Principles of Generalization for Learning Sequential Structure in Language

Michael C. Frank, Denise Ichinco, and Joshua B. Tenenbaum
{mcfrank, ithink, jbt}@mit.edu
Department of Brain and Cognitive Sciences, 43 Vassar Street
Cambridge, MA 02139 USA

Abstract

How do learners discover patterns in the sequential structure of their language? Infants and adults have surprising abilities to learn structure in simple artificial languages, but the mechanisms are unknown. Here we introduce a rule-based Bayesian model incorporating two principles: minimal generalization and representational parsimony. We apply our model to tasks in artificial language learning and inflectional morphology and show that it fits behavioral results from infants and adults and learns inflectional rules from natural data.

Keywords: Language acquisition; generalization; artificial language learning; inflectional morphology; Bayesian modeling.

Introduction

How do learners discover patterns in the sequential structure of their language? Experimental work on the unsupervised learning of sequential structure has suggested that infants and adults have access to flexible and powerful learning mechanisms which may be involved in language acquisition (Gomez, 2002; Marcus et al., 1999). However, both the particular mechanisms involved in these tasks and the aspects of acquisition to which they apply are at present unknown.

In our current work we attempt to address these questions by creating a computational model which embodies two principles suggested by this experimental literature: minimal generalization and representational parsimony. We show that these principles apply not only to artificial language tasks, but that they may also have applications to learning inflectional morphology, an important task facing language learners.

We first describe our model and how it embodies a tradeoff between these two principles within a hypothesis space expressive enough to capture many different types of rules. We next show how our model can be applied to artificial language experiments on learning identity-rules (Gerken, 2006; Marcus et al., 1999) and non-adjacent dependencies (Gomez, 2002). We then present an extension of our model to the case of inflectional morphology. Finally, we show preliminary data indicating that our model can be applied directly to learning inflectional rules in natural language.

The representations and learning mechanisms involved in the acquisition of inflectional morphology have been hotly debated in the literature on language acquisition. Two basic positions have been proposed: a single process of analogical learning (Rumelhart & McClelland, 1986) or a dual system consisting of both abstract rules and associative processes (Pinker, 1991). While this debate has been taken as representative of a wider debate over the format of mental representation, it has nevertheless tended to confound a number of independent computational issues.

Proponents of analogical or associative theories have emphasized the parsimony and neural plausibility of this type of proposal. In contrast, dual-route theorists have focused on representational or expressive limitations of the analogical approach. There are two dissociable issues captured by this debate: (1) the number of routes for morphology learning and (2) the algorithmic form and expressive power of those routes. For instance, recent work by Albright & Hayes (2003) compared an analogical model with a rule-based model and found that the greater expressivity of the rule-based model allowed for tighter generalization and better fit to human experimental data in a novel-word inflection wug task, despite the fact that both models had only one route for representation.

Under a more general definition of a rule as a systematic regularity, rules can be both broad (as in the regular rule for the past tense in English orthography, $+ed$, and narrow (as in the past tense rule for the verb go: $go \rightarrow went$. Within an expressive enough hypothesis space, a rule could even be formulated for analogical inferences like using inflections from stems with high similarity.[1] If we assume that the hypothesis space of rules is broad enough to capture many different types of regularities, the problem of how to find the right rule within this hypothesis space becomes more important.

Our current work is not directly concerned with the exact form of the representations used by human learners. Instead, we assume that learners are attempting to make generalizations from limited data within some hypothesis space and focus on the principles by which they find the best generalizations in that space. Following Albright & Hayes (2003), our hypothesis space consists of sets of explicit rules, both for their ease of interpretation and because Albright and Hayes' data show that this kind of representation provides a better fit to human generalizations. However, we take rules to be a representational convenience which we adopt at the highest of Marr's (1982) levels of analysis: the level of computational theory. Thus, we focus here not on testing different kinds of representations, but instead on making explicit and individually testing the principles of generalization by which particular rules are learned.

