An Entropy Model of Artificial Grammar Learning

Emmanuel M. Pothos (e.pothos@bangor.ac.uk)
School of Psychology, 39 College Road
Bangor, LL57 2DG, UK

Todd M. Bailey (todd@psy.ox.ac.uk)
Department of Experimental Psychology; South Parks Road
Oxford, OX1 3UD, UK

Abstract

We propose a model to characterize the type of knowledge acquired in Artificial Grammar Learning (AGL). In particular, we suggest a way to compute the complexity of different test items in an AGL task, *relative* to the training items, based on the notion of Shannon entropy: The more predictable a test item is from training items, the higher the likelihood that it will be selected as compatible to the training items. Our model is an attempt to formalize some aspects of inductive inference by providing a quantitative measure of the knowledge abstracted by experience. We motivate our particular approach from research in reasoning and categorization, where reduction of entropy has also been seen as a plausible cognitive objective. This may suggest that reducing (Shannon) uncertainty may provide a single explanatory framework for modeling as diverse aspects of cognition, as learning, reasoning, and categorization.

Introduction

Artificial Grammar Learning (henceforth AGL; Reber, 1989; Redington & Chater, 1996) is an experimental paradigm to study inductive inference. An artificial grammar is a set of rules that can be used to generate sequences of symbols. These sequences are labeled grammatical (G) to distinguish them from ungrammatical sequences (NG), which are sequences that violate the rules of the finite state language. Figure 1 shows an example of one such grammar. With this set of rules, while the string MSSV is legal, this would not be the case for string MMSV.

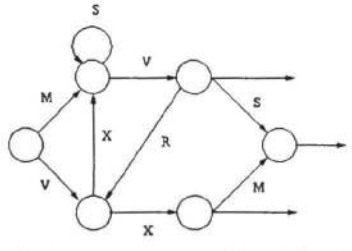

Figure 1: This is the grammar used in Reber & Allen, 1978, as well as Pothos & Chater (1998a, submitted), whose results are analyzed in this work.

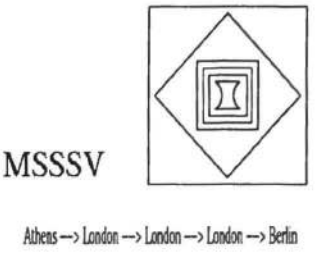

Figure 2: Examples of the types of stimuli used in Pothos & Chater (1998a, submitted). Letter strings were used in Experiment 1, arrangements of shapes in Experiment 2, and city sequences in Experiment 3.

In a typical AGL experiment the sequences of symbols are presented as letter strings, such as MSSV or MSSSV (but these sequences can also be, for example, graphical symbols or musical tones; see Figure 2). Participants are presented with a subset of the G strings in a training phase, and asked to observe them, but no other information is provided either about the nature of the strings, or about the subsequent test phase. After training, they are told that the strings they saw all complied to a set of rules and are then asked to identify the novel G strings in a set that contains both G and NG ones. A robust finding in the literature is that participants can identify the new G strings with above chance accuracy, while in many circumstances they are unable to fully articulate the basis on which they made their decisions.

Pothos and Chater (1998, submitted) provided results indicating that overall performance in AGL does not vary, regardless of whether the stimuli are letter strings (as is the standard condition), embedded shapes, or sequences of cities that correspond to the routes of an airline company (see also Pothos & Bailey, 1997; Bailey & Pothos, 1998).[1]

[1] Altmann, Dienes, and Goode (1995), as well as Whittlesea & Wright (1997) also present evidence that AGL-type of learning is possible with stimuli other than the standard AGL strings, but these investigators have not attempted a direct comparison of performance across the different conditions.

Performance was investigated in terms of overall accuracy in detecting G strings, as opposed to NG ones, and also patterns of error across the different sets the test items could be divided into.

The fact that performance does not appear to be different in conditions as different as the ones used by Pothos & Chater (1998a, submitted), appears to suggest that the type of learning observed with AGL reflects general properties of the learning process (that is, properties that do not depend on the particular experimental format used in different situations). Thus, the project of identifying an adequate theory of how participants generalize in an AGL task from training items to the test ones is an important one.

