Cognition 120 (2011) 360–371

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Three ideal observer models for rule learning in simple languages

Michael C. Frank[a,*], Joshua B. Tenenbaum[b]

[a] Department of Psychology, Stanford University, United States
[b] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, United States

ARTICLE INFO

Article history:
Available online 4 December 2010

Keywords:
Language acquisition
Artificial language learning
Bayesian modeling
Generalization
Infant development

ABSTRACT

Children learning the inflections of their native language show the ability to generalize beyond the perceptual particulars of the examples they are exposed to. The phenomenon of "rule learning"—quick learning of abstract regularities from exposure to a limited set of stimuli—has become an important model system for understanding generalization in infancy. Experiments with adults and children have revealed differences in performance across domains and types of rules. To understand the representational and inferential assumptions necessary to capture this broad set of results, we introduce three ideal observer models for rule learning. Each model builds on the next, allowing us to test the consequences of individual assumptions. Model 1 learns a single rule, Model 2 learns a single rule from noisy input, and Model 3 learns multiple rules from noisy input. These models capture a wide range of experimental results—including several that have been used to argue for domain-specificity or limits on the kinds of generalizations learners can make—suggesting that these ideal observers may be a useful baseline for future work on rule learning.

1. Introduction: from "rules vs. statistics" to statistics over rules

A central debate in the study of language acquisition concerns the mechanisms by which human infants learn the structure of their first language. Are structural aspects of language learned using constrained, domain-specific mechanisms (Chomsky, 1981; Pinker, 1991), or is this learning accomplished using more general mechanisms of statistical inference (Elman et al., 1996; Tomasello, 2003)? Recent experiments have provided compelling demonstrations of the types of abstract regularities that can be learned from short exposures to novel language stimuli (Gómez, 2002; Gómez & Gerken, 1999; Marcus, Vijayan, Bandi Rao, & Vishton, 1999; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Smith & Yu, 2008), suggesting that characterizing the learning mechanisms available to infants may lead to progress in understanding language acquisition more generally.

Experiments on generalization have provided particularly important evidence for these learning abilities, which in turn may be relevant to the acquisition of complex linguistic structures. In one experiment, Marcus et al. (1999) familiarized seven-month-olds to 2 min of syllable strings conforming to abstract rules like *ABA* (e.g., *ga ti ga*) or *ABB* (e.g., *ga ti ti*). When tested using a head-turn preference procedure, infants showed a preference for strings that violated the rule they had heard over strings that conformed to that rule, even when both sets of test strings were generated from syllables that had not yet been heard.

These experiments suggested that infants could abstract away from the perceptual particulars of the syllables in the familiarization sequence and learn something like an abstract rule, but they left many questions unanswered. What sort of rule do infants learn—for instance, a rule focusing on identity, like "first syllable is the same as the

* Corresponding author. Address: Department of Psychology, Stanford University, 450 Serra Mall, Jordan Hall (Building 420), Stanford, CA 94305, United States. Tel.: +1 650 724 4003.
E-mail address: mcfrank@stanford.edu (M.C. Frank).

0010-0277/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2010.10.005

third syllable," or one focusing on difference, like "second syllable different from third syllable"? From among the many rules that are consistent with the data, how do learners decide which should guide generalization? Do learners even acquire a "rule" at all, or instead, some kind of subsymbolic summary?

Subsequent studies of rule learning in language acquisition have addressed all of these questions, but for the most part have collapsed them into a single dichotomy of "rules vs. statistics" (Seidenberg & Elman, 1999). The poles of "rules" and "statistics" are seen as accounts of both how infants represent their knowledge of language (in explicit symbolic "rules" or implicit "statistical" associations) as well as which inferential mechanisms are used to induce their knowledge from limited data (qualitative heuristic "rules" or quantitative "statistical" inference engines). Formal computational models have focused primarily on the "statistical" pole: for example, neural network models designed to show that the identity relationships present in *ABA*-type rules can be captured without explicit rules, as statistical associations between perceptual inputs across time (Altmann, 2002; Christiansen & Curtin, 1999; Dominey & Ramus, 2000; Marcus, 1999; Negishi, 1999; Shastri, 1999; Shultz, 1999, but c.f. Kuehne, Gentner, & Forbus, 2000).

We believe the simple "rules vs. statistics" debate in language acquisition needs to be expanded, or perhaps exploded. On empirical grounds, there is support for both the availability of rule-like representations and the ability of learners to perform statistical inferences over these representations. Abstract, rule-like representations are implied by findings that infants are able to recognize identity relationships (Tyrell, Stauffer, & Snowman, 1991; Tyrell, Zingaro, & Minard, 1993) and even newborns have differential brain responses to exact repetitions (Gervain, Macagno, Cogoi, Peña, & Mehler, 2008). Monkeys (Wallis, Anderson, & Miller, 2001), rats (Murphy, Mondragón, & Murphy, 2008), and honeybees (Giurfa, Zhang, Jenett, Menzel, & Srinivasan, 2001) can recognize and generalize the same sorts of relations that infants can, though the tasks that have been used to test this kind of relational learning vary widely across populations. Learners are also able to make statistical inferences about which rule to learn. For example, infants may have a preference towards parsimony or specificity in deciding between competing generalizations: when presented with stimuli that were consistent with both an *AAB* rule and also a more specific rule, *AA di* (where the last syllable was constrained to be the syllable *di*), infants preferred the narrower generalization (Gerken, 2006, 2010). Following the Bayesian framework for generalization proposed by Tenenbaum and Griffiths (2001), Gerken suggests that these preferences can be characterized as the products of rational statistical inference.