Model Design

We formalize the idea of a rule as a set of restrictions on the features of a string. For instance, Marcus et al. (1999) presented infants with strings like *wo fe fe* (three-syllable strings where the last two syllables were the same). In our

[1] For instance, though the hypothesis space of our current model does not allow similarity-based rules (e.g., "strings within some edit distance of X"), it would be relatively simple to add such rules.

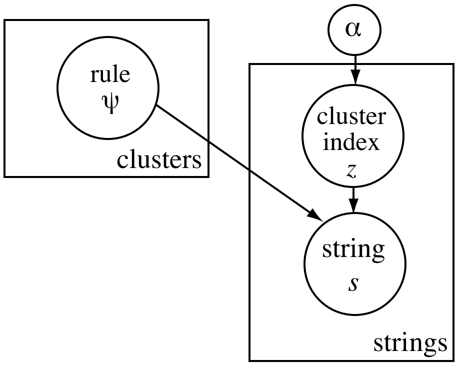

Figure 1: A schematic representation of the generative process for our model. This process defines a distribution over sets of strings by showing how they could be generated by a set of underlying rules. In practice, we invert this process through Bayesian inference, calculating the posterior probability that an observed set of strings was generated by a particular set of rules. Circles represent variables, arrows represent dependencies, and rectangles (plates) group sets of elements that are repeated.

F1	F2	F3	F12	F13	F23	Translation
wo	*fe*	*fe*	*	*	*	only *wo fe fe*
*	*fe*	*fe*	*	*	*	ends *fe fe*
wo	*	*fe*	*	*	*	begins *wo*, ends *fe*
wo	*	*	*	*	=	*wo BB*
*	*	*	*	*	=	*ABB*
*	*	*	*	*	*	any string

Table 1: Some of the rules consistent with the string *wo fe fe*, from Marcus et al. (1999). F1, e.g., refers to those features which describe the first element of the string, while F23, e.g., refers to features that relate the second and third element of the string. A * denotes no restriction on a particular feature.

models of artificial language tasks, we define features over both individual elements (syllables, phonemes, or words depending on the experiment) and pairs of elements (as in Table 1). For individual elements, our features simply denote whether an element has a particular value (e.g., the first syllable is *wo*). For pairs of elements, we restricted our hypothesis space to contain a single binary feature: the identity relationship in which two elements have exactly the same value regardless of what it is. By picking combinations of features we can generate more complex rules like *AAx* (the first two elements are the same and the third is *x*)—the rule used by Gerken (2006).[2] The conjunction of a set of features makes a rule; a rule is true of a particular string only if the string contains all the features included in that rule.

The goal of our model is to find one or a small number of rules which tightly describe the available data. Imagine the set of strings *abc abd abe*. One description of these strings might be "*a* followed by any two letters." However, intuitively it seems as though the less general rule, "*ab* followed by any one letter" is more likely. Our model formalizes this principle of minimal generalization, known as the "size principle" (Tenenbaum & Griffiths, 2001) by assigning probabilities to rules depending on how tightly they fit an observed set of strings.

Consider a second set of strings: *abc abd abe mnp mnq mnr*. Again, one description of this set of strings might be "any set of three letters." But there is a less general description: "*ab* followed by any letter," or "*mn* followed by any letter." In this case, the better description seems to contain two specific rules rather than one more general rule. Taking this principle to its logical conclusion, however, produces a very unparsimonious set of rules: "*abc*," "*abd*," "*abe*," "*mnp*," "*mnq*," or "*mnr*." Though this description is very specific, it includes too many rules and fails to identify the generalization linking subsets of the strings together. We formalize this intuition by including a prior on the number of rules used to describe a set of strings. This prior is known as the Chinese Restaurant Process (CRP) (Rasmussen, 2000).

The Bayesian framework we use here gives us a principled method for trading off minimal generalization (which prefers more specific rules, even if there are more of them) and representational parsimony (which prefers fewer rules, even if they are more general).