Investigators have proposed accounts of AGL performance in the context of rules, stimulus fragments (parts), or similarity. The original claim by Reber and his colleagues (see Reber, 1989, for a review) has been that participants learn in training something of the abstract, rule structure of the finite state language used to create the stimuli. This view has been corroborated by "transfer" experiments, where the symbols used in training were different from the symbols used in test. However, one has to observe that the actual artificial grammar used in different experiments is an object defined entirely by the experimenter; there is no reason to expect *a priori* that it will be psychologically relevant. In this sense, Dulany, Carlson, & Dewey's (1984) theory would appear more realistic. These investigators have instead argued that participants acquired "correlated grammars," that is a set of "microrules" which generally approximated the true grammar, but might at the same time include unrepresentative or even wrong rules.

Perruchet and Pacteau (1990) asked what is the minimal type of knowledge that could be used by participants and lead to the observed levels of accuracy. They suggested that all that is learned is information about the legal bigrams, that is which pairs of symbols have been observed in the training items (see Gomez and Schvaneveldt, 1994, and Johnstone & Shanks, in press, for an extension of this approach; Redington & Chater, 1996, for a re-evaluation of these results).

Other theorists suggested that an important factor is similarity: That is, whether a test item is selected as G or NG will depend, to some extent, on how similar it is to training items. Similarity has been operationalized in different ways, for example, as symbol differences between test and training items (Brooks and Vokey, 1991), or empirically computed on the basis of direct similarity data from participants (Bailey & Pothos, 1998; Pothos & Bailey, 1997). A similar approach by Knowlton and Squire (1996) differs in that these investigators computed item similarity in the context of an instantiation of Servan-Schreiber's (1991; see also Servan-Schreiber & Anderson, 1990) "chunking-hypothesis," which is a general theory of learning; the main finding of previous investigations, that similarity is an important predictor of grammaticality performance, has been replicated.

What do all the above theories share? Unfortunately very little. Theories such as the above can be used to make predictions as to which items would be more likely to be selected as G in the test part of an AGL task (that is, predictions about "grammaticality endorsements"). While the actual predictions made by different models in practice often correlate very highly (e.g., see Johnstone & Shanks, in press), there is little theoretical insight as to the extent to which these models are supposed to be mutually exclusive (in terms of representing different *hypotheses* about learning processes) or not. For example, is the microrules approach (Dulany et al., 1984) the same as Perruchet & Pacteau's (1990) bigram proposal? They both suggest that the knowledge acquired in an AGL task is of the form: If there is an M then a V must follow; although there are some qualifications, these theories would still probably be compatible in terms of their predictions. However, one theory is in terms of rules, while the other might be more reminiscent of exemplar models of classification. Furthermore, if one accepts that AGL is supposed to be a small scale, experimental version of real-life learning tasks (and the utility of investigating AGL would be arguable otherwise), in most of the above cases one cannot readily see how the explanations proposed could generalize to other learning situations, or relate to existing accounts of other aspects of cognition.

Motivation

Our aim in this research is to derive a model of AGL performance from the same *computational* principles that have been seen as relevant in research in reasoning and categorization, namely the assumption that uncertainty is quantified via Shannon entropy and that the cognitive system operates in a way to reduce this uncertainty. Whether the same principles underlie cognitive performance in areas as diverse as reasoning, categorization, and learning is arguable; however, here we suggest as a plausible hypothesis that these processes reflect the same, basic, problem of inductive inference (that is the successful generalization from previously seen instances to future events). We begin with a brief presentation of the models in reasoning and categorization that motivate the present work.

For several years investigators assumed that human reasoning is mediated by the rules of classical logic (e.g., Braine et al., 1995; Evans, 1991). The observation that people often fall prey to an alarmingly large number of logical and probabilistic fallacies, and recent theoretical investigations criticizing the appropriateness of logic for everyday reasoning (Chater & Oaksford, 1993), have led theorists to pursue alternative approaches. The Wason selection task is a simple problem where people are asked to examine whether a conditional rule is true or false, by selecting among a set of cards (the cards are labeled with

one clause of the conditional, and contain hidden information about the other clause of the conditional). Oaksford & Chater (1994) suggest that people select these cards that minimize the expected *uncertainty* in deciding whether the rule is true or not. Uncertainty is quantified using the notion of Shannon entropy: If there are N events, that can occur with probability p_i, then the entropy in trying to guess which one will actually occur is given by $entropy = -\sum_{i=1}^{N} p_1 \log(p_i)$. As will become clear in the presentation of our model of AGL performance to follow, we suggest that the items people will be selecting as G in the test part of an AGL task, are the ones that are most predictable in terms of the training items. That is, we predict that the strings selected as G are the ones that minimize the entropy of specifying them, relative to the training items. This is a strategy very similar to Oaksford and Chater's (1994), who claimed that people select these cards that minimize the entropy of selecting the right hypothesis.

