Exploring the Role of Representation in Models of Grammatical Category Acquisition

Ting Qian (tqian@bcs.rochester.edu)[1]
Patricia A. Reeder (preeder@bcs.rochester.edu)[1]
Richard N. Aslin (aslin@cvs.rochester.edu)[1]
Josh B. Tenenbaum (jbt@mit.edu)[2]
Elissa L. Newport (newport@bcs.rochester.edu)[1]
[1]Department of Brain & Cognitive Sciences, University of Rochester
[2]Department of Brain & Cognitive Sciences, MIT

Abstract

One major aspect of successful language acquisition is the ability to generalize from properties of experienced items to novel items. We present a computational study of artificial language learning, where the generalization patterns of three generative models are compared to those of human learners across 10 experiments. Results suggest that an explicit representation of word categories is the best model for capturing the generalization patterns of human learners across a wide range of learning environments. We discuss the representational assumptions implied by these models.

Introduction

Learning the grammar of a language consists of at least two important tasks. First, learners must discover the cues in the linguistic input that are useful for constructing the grammar of the language. Second, learners must represent their knowledge of the grammar in a form that makes it possible to assess the grammaticality of future input. With an appropriate representation of the grammar, learners can generalize from properties of the small set of experienced items to predicted properties of novel items. This ability for generalization is crucial for language acquisition, as the input for learning is naturally limited. Such generalization should extend to only the novel items that are actually licensed by the language, no more (over-generalization) and no less (under-generalization).

Previous research has offered several hypotheses regarding the cues that learners use and the representations of grammar they form. In the realm of syntactic category acquisition, one hypothesis is that the categories (but not their contents) are innately specified prior to receiving any linguistic input, with the assignment of words to categories accomplished with minimal exposure (e.g. McNeill, 1966). On this view, both the cues and the representations are predefined and independent of linguistic input. A contrasting view states that grammatical categories are learned, though different hypotheses appeal to the importance of different cues or cue combinations during the learning process (such as semantic cues, e.g., Bowerman, 1973). Within this class of non-nativist hypotheses, several studies have suggested that distributional cues may be sufficient for extracting the grammar of the input language (e.g., Braine, 1987; Maratsos & Chalkley, 1980; Mintz et al., 2002). Distributional cues are defined over patterns in the linguistic input, such as token frequencies, co-occurrence statistics, and latent structural dependencies between linguistic elements. Although studies have shown that human learners and computational models can successfully learn grammatical categories when only these cues are available, the question of representation still remains poorly understood. How do learners represent the knowledge of previously encountered linguistic items in order to generalize to novel ones?

The aim of the present work is to ask what types of representations are used by human learners in an artificial grammar learning (AGL) task that includes many of the distributional properties of spoken language. We focus on how learners induce grammatical categories and assign words to them. Our approach involves computational modeling, comparing the simulated learning outcome of three different models, each of which makes a different assumption about how learners represent the learned grammar. We assess the models by comparing the generalization patterns of each model and those of human learners. Our experimental data come from our previous findings across 10 AGL experiments (Reeder et al., in review; Schuler et al., in prep). In the next section, we first provide a brief summary of these results. Importantly, the goal of our modeling work is not to mirror every detail of human behavior in AGL experiments: to do so, one must consider psychological variables such as memory and attention, which are currently not included in our models. Instead, we are interested in exploring the representational assumptions that human learners have adopted in our experiments.

Background on Behavioral Results

The behavioral data come from a series of 10 experiments with adult participants in which we created an artificial grammar with the structure (Q)AXB(R). Each letter represents a category of nonsense words. Q and R words served as optional categories that made sentences of the language vary in length from 3 to 5 words and made words of the language observe patterning in terms of relative order but not fixed position. The sizes of the categories varied across experiments, leading to different numbers of possible sentences in the language. For ease of presentation, we will number the experiments. In Experiments 1-4 (Reeder et al., 2009), there were 108 possible sentences that could be created from this grammar; in Experiment 5 (Reeder et al., 2009), there were 576 possible sentences; in Experiments 6-10 (Reeder et al., 2010;

Schuler et al., in prep), there were 144 possible sentences.