Model details

A generative process (such as the one in Figure 1) is a sequence of steps which jointly define a probability distribution. By defining our model generatively we can use Bayesian inference to calculate the posterior probability that a set of unobserved states—in our case, a set of rules and clusters—generated an observed product: a set of strings.

Following the arrows in Figure 1, in order to generate a string, we first decide what cluster c it belongs to (each cluster has one rule associated with it) by giving it a cluster index z. If this is the first string we have generated, then the string must go in its own cluster; if we have generated some strings already, we can decide whether the new string will fall in one of these pre-existing clusters or go in a cluster of its own. This process, the CRP, is governed by a concentration parameter α which controls how likely a new string is to go in its own cluster.

Once we have decided on which cluster the string belongs to, we then either use the rule ψ already assigned to that cluster or—in the case of a new cluster—randomly pick a rule to go with it out of the space of rules Ψ. We then pick a string s

[2] In this type of hypothesis space it is possible to define inconsistent rules (e.g.,"first element is *wo*, second element is *fe*, and first and second elements are the same"). We deal with this by excluding inconsistent hypotheses from consideration and renormalizing the probability of the remaining rules in the hypothesis space.

uniformly from the set of strings that are consistent with ψ.

Formally, the joint probability of a full corpus of strings S and a partition Z of those strings into rule clusters is given by

$$P(S,Z|\alpha) = P(S|Z) \cdot P(Z|\alpha) \tag{1}$$

The probability $P(Z|\alpha)$ of a partition is given by the CRP with concentration parameter α. The probability of the corpus given the cluster assignments is the product of independent terms for each string (corresponding to the plate over strings in Figure 1):

$$P(S|Z) = \prod_i P(s_i|z_i) \tag{2}$$

Because strings in each cluster c are generated by a particular rule ψ_c for that cluster, we group the terms in Equation 2 into a product over clusters and then a separate product over strings in that cluster:

$$P(S|Z) = \prod_c \prod_{i:z_i=c} \sum_{\psi_c} P(s_i|\psi_c) \cdot P(\psi_c) \tag{3}$$

However, because ψ_c is not known, in computing the probability of observing the strings associated with cluster c we integrate the predictions of all rules congruent with the strings in the cluster, weighted by their prior $P(\psi_c)$. For simplicity we take this prior to be uniform, equal to the inverse of the number of possible rules in our description language. The likelihood function for a rule is then given by

$$P(s_i|\psi_c) = \begin{cases} \frac{1}{|\psi_c|} \text{if } s_i \text{ consistent with } \psi_c \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

For each string, this probability is simply the probability of a string being chosen uniformly from the set of strings consistent with the rule for that cluster (the size principle). Put another way, the size of a rule $|\psi|$ is given by the rule's extension: the number of ways the symbols—syllables, letters, or phonemes—of a language can be combined that are congruent with the rule. Since larger rules will be less likely to generate a particular example, our model favors minimal generalization by giving highest probability to the smallest rule that could have generated the observed data. The CRP prior in turn ensures representational parsimony by giving higher probability to hypotheses with fewer different rules, whatever their size.

Because we pose our model as a generative process, we are able to invert that process using Bayes' rule and compute the posterior probability of a hypothesis (a partition of strings into clusters and rules to go with those clusters) given a set of strings. In practice, we use a Gibbs sampler to search for the best partition of strings into clusters (MacKay, 2003).

All simulations were conducted using types rather than tokens. Accordingly, only one example of a particular string was included in the training set for our model, even if strings were presented multiple times in the original experiment. Since we used the size principle to determine the likelihood of a particular rule, we made this choice because using tokens rather than types would make a rule less probable with each repetition of the same string (intuitively, an undesirable consequence). One possible extension of our model to deal with this issue would include another step in the generative process which generated tokens from types (Goldwater et al., 2006).

sample rule	# of clusters	$\alpha = .9$	.09	.009
ABA	1	**-75.70**	**-73.47**	**-73.21**
le B le	4	-77.84	-82.51	-89.15
A di A	4	-83.38	-88.05	-94.70
le di le	16	-112.59	-144.89	-179.16

Table 2: Log posterior probability of different hypotheses (shown with an example of the maximum likelihood rule for one of the clusters for that hypothesis along with the total number of clusters in that hypothesis) under different CRP parameter values. While the absolute probability of the different clusterings changes relative to the value of α, the single cluster/rule hypothesis was always preferred.

