new data and a bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia*, *142*, 107427.

Nicol, J., & Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *Journal of psycholinguistic research*, *18*, 5–19.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... others (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, *7*, e33468.

Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of cognitive neuroscience*, *18*(7), 1098–1111.

Oberauer, K. (2009). Design for a working memory. *Psychology of learning and motivation*, *51*, 45–100.

Oberauer, K., Farrell, S., Jarrold, C., Pasiecznik, K., & Greaves, M. (2012). Interference between maintenance and processing in working memory: The effect of item–distractor similarity in complex span. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 665.

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of memory and language*, *55*(4), 601–626.

Omaki, A., Lau, E. F., Davidson White, I., Dakan, M. L., Apple, A., & Phillips, C. (2015). Hyper-active gap filling. *Frontiers in psychology*, *6*, 384.

Parker, D. (2017). Processing multiple gap dependencies: Forewarned is forearmed. *Journal of Memory and Language*, *97*, 175–186.

Parker, D., & Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition*, *157*, 321–339.

Patil, U., Vasishth, S., & Lewis, R. L. (2016). Retrieval interference in syntactic processing: The case of reflexive binding in english. *Frontiers in Psychology*, *7*, 167541.

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: a critical review. *Psychological bulletin*, *134*(3), 427.

Potts, C. (2002). The syntax and semantics of as-parentheticals. *Natural Language & Linguistic Theory*, *20*, 623–689.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr, C. (2012). *Psychology of reading*. Psychology Press.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1290.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*(4), 504–509.

R Core Team, R., et al. (2013). R: A language and environment for statistical computing.

Rich, S., Balachandran, L., Duff, J., Kogan, M., Van Handel, N., Wax Cavallaro, M., & Wagers, M. (2022). *"it's alive!" animacy-based structural predictions update with context.* (The 35th Annual Conference on Human Sentence Processing)

Rich, S., & Harris, J. A. (2023). Global expectations mediate local constraint: evidence from concessive structures. *Language, Cognition and Neuroscience*, *38*(3), 302–327.

Rissman, L., & Rawlins, K. (2017). Ingredients of instrumental meaning. *Journal of Semantics*, *34*(3), 507–537.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, *12*, 1–8.

Ross, J. R. (1967). Constraints on variables in syntax.

Slioussar, N., & Malko, A. (2016). Gender agreement attraction in russian: production and comprehension evidence. *Frontiers in psychology*, *7*, 166019.

Smith, G. (2018). A theory of timing effects in a self-organizing model of sentence processing.

Smith, G., Franck, J., & Tabor, W. (2018). A self-organizing approach to subject–

verb number agreement. *Cognitive science*, *42*, 1043–1074.

Smith, G., Franck, J., & Tabor, W. (2021). Encoding interference effects support self-organized sentence processing. *Cognitive Psychology*, *124*, 101356.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and linguistics compass*, *9*(8), 311–327.

Stowe, L. A. (1986). Parsing wh-constructions: Evidence for on-line gap location. *Language and cognitive processes*, *1*(3), 227–245.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of memory and language*, *48*(3), 542–562.

Sturt, P., Keller, F., & Dubey, A. (2010). Syntactic priming in comprehension: Parallelism effects with and without coordination. *Journal of Memory and Language*, *62*(4), 333–351.

Sturt, P., & Martin, A. E. (2016). Using grammatical features to forecast incoming structure: The processing of across-the-board extraction. In *The 29th cuny human sentence processing conference.*

Swinney, D., Ford, M., Frauenfelder, U., & Bresnan, J. (1988). On the temporal course of gap-filling and antecedent assignment during sentence comprehension. *Unpublished manuscript*, *15*, 17.

Tanenhaus, M. K., Carlson, G. N., & Seidenberg, M. S. (1985). Do listeners compute linguistic representations. *Natural language parsing: Psychological, computational, and theoretical perspectives*, *10*.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of memory and language*, *47*(1), 69–90.

Traxler, M. J., & Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, *35*(3), 454–475.

Traxler, M. J., Williams, R. S., Blozis, S. A., & Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language*, *53*(2), 204–224.

Uitvlugt, M. G., & Healey, M. K. (2019). Temporal proximity links unrelated news events in memory. *Psychological science*, *30*(1), 92–104.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 407.

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*(3), 285–316.

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of memory and language*, *55*(2), 157–166.

Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of memory and language*, *65*(3), 247–263.

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4), 685–712.

Villata, S., Tabor, W., & Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. *Frontiers in psychology*, *9*, 310880.

Wagers, M. (2008). *The structure of memory meets memory for structure in linguistic cognition.* University of Maryland, College Park.

Wagers, M., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of memory and language*, *61*(2), 206–237.

Wagers, M., & Pendleton, E. (2016). Structuring expectation: Licensing animacy in relative clause comprehension. In *Proceedings of the 33rd west coast conference on formal linguistics* (pp. 29–46).

Wagers, M., & Phillips, C. (2009). Multiple dependencies and the role of the grammar in real-time comprehension1. *Journal of Linguistics*, *45*(2), 395–433.

Wang, J., Cherkassky, V. L., Yang, Y., Chang, K.-m. K., Vargas, R., Diana, N., & Just, M. A. (2016). Identifying thematic roles from neural representations measured by functional magnetic resonance imaging. *Cognitive neuropsychology*, *33*(3-4), 257–264.

Wanner, H., & Maratsos, M. P. (1978). An atn approach to comprehension. In *Linguistic theory and psychological reality.* mit Press.

Wassenberg, R., Hendriksen, J. G., Hurks, P. P., Feron, F. J., Vles, J. S., & Jolles, J. (2010). Speed of language comprehension is impaired in adhd. *Journal of Attention Disorders*, *13*(4), 374–385.

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(4), 442.

Wen, T., & Egner, T. (2022). Retrieval context determines whether event boundaries impair or enhance temporal order memory. *Cognition*, *225*, 105145.

Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, cognition and neuroscience*, *30*(6), 648–672.

Yadav, H., Smith, G., Reich, S., & Vasishth, S. (2023). Number feature distortion modulates cue-based retrieval in reading. *Journal of Memory and Language*, *129*, 104400.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, *123*(2), 162.