Participants in these experiments were first exposed to a carefully selected subset of the possible sentences of the grammar. The exposure strings were chosen to test whether specific distributional cues enabled learners to form a category of lexical items and generalize to novel words, or to allow exceptions that maintain lexical specificity. In particular, different experiments tested learners sensitivity to the *contexts* of individual words and their individual frequencies, the *sparsity* of sampling the language, the *overlaps* among contexts across words, the non-overlap of contexts (or *systematic gaps* in information), and the size of the exposure set. In each experiment, a portion of the possible strings was withheld in order to create different kinds of "gaps" in the input to participants.

After exposure, subjects completed a grammaticality rating task, where they rated strings on a scale from 1-5, with larger values indicating higher grammaticality. The test was comprised of three types of test strings: familiar AXB sentences (presented during exposure), novel AXB sentences (withheld from the exposure set), and ungrammatical strings that violated the AXB word order (i.e., "A1X1A2" or "A1A2B3"). Importantly, in order to understand how learners generalized from training sentences and the type of knowledge representations suggested by their generalization behaviors, we varied the way the presentation set and the gaps occurred in each experimental condition, as summarized in Table 1. In Experiments 1-2, we varied the sparseness of sampling the language, but learners heard all AX and XB bigrams. In Experiments 3-4, we varied the overlap of contexts across X words: each X was heard with only 2 of the 3 As and Bs. Experiments 6-9 included a new X word (called "X4") that appeared in only one sentence frame in the training subset. The purpose was to test whether learners would generalize to X4 as one of the X words (and therefore able to occur in all X-word contexts), despite its own extremely limited exposure and minimal overlap with the other X-word contexts. Experiment 5 created subcategories in the language, with distinct occurrence privileges for X words and contexts words: half of the X words only occurred with half of the A words and half of the B words, while the remaining X words occurred with the remaining As and Bs.

In experiments 1-9, the bigram statistics were carefully balanced: all grammatical bigrams were presented equally often (with the exception of the X4 bigrams in Experiments 6-9). Under this balanced design, one possible strategy for judging grammaticality could be simply to keep track of bigram statistics. To examine this, we ran Experiment 10, where the bigram statistics were not balanced.

By definition, the generative grammar used in all these experiments is the same: (Q)AXB(R). However, our distributional manipulations across all of these experiments led human subjects under certain circumstances to restrict generalization to be maximally compatible with the input, while in other circumstances to generalize to the full grammar.

Table 1: Descriptions of the sampling bias in each experiment

Experiment	**Sampling bias**
Expt 1	Uniformly Distributed Gaps, Dense Sampling (1/3 withheld): Every X-word heard with every A- and B-word
Expt 2	Uniformly Distributed Gaps, Sparse Sampling (2/3 withheld): Every X-word still heard with every A- and B-word
Expt 3	Systematic Gaps, Sparse Sampling: Each X-word heard with a subset of possible A- and B-words
Expt 4	Extended Exposure to Systematic Gaps: Same as Experiment 3, but exposure was tripled
Expt 5	Subcategorization: Gaps were inserted such that a clear divide segregated X-words and contexts words into two subcategories
Expts 6-9	Same as Experiments 1-4, but included a very minimally overlapping X-word (X4); X4 seen in just one sentence frame in each condition
Expt 10	Same as Experiment 3, but bigram statistics are not balanced because words varied in frequency

Learners rated novel grammatical sentences as high as familiar grammatical strings in Experiments 1 and 2, showing a strong tendency to generalize across the words within a category. In Experiments 3 and 10, where a systematically-gapped training set was presented (balanced or not), learners became more conservative and treated novel grammatical sentences as somewhat less grammatical than familiar ones (but still more grammatical than ungrammatical ones). Generalization was further reduced in Experiment 4, when the exposure to systematic gaps was increased. In the subcategorization experiment (Experiment 5), learners did not fully generalize across the gaps created by the subcategory structure, indicating that they used the distributional information to learn that there were two subcategories within the X category. Lastly, the results of Experiment 6 showed that when learners were given a dense sampling of a language with almost complete overlap of contexts for several words in the X category, learners generalized a novel word (X4) to the full range of grammatical contexts of the other X-words, even when they heard X4 in only one of those contexts. When contexts were more sparse (Experiment 7) and there were significantly more systematic gaps in the input (Experiments 8 & 9), learners did not fully transfer their knowledge of X-category structure to the minimally overlapping X4 word. In all experiments, ungrammatical sentences were rated significantly lower than any novel grammatical test string.