In categorization, Pothos & Chater (1998b) suggested a model whereby people's classifications on a set of items were such so as to reduce the *description length* (used here in a technical sense) of the items as much as possible. In other words, categories were seen as a means to simplify the description of a set of items as much as possible. Pothos (1998) illustrated that the mathematical framework of Pothos & Chater (1998) is equivalent to an entropy minimization one: That is, the preferred classification of a set of items is assumed to be the one that reduces the uncertainty in predicting the similarity structure of these items.

The very brief exposition above can only provide a presentation of the models in question at a very crude, qualitative level. At such a level, it might appear that the theoretical coherence afforded by terms like "reduction of uncertainty," or "entropy minimization," is only an artifact of the fact that the mathematical specification of models based on such notions is relatively loose; so that conceptually different models can still be instantiated in a way that would appear consistent with an entropy maximization process. This is far from true. Although there can be several different entropy maximization procedures to address the same cognitive problem, such alternatives still need to share the same foundation (a specific use of probabilities, quantifying uncertainty in a certain way, etc.), that would make them much more similar, as a class of models, compared to others.

An entropy model of AGL

This model is an attempt to quantify what exactly is learned in training in an AGL task. In such a task the test items are evaluated in terms of whether they are compatible with the training items or not. What can this mean? We suggest that each test item is given a complexity measure according to how "specifiable" it is from training items. This complexity measure is computed by dividing the item into parts, and seeing how "determinable" the continuation to each of these parts can be on the basis of information from training.

First, each test string is broken down into all constituent fragments, "anchored" at the beginning or end of the string. Letting symbols "b" and "e" stand for the beginning and the end of a string, test string MSV would be broken into [b, bM, bMS, bMSV] in the forward direction, and fragments [e, Ve, SVe, MSVe] in the reverse. We consider these fragments as relevant, on the simple assumption that symbol sequences are likely to be parsed/ encoded by the cognitive system in a simple forward and reverse direction. For a given test item, we ask what is the expected difficulty of specifying a continuation, given what one has seen in training, and in this way we compute the S-measures for each fragment (for the reverse chunks, we ask how likely a given symbol is to precede a particular fragment; for simplicity, we use continuation to refer to both, when discussing general properties of the S-measures). In particular, if there are N possibilities for a continuation, each occurring with a probability p_i from training, then the entropy associated with specifying the next symbol in the string is given by $S(fragment) = -\sum_{i=1}^{N} p_1 \log(p_i)$.

For example, suppose that the training items consist only of strings MSSV, MSSSSX, and MSVRV. When we see MSV in test, then to compute the overall complexity of this string, relative to the training items, we need to consider, first, S(b): How hard is it to guess what the next symbol is, for the first symbol in a string, from training? All training strings start with an M, thus, we have S(b) = 0. Likewise, S(bM) = 0. To compute S(bMS), note that after fragment bMS in training, we have an "S" continuation with a probability of 2/3 and a "V" one (the observed continuation in test) with probability 1/3. Thus, S(bMS) would be 1/3log(1/3)-2/3log(2/3). Taking an example for the reverse S measures, S(e) would be computed by noting that in the training items an end symbol is preceded by a V symbol with a probability 2/3 and an X one with probability 1/3.

In cases where there is a novel symbol in a test item, or a fragment not seen in training, we compute S by assuming that all the possible symbols are equiprobable. In the above example, if we had a test string QM, then forward S(bQ) would be given by $-5 \times \frac{1}{5} \log_2 \frac{1}{5}$ (we have five possible symbols, M, S, V, Q, and "e," and since the fragment bQ has not been observed in training, all possible continuations are equiprobable). The underlying hypothesis is that if there is no information from training about a given sequence, the cognitive system will operate as if all possible continuations were equiprobable.

How would the S-measures corresponding to the different fragments lead to an overall complexity measure for a string? We suggest that all forward and reverse S-measures

are averaged, and that the resulting number reflects how familiar a given test string is relative to the training items. Without going into too much detail here, using an average, instead of some other way of combining *S*-measures, has the advantage that the computation of the overall complexity of an item is more balanced across items, and also for individual items that contain a single violation of regularity relative to the training items (e.g., in the above example, think of an item like VMSV), the effect of this single violation is moderated by the presence of regular fragments (regular with respect to the training items).