Experiment 1: Learning Identity Rules

In our first set of simulations we ran our model on the stimuli from two sets of experiments on artificial rule learning with infants. The first were those of Marcus et al. (1999), who familiarized infants to rules of the forms *ABB*, *AAB*, and *ABA* and tested them on their ability to discriminate strings of this form from strings of an alternate form (e.g., *ABB* vs. *AAB* as in the example above). The Marcus et al. stimuli were composed of a vocabulary of eight syllables, of which 4 were designated as *A* elements and 4 were designated as *B* elements, creating a total set of 16 tokens.

The second set of stimuli came from Gerken (2006), who tested infants on sets of four strings drawn either from an *AAB* rule or a narrower *AAx* rule (the first two elements the same followed by *x*). Gerken found that even though both sets of strings were consistent with the broader *AAB* rule, infants showed evidence of learning the *AAx* rule when all the evidence was consistent with the narrower generalization.

For each of the three Marcus et al. (1999) rules and for the two Gerken (2006) rules, our model assigned the highest posterior probability to the hypothesis that all the strings were generated by the same rule; the rule with the highest likelihood for this cluster was the rule posited by the researchers (e.g., *ABA*). Although the likelihood of a partition of the strings into several more specific rules, e.g. "*le B le*," "*wi B wi*," "*ji B ji*," or "*de B de*" was higher, the prior was considerably lower, leading to a consistent preference for the single cluster hypothesis (Table 2).

Experiment 2: Using Variability to Generalize

Does increasing type variability strengthen generalizations? Gomez (2002) presented learners with three-word strings containing an invariant dependency between the first and third

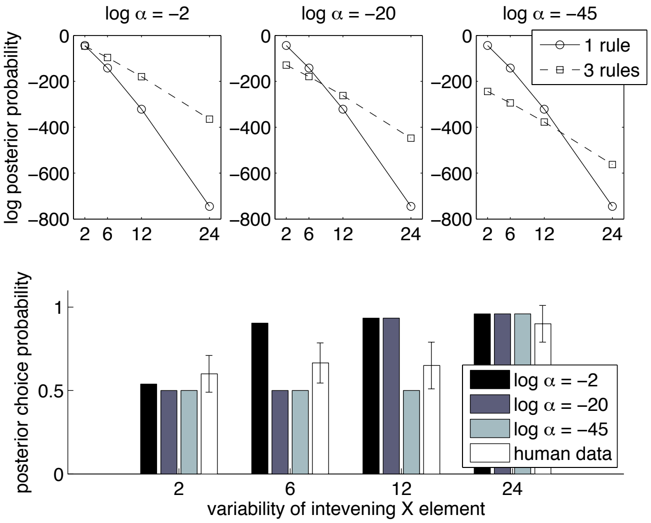

Figure 2: Results from our simulations of experiments by Gomez (2002). Top: log posterior probability of clusterings with either 1 or 3 rules (corresponding either to a learn nothing rule or a correct generalization) at three different values for α, the CRP parameter. Bottom: model performance at test compared with human data reported by Gomez (2002). Error bars show standard error of the mean.

elements (e.g., generated by the rule *aXb*, where the identity of the *X* element varied). They manipulated how many elements could appear in the *X* position of the string (data shown in Figure 2) and found that participants were able to learn the specific rules of the language only when the variability of the *X* element was greater than 12 elements, concluding that variability in adjacent dependencies might lead to greater attention to non-adjacent dependencies.