Models

We use a generative model-based framework to develop our three models. The structures of these generative models make explicit the assumptions about knowledge representations. The goal of our modeling effort is to understand what elements must be included in the representation of the QAXBR grammar so that the models' generalization behavior will be

most compatible with human behavior across all 10 experiments. The answer to this question is related to the types of distributional cues that human learners attend to in the experiments. For the models reported in this paper, we make the simplifying assumption that learners only attend to local bigram information in the input, the bare minimum to capture the sequential dependencies within QAXBR sentences (although our models can easily be extended to use other distributional cues). A successful model should assign high probabilities to grammatical sentences and low probabilities to ungrammatical ones. Crucially, a successful model should assign probabilities to novel grammatical sentences that match the ratings of human learners.

Word Bigram Model

The first model is the *word bigram model*: the probability of a sentence is simply the product of the probabilities of its ordered word pairs, where the probability of each word w_i is conditioned on the preceding word w_{i-1}:

$$p(s) = \prod_{w_i \in s} p(w_i|w_{i-1}) \quad (1)$$

Equation (1) can be interpreted to suggest that a word bigram model represents the grammar with a set of multinomial distributions. Each distribution specifies the probabilities that a word will be followed by any other words in the vocabulary. The parameters of these distributions are typically estimated from training data with maximum likelihood estimation (MLE). However, the standard MLE algorithm is insensitive to sample size, which is a crucial variable of interest in several experiments. When comparing Experiments 3 and 4, for example, our subjects exhibited different generalization patterns as a result of the change in the amount of exposure to the same set of training data (i.e., a change in sample size only). Therefore, we adopt a Bayesian approach that is sensitive to sample size. The fully derived form for estimating the probability of a word is:

$$p(w_i|w_{i-1}, \text{all previous bigrams}) = \frac{n_{w_{i-1},w_i} + \beta}{\sum_k n_{w_{i-1},w_k} + v\beta} \quad (2)$$

where v is the vocabulary size, n_{w_{i-1},w_i} is the frequency of bigram (w_{i-1}, w_i), and β is a free parameter. The β parameter determines whether certain parameter settings of the multinomial distributions are favored. Here, we report results with β set to 1, which is a non-biased prior.

Simulation Procedure In each experiment, the word bigram model first estimates its model parameters according to the training sentences. In experiments where the length of exposure is a predictor of interest (Expts. 3, 4, 8 & 9), we duplicate the training data to simulate the effect of extended exposure. Unlike human subjects, however, the model is given information regarding the size of the vocabulary, and does not have memory limitations.

Word Bigram Mixture Model

The word bigram model implies that there is one single representation that corresponds to the grammaticality of a sentence. Natural languages, however, are usually more flexible: a sentence can have many different types of grammaticality (or ungrammaticality), such as Noun-Verb agreement, as well as lexical restrictions, such as transitive/intransitive verbs. We address this problem by developing a word bigram mixture model, where multiple patterns of grammaticality can be modeled simutaneously. Each component in the mixture is a word bigram model. A grammatical sentence is generated from a component grammar, which is in turn generated from a stochastic process (the model can be viewed as a Dirichlet process mixture model; Ferguson, 1973). We can describe the process of generating a sentence s in two steps:

- (a) $p(s \text{ is generated by an existing component } k) = \frac{n_k}{n+\alpha}$
  (b) $p(s \text{ is generated by a new component}) = \frac{\alpha}{n+\alpha}$
- If (a), $p(s = w_1, \ldots, w_m) = p(w_1, \ldots, w_m|\mathcal{B}_k)$
  If (b), $p(s = w_1, \ldots, w_m) = p(w_1, \ldots, w_m|\mathcal{B}_{new})$

where n_k is the number of sentences that have been generated as instances of component grammar k, n is the number of sentences that have already been generated, α is a free parameter of the model (a larger α leading to more new clusters), and $\mathcal{B}$ refers to the parameters of a component bigram model (as described in the previous section). Combining these two steps, the probability that s will be generated by a word bigram mixture model is

$$p(s = w_1, \ldots, w_n) = \frac{\sum_k p(s|\mathcal{B}_k) n_k + p(s|\mathcal{B}_{new})\alpha}{n + \alpha} \quad (3)$$

Simulation Procedures Equation (3) describes a generative model, with which we can assess the probability that a sentence is generated by an existing representation of the grammar. However, the learner faces the opposite problem: they must infer the representation given the observed sentences. We used the Gibbs sampling method to infer these parameters (the exact details are not described due to space limits). We run the model on the training data used in the experiments. The first 500 samples of each run are discarded (which may be biased towards initial values). Due to the small scale of our artificial language, the sampler converges quickly, well within the discarded 500 samples. Each of the remaining posterior samples is considered as a candidate representation of the grammar. For experiments with longer exposure, we also run the sampler longer to approximate the effect. The average probability that a sentence is generated by these posterior representations is taken as a measure of the grammaticality of the sentence.

Category Bigram Mixture Model

A notable feature of the two models presented so far is the lack of explicit representation for grammatical categories.

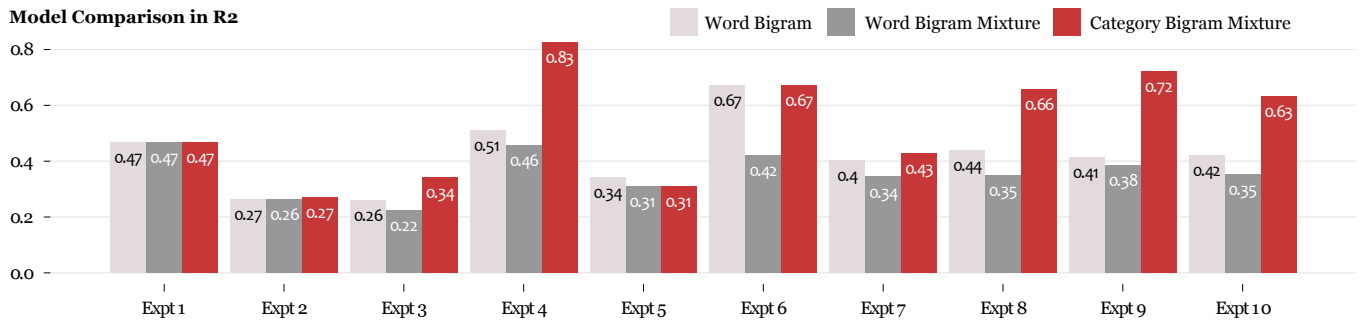

Figure 1: Grammaticality predictions made by the category bigram mixture model best approximate subject ratings in most experiments. R^2 is calculated by regressing subject ratings against model predictions.

In both models, bigram statistics are based on word tokens. However, a crucial component of language acquisition involves organizing words into grammatical categories and discovering relations between them. To investigate whether human learners were in fact organizing the words into categories, we also developed the category bigram mixture model. The category bigram mixture model preserves the notion of multiple component grammars and introduces a bigram-based word category discovery process nested within each component grammar. In other words, the component grammars in the category bigram mixture model are themselves infinite mixtures of bigram models on categories. Therefore, generating a sentence under the grammar is a two-step process with the second step containing another two-step process:

- (a) $p(s \text{ is generated by an existing component } k) = \frac{n_k}{n+\alpha}$
  (b) $p(s \text{ is generated by a new component}) = \frac{\alpha}{n+\alpha}$
- If (a), for the category of each word, c_i:
  - (i) $p(c_{i-1}, c_i \text{ belongs to existing bigram } l) = \frac{n_l^k}{n^k+\alpha_0}$
    (ii) $p(c_{i-1}, c_i \text{ is novel}) = \frac{\alpha_0}{n^k+\alpha_0}$
  - If (i), $p(w_i) \sim MultiNomial(c_i|c_{i-1})$
    If (ii), $p(w_i) \sim MultiNomial(c_{new}|c_{i-1})$

  If (b), for the category of each word, c_i:
  - (i) $p(c_{i-1}, c_i \text{ belongs to existing bigram } l) = \frac{n_l^{new}}{n^k+\alpha_0}$
    (ii) $p(c_{i-1}, c_i \text{ is novel}) = \frac{\alpha_0}{n^k+\alpha_0}$
  - If (i), $p(w_i) \sim MultiNomial(c_i|c_{i-1})$
    If (ii), $p(w_i) \sim MultiNomial(c_{new}|c_{i-1})$

where in the top-level process, n_k is the number of sentences that have been generated by component grammar k, n is the total number of generated sentences, α is the free parameter (as in the word mixture model) influencing the tendency of creating more component grammars. For clarity, we write c_{i-1}, c_i as the bigram label that each bigram l is associated with in the nested process: n_k^l is the frequency of category bigram l in component grammar k; n_k is the total number of category bigrams in component grammar k, and α_0 is a free parameter (a larger value leading to more category bigrams). Finally, the probability that each word is generated from its category c_I, conditioned on the category of its preceding word c_{i-1}, is modeled as a multinomial distribution.

Relation to other models The problem of discovering categories for word tokens in a language is analogous to the problem of part-of-speech tagging in computational linguistics, which has been under active research for several decades. Our category bigram mixture model is most similar to the Bayesian unsupervised tagging algorithm developed by Goldwater & Griffiths (2007). While our approach is not fully Bayesian (in the sense that hyper-parameters are treated as free parameters), it has the flexibility of discovering multiple part-of-speech sequence patterns (i.e. component grammars) and creating as many part-of-speech tags as needed (due to the nested Dirichlet Process).

Simulation Procedure As in the case of the word bigram mixture model, Gibbs sampling is applied to the inference problem to find samples of the posterior distribution. Each of the remaining posterior samples is considered as a candidate representation of the grammar under the category bigram mixture model. The average probability that a test sentence will be generated by these representations is taken as a measure of the grammaticality of the sentence.

Results and Discussion

Model predictions are in the format of probability estimates. A higher probability estimate means that a sentence is more grammatical. The quality of model predictions is determined by examining how well they correlate with subject ratings. To ensure that subject ratings are maximally comparable with model predictions, we transformed discrete ratings into z-scores within each subject and experiment, so that the ratings of subjects with consistent biases (consistently high or consistently low ratings) were normalized. We computed the R^2 metric for each group using a linear regression where model predictions were used to predict subject ratings. A model with a high R^2 indicates that the particular model explains a significant amount of variance in subject ratings (see Fig 1). Overall, the category bigram mixture model best captures hu-

man behavior across all 10 experiments combined (R^2 Word Bigram = 0.4, R^2 Word Bigram Mixture = 0.35, R^2 Category Bigram Mixture = 0.47).

The general advantage of the category bigram mixture model suggests that our human learners may have acquired a representation of an X category, and not just a set of simple word co-occurrences. In X4-related experiments (Expts. 6-9), we asked whether learners could extend their knowledge of a target category to a very infrequently presented word for which they only had minimal context information. We found that there was a point in learning where hearing just one context for the minimally overlapping X4 word was enough to generalize full category privileges for that word (Expt 6). Simple word co-occurrence and bigram counts will not achieve this outcome. The category bigram mixture model, however, has the appropriate representation for supporting such a learning outcome. Indeed, in Experiment 6, X4 gets assigned to the same category as all other X-words, thus enabling the generalization to novel X4 sentences (the effect of X4 sentences on overall R^2 is reduced by the extremely small number of X4 sentences in the testing phase).