Investigation of the model

The limited exposition of our model cannot address all the relevant theoretical issues. In this section, we aim to alleviate some concerns by fitting the model to results from three AGL experiments reported in Pothos & Chater (1998a, submitted). They utilized the Reber & Allen (1967) grammar to create AGL stimuli that were standard letter strings (Experiment 1), nested arrangements of geometric shapes (Experiment 2), and sequences of cities (Experiment 3). The average *S*-measure reflects how specifiable each test item is on the basis of training items. That is, a high *S*-measure indicates that a particular string is not "intuitive" relative to the training items. Thus, we predicted that the extent to which different test items would be selected as G would negatively correlate with the average *S*-measure of these items.

Table 1 shows the correlation of the average *S*-measure for the test strings of the Reber and Allen (1967) grammar, with the probability that these strings would be selected as G in each of the three experiments of Pothos & Chater (1998a, submitted). That is, for each of these experiments, we averaged the total number of times (across participants) each test items was selected as G. Considering the low number of participants in these studies (ten participants each) compared to the high number of endorsements we are trying to predict (50 items), the fact that all the correlations are in the predicted direction and highly significant provides important support for our model.

Table 1: The correlation of the average *S*-measure of individual test items complexity, with the number of times each test item (50 test items, in total) was selected as G in the three experiments of Pothos & Chater (1998a, submitted).

	Letters	Shapes	Cities
Average S-measures			
Pearson Correlation	-0.577	-0.613	-0.381
p-value	0.000	0.000	0.006

This work is not meant to disconfirm any of the existing models. Our objective has been to propose a computation of "what is learned in AGL," in a wider theoretical context, that is, in a way that relates to research in other areas of cognition. That is, we hope to provide a model of AGL that would still capture as much of the intuitions seen as relevant in other AGL accounts, while the model itself would be inspired from more general computational principles. In this respect, finding that the average *S*-measure correlates with the predictions from other models of AGL performance, would provide important support for our approach.

Table 2 shows the correlation of the *S*-measure with grammaticality, global associative chunk strength (Knowlton and Squire, 1996), associative chunk strength at anchor positions (same as before, but computed only for fragments at the beginning or end of a string, that is the anchor positions), novel chunk strength (the fraction of novel fragments in a string; see Meulemans & Van der Linden, 1997), and anchor novel chunk strength (same as before, but one looks at the proportion of novel fragments at the anchor positions). Table 2 reveals that our measure correlates very highly with almost all the above measures.

Table 2: The number in each of the rows represents the correlation of the Average *S*-measure with the AGL performance measure in each row. The "*" flags correlations significant at the 0.01 level or less. All correlations were computed over the 50 test items in the Pothos & Chater (1998a, submitted), work.

	Average S-measure	
grammaticality	-.630	*
global chunk strength	-.436	*
anchor chunk strength	-.515	*
novel chunk strength	-.378	*
anchor novel chunk strength	-0.022	

Discussion and future direction

We have tried to quantify the amount of information that is available in the test part on an AGL task from training. In this respect, we proposed the average *S*-measure which, for each string, provides us with a number reflecting how specifiable the string is (that is, how easily it can be determined), given a particular set of training items.

To examine our model, we computed average *S*-measures for all the strings of the Reber and Allen (1978) grammar and showed that these correlated significantly with grammaticality endorsements from three AGL experiments reported by Pothos & Chater (1998a, submitted). Moreover, we showed that the average *S*-measure captures many aspects of previously proposed measures of AGL performance, that made different assumptions about what is learned in an AGL task.

The underlying motivation for the present work was to provide a quantitative measure of knowledge acquired in an AGL task, in a broad theoretical context. Thus, our model can be seen as having a foundation very similar to Oaksford & Chater's (1994) model of reasoning in the selection task, and Pothos & Chater's (1998) model of categorization. In all these cases, it is assumed that reduction in uncertainty (quantified via Shannon's entropy) is the objective for the cognitive system in processing information about the world. Further fleshing out the formal relation between such seemingly diverse aspects of cognition is an important future objective.

The average *S*-measure can be employed to model on-line generalization, that is generalization patterns from individual items, in the sense that the first item can be used to compute the complexity of the second one, the first and second together, the complexity of the third one, and so forth. Moreover, one can manipulate the "total information available from training," and thus make predictions about the overall level of grammaticality accuracy in an AGL task, when the actual number of training items presented is always the same. Both the above considerations represent possible simple extensions of the present work, to further test the psychological plausibility of the average *S*-measure of generalization.