We tested whether our principle of representational parsimony (embodied in the CRP prior on the number of rules the model learns) could be responsible for the results they observed. We ran our model on the same set of strings and found that the model showed the same qualitative tradeoff as participants, switching between parsimony of representation and minimal generalization by learning only a single rule (“accept any string”) for $|X| = 2$, but quickly moving to the correct generalization (learning three rules, “*a_b*,” “*c_d*,” or “*e_f*”) for variability greater than 2.

Why does the model prefer to generalize at such a low rate of variability compared with the human participants? One reason might be the memory limitations of human learners: perhaps human learners can only appreciate some of the evidence for a particular inference at any given time. In other work, we have used a memory decay function over tokens to simulate this kind of limited use of evidence. Given the simplicity of the current experiment, however, we chose to simulate the limited use of evidence by lowering the α parameter on the CRP until the model strongly dispreferred hypotheses with more rules. We then modeled the forced-choice task of human participants by calculating the probability of choosing a correct string over an incorrect string via a Luce choice rule (Luce, 1963) comparing the posterior probability of correct and incorrect strings under the model (Figure 2). While the level of variability at which the model was able to discriminate strings correctly varied widely with different values of α, the qualitative trend remained constant: a tradeoff between preferring representational parsimony (prior) with less evidence and minimal generalization (likelihood) as the amount of available evidence increased.

Experiment 3: Artificial Inflectional Morphology

In order to test the performance of the model in fitting a more complex range of human data (including production data as opposed to forced choice accuracies), we conducted a simple experiment with adults using artificial morphological stimuli.

Experimental paradigm

Participants Twenty-one participants from the MIT community participated for payment.

Materials and Methods Participants were told they were learning about the language of a remote island and were given sets of 20 index cards (each of which had on it a noun from the island’s language paired with its plural form). They were told that they could spread out the cards and rearrange them any way they wanted in order to learn the language best. They were then given a sheet with fifteen novel nouns and asked to fill in the plural form for each noun and give a confidence rating. Participants in all three conditions received the same test materials.

Participants saw index cards from one of three conditions, which we called multiple rules, reduplicative rules, and rule plus exceptions. In all conditions, rules were suffix rules which required adding material to the end of a stem; no rules conflicted in their application—in both the training and the test sets, only one rule applied to each stem. Stems were multi-syllabic, pronounceable non-words that did not sound recognizably English-like.

In the multiple rules condition, we defined five rules, each of which was attested in four examples. Each rule applied only to stems with a particular ending: for instance, +*ene* / __*ij* (“add *ene* if the word ends with *t*”) was a rule that applied to the words *gimij*, *varij*, *ipij*, and *haspadij*.

In the reduplicative rules condition, there were two rules, each with 10 examples: +*em* and *reduplicate last syllable* / __a_ (“repeat the last syllable of words with an a in the second-to-last position,” as in the stem-inflected form pair *vigutap* → *vigutaptap*).

The final condition was the rules-plus-exceptions condition, in which participants saw 17 examples of one suffix rule, two examples of another, and then a third irregular form, meant to simulate a system like the English plural or past tense where there is an overwhelmingly frequent rule with only a relatively small number of exceptions.