Limitations of the category bigram mixture model

While the category bigram mixture model best approximates human generalization patterns across the 10 experiments, it does no better than the other two models in capturing human performance in the subcategorization experiment. Indeed, the two mixture models acquire the subcategory structure, but fit human performance no better than the simplest word bigram model. This paradoxical result is due to the experimental design: all bigrams in the training subset conform to the subcategory boundaries and are presented equally often. At test, novel subcategory-conforming items are rated as high as familiar ones because they contain only bigrams that have been presented (thus indistinguishable from familiar ones). Test strings violating the subcategory structure are rated low by the word bigram model simply because they contain one or two bigrams which are never seen in the training data. The balanced presentation of all within-subcategory bigrams enables the word bigram model to distinguish between subcategory conforming and violation items without learning the existence of two subcategories. As a result, even though the two mixture models successfully discover the existence of two subcategories, the additional advantage of such discoveries is minimal.

The category bigram mixture model also tends to overgeneralize in experiments with systematic gaps. This is most clearly demonstrated in Experiments 4 and 9 (see Fig 2). In those experiments, subjects were exposed to a language with frequent systematic gaps in the input. Human learners gave novel grammatical sentences a significantly lower rating than familiar grammatical strings, especially when the training materials were presented multiple times. We view this restriction of generalization as a rational behavior that prevents human learners from over-generalizing when systematic and persistent gaps occur in the input.

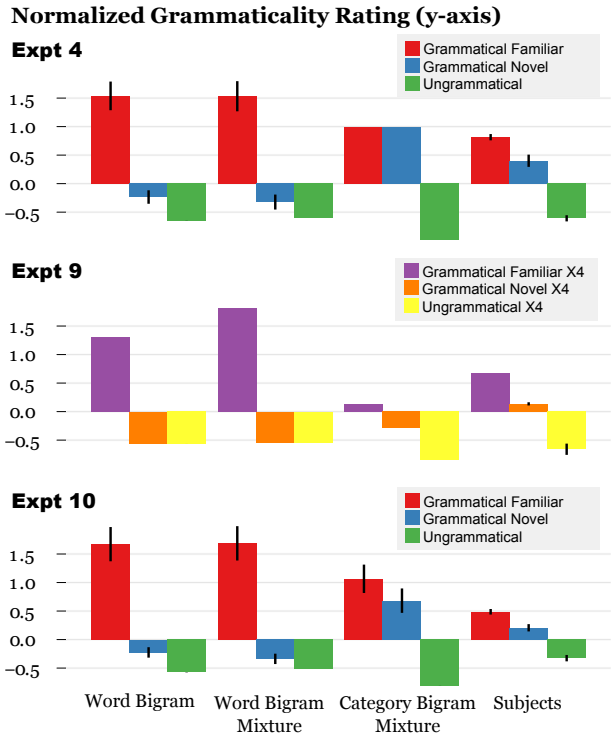

Figure 2: Z-score normalized grammaticality ratings by models and human subjects in 3 experiments where the Category Bigram Mixture model had highest R^2. Error bars = SE. Regarding the difference between familiar and novel ratings, Expt 4 shows overgeneralization by the category bigram mixture model, Expt 9 shows overly conservative behavior of the category bigram mixture model, and Expt 10 shows the category bigram mixture model best capturing human behavior, despite systematic gaps and variable bigram frequencies.

In the word-based models, the reduced generalization is captured because an increase in the probabilities of observed word bigrams necessarily leads to a decrease in the probabilities of unobserved ones, thus producing restricted generalization. However, this effect is much weaker in the category bigram mixture model, where repeated exposure to training sentences only strengthens the category bigram dependencies. When a novel grammatical sentence is presented to this model, its category bigrams have been observed many times during training and the sentence will receive a relatively high rating as a result (despite this, R^2 is relatively higher in the category bigram mixture model because its predictions are qualitatively closer to subject ratings; see Fig 2). This is an indication that learners have slightly different constraints on learning and/or a slightly different strategy from the category bigram mixture model. We are currently exploring other possible models that build on the idea of an underlying category representation, but incorporate learning constraints that more closely mimic human learning (e.g. incremental models of learning) and lead to the construction of a more restrictive grammar that is still compatible with the input.