Acknowledgments

We would like to thank Nick Chater for valuable comments on this work. Emmanuel Pothos was supported by ERSC grant ref. R000222655 to Emmanuel Pothos and Nick Chater. Todd Bailey was supported by grants from the McDonnell-Pew Center for Cognitive Neuroscience, Oxford, and a grant to Kim Plunkett from the Biotechnology and Biological Sciences Research Council, UK

References

Altmann, G. T. M., Dienes, Z. & Goode, A. (1995). Modality Independence of Implicitly Learned Grammatical Knowledge. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21,* 899-912.

Bailey, T. M. & Pothos, E. M. (1998). Unconfounding similarity and rules in artificial grammar learning. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 96-101, LEA: Mahwah, NJ.

Braine, M. D. S., O'Brien, D. P., Noveck, I. A., Samuels, M. C., Lea, B. L., Fisch, S. M., Yang Y. (1995). Predicting Intermediate and Multiple Conclusions in Propositional Logic Inference Problems: Further Evidence for a Mental Logic. *Journal of Experimental Psychology: General, 124,* 263-292.

Brooks, L. R. & Vokey, J. R. (1991). Abstract Analogies and Abstracted Grammars: Comments on Reber (1989) and Mathews et al. (1989). *Journal of Experimental Psychology: Learning, Memory and Cognition, 120,* 316-323.

Chater, N., & Oaksford, M. (1993). Logicism, mental models and everyday reasoning. *Mind and Language, 8,* 72-89.

Evans, St B. T. J. (1991). Theories of Human Reasoning: The Fragmented State of the Art. *Theory & Psychology, 1,* 83-105.

Gomez, R. L., Schvaneveldt, R. W. (1994). What Is Learned From Artificial Grammars? Transfer Tests of Simple Association. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20,* 396-410.

Johnstone, T. & Shanks, D. (submitted). Two Mechanisms in Implicit Grammar Learning? Comment on Meulemans and Van der Linden (1997).

Knowlton, B. J., Squire, L. R. (1996), Artificial Grammar Learning Depends on Implicit Acquisition of Both Abstract and Exemplar-Specific Information. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22,* 169-181.

Meulemans, T., Van der Linden, M. (1997). Associative Chunk Strength in Artificial Grammar Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1007-1028.

Oaksford, M. & Chater, N. (1994). A Rational Analysis of the Selection Task as Optimal Data Selection. *Psychological Review, 101,* 608-631.

Perruchet, P. & Pacteau, C. (1990). Synthetic Grammar Learning, Implicit Rule Abstraction or Explicit Fragmentary Knowledge?. *Journal of Experimental Psychology, General, 119,* 264-275.

Pothos, E. M. (1998) Aspects of Generalisation. Unpublished D.Phil thesis, University of Oxford.

Pothos, E. M. & Bailey, T. M. (1997). Rules vs. Similarity in Artificial Grammar Learning. In *Proceedings of the Similarity and Categorisation Workshop 97,* 197-203, University of Edinburgh.

Pothos, E. M. & Chater, N. (1998a). Generality of the Abstraction Mechanisms in Artificial Grammar Learning. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 854-858, LEA: Mahwah, NJ.

Pothos, E. M. & Chater, N. (1998b). Rational Categories. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 848-853, LEA: Mahwah, NJ.

Pothos, E. M. & Chater, N. (submitted). Generality of the Abstraction Mechanisms in Artificial Grammar Learning. *European Journal of Cognitive Psychology.*

Reber, A. S. (1989). Implicit Learning and Tacit Knowledge. *Journal of Experimental Psychology, General, 118,* 219-235.

Reber, A. S., Allen R. (1978). Analogic and abstraction strategies in synthetic grammar learning, A functional interpretation. *Cognition, 6,* 189-221.

Redington, M. & Chater, N. (1996). Transfer in Artificial Grammar Learning, Methodological Issues and Theoretical Implications. *Journal of Experimental Psychology, General, 125,* 123-138.

Servan-Schreiber, E. (1991). *The Competitive Chunking Theory: Models of Perception, Learning, and Memory.* Doctoral Dissertation, Department of Psychology, Carnegie-Mellon University.

Servan-Schreiber, E., Anderson, J. R. (1990). Learning Artificial Grammars With Competitive Chunking. *Journal of Experimental Psychology: Learning Memory and Cognition, 16,* 592-608.

Whittlesea, B. W. & Wright, R. L. (1997). Implicit (and explicit) learning, Acting adaptively without knowing the consequences. *Journal of Experimental Psychology, Learning, Memory and Cognition, 23,* 181-200.