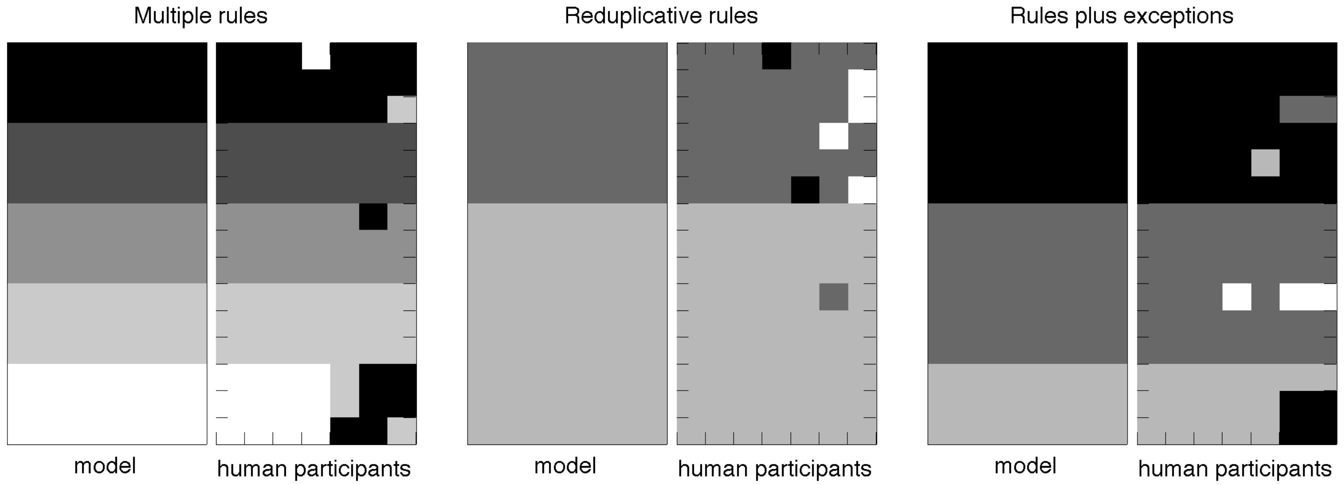

Figure 3: For each condition of Experiment 3, we show the clusters found by our model (left side) and the clusters in participants responses (right side). Each inflection (e.g., $+em$) was given a different grayscale value. For the human participants, the 15 rows in each plot represent the items in the generalization test and the 7 columns represent the 7 participants in each condition.

Results Average pairwise similarity between participants in the multiple rules condition was 81.9%; in the reduplicative rules condition, 83.4%; and in the rules plus exceptions condition, 86.8% (Figure 3). These pairwise similarities differed significantly from chance (computed via permutation of participants' responses): all ps $< .0001$, all ts > 18.

Inflectional Model

To adapt our model to inflectional data, we added a step to the generative process. Each cluster was assigned both a rule schema (what we referred to as a rule in the initial model: a set of features) and an inflectional rule (a procedure for modifying a stem—what we called a string in the initial model—to create an inflected form). The rule schemata in this model were defined over phonemes and positions in the stem (counting backwards from the end of the stem). For instance, possible features could be the last phoneme is *e* or the second to last phoneme is *t*.

Inflectional rules were defined as a set of deterministic transformations to be performed on each stem. For instance, if the inflectional rule were $+ed$ / $\_\_t$, the full rule (schema and inflectional rule) would be consistent if all stems in the cluster ended with *t* and all inflected forms were suffixed by *ed*. In practice we included four possible operations for inflection, which could be combined as necessary to create the proper inflected form: adding a suffix, reduplicating a suffix, substituting a vowel, and substituting an entire word. Because the space of rules in this and the next experiment was larger than the space in the first two experiments (due to the greater length of strings), we calculated only the highest-probability schema and inflectional rule for each cluster.[3]

Fit to data

To test the fit of our model to the data we collected, we ran the model on each training set. For each of the three conditions, the clustering with the maximum posterior probability was the one we intended; the maximum likelihood rule (schema and inflectional rule) for each cluster similarly matched our intended design. To model generalization to novel test items in each condition, we chose the maximum a posteriori hypothesis in the model and used it to generate the maximum a posteriori inflected form for each stem (Figure 3).

The sets of rules preferred by the inflectional models produced generalizations that were highly similar to those of our human participants (and performance was robust to manipulation of the CRP parameter α). In the multiple rules condition, the model produced the same form as the human participants in 93 of 105 cases (88.6%); in the reduplicative rules condition, 98 of 105 cases (93.3%); and in the rules plus exceptions condition, 95 of 105 cases (90.5%). In each of the three experiments, three participants produced exactly the same pattern of judgments as the model while the other four differed by no more than 4 of 15 judgments. These results suggest that the model effectively recovered the same structure from the training data as the human participants.