This pattern can be contrasted with model performance on Expt 10, where not all grammatical bigrams are seen equally often during exposure. Results from this experiment make clear that the category bigram mixture model is the most robust to manipulations of the bigram distribution (see Fig 2). The other two models rate grammatical novel sentences almost as low as ungrammatical sentences, since novel grammatical sentences contain novel and low-frequency bigrams. By definition, these are less grammatical to the word-based models due to having a lower probability. The category mixture model, on the other hand, is not negatively influenced by the unbalanced design due to the abstraction of word categories.

General Discussion

Across 10 experiments, we compared the grammaticality predictions of three different models to human subject ratings. Our primary interest was to find the representational elements of the grammar that are most compatible with the generalization behaviors displayed by the learners. Generalization depends on the ability to abstract over categories, which is fundamental to linguistic productivity. A number of researchers have asked whether there is adequate distributional information in the input to form linguistic categories. Previous work uses hierarchical clustering and a computational learning mechanism to attempt to deduce grammatical categories from corpora of child directed speech based solely on distributional analyses of the input (e.g. Mintz et al., 2002; Redington et al., 1998). These models have been able to use co-occurrence statistics among words to achieve relatively good categorization performance for frequent target words, indicating the utility of these types of distributional cues for categorization.

The behavioral experiments that this work is built upon suggest that the patterning of word tokens in a substantial corpus of linguistic input appears to be sufficient to extract the underlying structural categories in a natural language, given an appropriately capable learner. Our modeling results have further explicated the representational assumptions for extracting the knowledge of a grammar. Of the three models, two models are based on simple word bigrams collected from training data. While word bigrams are useful for capturing the lexical dependencies of the grammar, they cannot explain how human learners could generalize from experienced examples to novel items, especially when the prior experience is minimal (i.e., Expts 6-9). Such rapid and automatic generalization behavior calls for a richer representation, in which the grammar of the artificial language is organized around potential categories of vocabulary words. The category bigram mixture model introduces the notion of categories as a representational assumption, which led to model predictions that better approximated the behavior of human learners in almost all experiments. A limitation of the category bigram mixture model, however, is that it overgeneralizes compared to human performance. A fourth type of model could add a generative component that asks how likely it is that a string is absent given a random sampling process. If that probability is low, then it would penalizes the probability even further by downweighting it in the grammar. We are exploring this direction, in conjunction with other mixture models that may more closely mirror the constrained learning environment that human learners face during natural grammatical category acquisition.

Acknowledgments

This research was supported by NIH Grants HD037082 to RNA and DC00167 to ELN, and by an ONR Grant to D. Bavelier at the University of Rochester.

References

Bowerman, M. (1973). Structural relationships in childrens utterances: Syntactic or semantic? In T. Moore (Ed.), *Cognitive development and the acquisition of language.* Harvard University Press.

Braine, M. (1987). What is learned in acquiring word classes a step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (p. 65-87). Lawrence Erlbaum Associates.

Ferguson, S. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*, 209-230.

Goldwater, S., & Griffiths, T. (2007). A fully bayesian approach to unsupervised part-of-speech tagging. In *ACL.*

Maratsos, M., & Chalkley, M. A. (1980). The internal language of childrens syntax: The ontogenesis and representation of syntactic categories. In K. Nelson (Ed.), *Childrens language* (Vol. 2). Gardner Press.

McNeill, D. (1966). Developmental psycholinguistics. In *The genesis of language: A psycholinguistics approach* (p. 69-73). MIT Press.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, *26*, 393-425.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 435-469.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2009). The role of distributional information in linguistic category formation. In *CogSci 2009* (p. 2564-2569).

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2010). Novel words in novel contexts: The role of distributional information in form-class category learning. In *CogSci 2010* (p. 2063-2068).

Reeder, P. A., Newport, E. L., & Aslin, R. N. (in review). *From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes.*

Schuler, K., Reeder, P. A., Newport, E. L., & Aslin, R. N. (in prep). *The effects of uneven frequency information in linguistic category formation.*