Experiment 4: Natural Morphology

In order to test the generality of the inflectional form of our model, we applied the version described in the previous section to the problem of learning the English past tense. We carried out preliminary simulations using a phonetically-

[3] It was not possible to calculate the number of phonetically legal words congruent with a particular schema (as we did in the artificial language examples). To compute the likelihood of a string given a rule (Equation 4), we approximated the size of a schema by counting the number of words in the training data that were congruent with that schema. Provided that the training data is a representative sample of the overall corpus, the relative values of $|\psi|$ should be comparable using this estimate.

Frequency	Rule	Example stem (past)
63	+d / __p__	appear (appeared)
23	+əd / __t	want (wanted)
12	+d	show (showed)
11	+əd / __d	need (needed)
9	+t / __k	look (looked)
9	+t / __s	increase (increased)
8	+t / __p	stop (stopped)
5	+t / __ ʃ	watch (watched)
4	Ø / __t	put (put)
3	o → u / __o	know (knew)
	…	
1	go → wɛnt	go (went)
1	gɛt → gat	get (got)

Table 3: A sample of the most frequent rules found by the inflectional model (Experiment 4) when run on the 200 most frequent English present-past verb form pairings.

transcribed corpus of present-past verb form pairs.[4]

We trained the model on the 200 most frequent past-tense phonological forms in English. Results for the most frequently applied rules are shown in Table 3. Since our model was only able to restrict particular elements of a string to one value (rather than a class of values, e.g., unvoiced phonemes), it was not able to capture the specific selectional regularities of the English past. Despite this, when we tested it on its generalization to the other forms in the corpus, it successfully inferred the correct form 88.5% of the time (1753 of 1981 forms correct). Despite the limits on the hypothesis space for rule schemata, the rules that the model learned on this small training set were similar to those that might be written in a phonology text (Table 3). In future work we hope to further increase the expressiveness of our hypothesis space in order to evaluate the generalization performance of the model.

General Discussion

On the basis of previous experimental work, we proposed two principles for sequential generalization in language: minimal generalization and representational parsimony. We formalized these principles in a Bayesian model. In Experiment 1, we showed that the principle of minimal generalization allowed our model to fit data on infant rule learning. In Experiment 2, we showed that increasing evidence via variability gave greater support for generalization, suggesting a bias for representational parsimony. In Experiment 3, we further tested these principles by altering our model to handle inflectional tasks and comparing this new model with human performance on three simple artificial inflection systems. We found a tight correspondence between the performance of the model and the productions of human participants. Finally, in Experiment 4, we ran our model on a subset of English past tense forms and found that the model acquired linguistically plausible rules and generalized them with relative accuracy. Taken together, these data suggest that the general principles of minimal generalization and representational parsimony—combined within an expressive hypothesis space—may be sufficient to account for a wide range of phenomena in sequential linguistic generalization.

Acknowledgments

The first author was supported by a Jacob Javits Grant. The authors thank Adam Albright, Noah Goodman, Vikash Mansinghka, Steve Piantadosi, Gary Marcus, and Ed Vul for valuable discussion.

References

Albright, A., & Hayes, B.(2003). Rules vs. analogy in english past tenses: a computational/experimental study. *Cognition*, *90*(2), 119-161.

Gerken, L. (2006). Decisions, decisions: infant language learning when multiple generalizations are possible. *Cognition*, *98*(3), B67-B74.

Goldwater, S., Griffiths, T. L., & Johnson, M.(2006). Interpolating between types and tokens by estimating power-law generators. *Advances in Neural Information Processing Systems*, *18*.

Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431-436.

Luce, R. D.(1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology.* New York: Wiley.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York: Henry Holt and Co., Inc.

Pinker, S.(1991). Rules of language. *Science*, *253*, 530-535.

Rasmussen, C. E. (2000). The infinite gaussian mixture model. *Advances in Neural Information Processing Systems*, *12*.

Rumelhart, D. E., & McClelland, J. L.(1986). On learning the past tense of english verbs. In J. McClelland & D. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629-640.

[4] Corpus data were obtained from the website of Bruce Hayes (http://www.linguistics.ucla.edu/people/hayes/learning/).