On theoretical grounds, we see neither a pure "rules" position nor a pure "statistics" position as sustainable or satisfying. Without principled statistical inference mechanisms, the pure "rules" camp has difficulty explaining which rules are learned or why the *right* rules are learned from the observed data. Without explicit rule-based representations, the pure "statistics" camp has difficulty accounting for what is actually learned; the best neural network models of language have so far not come close to capturing the expressive compositional structure of language, which is why symbolic representations continue to be the basis for almost all state-of-the-art work in natural language processing (Chater & Manning, 2006; Manning & Schütze, 2000).

Driven by these empirical and theoretical considerations, our work here explores a proposal for how concepts of "rules" and "statistics" can interact more deeply in understanding the phenomena of "rule learning" in human language acquisition. Our approach is to create computational models that perform statistical inference over rule-based representations and test these models on their fit to the broadest possible set of empirical results. The success of these models in capturing human performance across a wide range of experiments lends support to the idea that statistical inferences over rule-based representations may capture something important about what human learners are doing in these tasks.

Our models are *ideal observer* models: they provide a description of the learning problem and show what the correct inference would be, under a given set of assumptions. The ideal observer approach has a long history in the study of perception and is typically used for understanding the ways in which performance conforms to or deviates from the ideal (Geisler, 2003).[1] On this approach, the ideal observer becomes a baseline from which predictions about human performance can be made. When performance deviates from this baseline, researchers can make inferences about how the assumptions of the model differ from those made by human learners (for example, by assuming perfect memory for input data or perfect decision-making among competing alternatives).

Our models are not models of development. While it is possible to use ideal observer models to describe developmental changes (e.g., Kiorpes, Tang, Hawken, & Movshon, 2003), the existing data on rule learning do not provide a rich enough picture to motivate developmental modeling. With few exceptions (Dawson & Gerken, 2009; Johnson et al., 2009), empirical work on rule learning has been geared towards showing what infants can do, rather than providing a detailed pattern of successes and failures across ages. Thus, rather than focusing on the capabilities of learners at a particular age, we have attempted to capture results across the developmental spectrum. It is likely that as more developmental patterns are described empirically, the models we present will need to be modified to take into account developmental changes in cognitive abilities.

In the first section of the paper, we describe the hypothesis space for rules that we use and propose three different ideal observer models for inferring which rule or rules generated a set of training data. These models build on, rather

[1] This approach to modeling learning is also sometimes referred to as a "computational level" analysis, after Marr (1982), because it describes the computational structure of the task rather than the algorithms or mechanisms necessary to perform it. Models at the computational level (including ideal observer models) typically make use of Bayesian methods to compute normative statistical inferences.

than competing with, one another so as to identify which assumptions in each model are crucial for fitting particular phenomena. In the second section, we apply these models to a range of experiments from the literature on infant rule learning.

2. Models

We first create a hypothesis space which defines the set of possible rules that our models could learn and then use Bayesian inference to decide which of these rules best fits the available training strings. The hypothesis space is constant across all three models, but the inference procedure varies depending on the assumptions of each model. This section describes the hypothesis space first, then the inference procedures for each model, and then our methods for linking model simulations to the results of experiments with human infants.[2]

Our approach is to make the simplest possible assumptions about representational components, including the structure of the hypothesis space and the prior on hypotheses. As a consequence, the hypothesis space of our models is too simple to describe the structure of interesting phenomena in natural language, and our priors do not capture any of the representational biases that human learners may brings to language acquisition.

Nevertheless, our hope is that this approach will help in articulating the principles of generalization underlying experimental results on rule learning. While a visit to the lab is surely too short to acquire representations with the semantic or syntactic complexity of natural language, artificial language learning tasks are nevertheless useful tools for investigating the principles by which both simple and complex structures can be learned (Gomez & Gerken, 2000). Our current models are designed around the same logic: they are attempts to characterize the principles that allow learners to succeed in learning, rather than realistic sketches of the representations that are being learned.

2.1. Hypothesis space

Although the hypothesis space for sequential rules could be infinitely large in principle, in practice describing the available empirical data requires only a relatively small set of hypotheses, due to the simplicity of the languages used in infant experiments. This hypothesis space is based on the idea of a rule as a restriction on strings. We define the set of strings S as the set of ordered triples of elements s_1, s_2, s_3 where all s are members of vocabulary of elements, V. There are thus $|V|^3$ possible elements in S. (All of the experiments we address here make use of three-element sequences, but this convention is easily extended to longer strings).

For each set of simulations, we define S as the total set of string elements used in a particular experiment. For example, in the training portion of the first experiment conducted by Marcus et al. (1999), they made use of the set of elements $S = \{ga, gi, ta, ti, na, ni, la, li\}$. These elements are treated by our models as unique identifiers that do not encode any information about phonetic relationships between syllables.

A rule defines a subset of S. Rules are written as ordered triples of primitive functions (f_1, f_2, f_3). Each function operates over an element in the corresponding position in a string and returns a truth value. For example, f_1 defines a restriction on the first string element, x_1. The set F of functions is a set which for our simulations includes $\cdot$ (a function which is always true of any element) and a set of functions is $_y(x)$ which are only true if $x = y$ where y is a particular element. The majority of the experiments addressed here make use of only one other function: the identity function $=_a$ which is true if $x = x_a$. For example, in Marcus et al. (1999), learners heard strings like *ga ti ti* and *ni la la*, which are consistent with $(\cdot, \cdot, =_2)$ (*ABB*, or "second and third elements equal"). The stimuli in that experiment were also consistent with another regularity, however: $(\cdot,\cdot,\cdot)$, which is true of any string in S. One additional set of experiments makes use of musical stimuli for which the functions $>_a$ and $<_a$ (higher than and lower than) are defined. They are true when $x > x_a$ and $x < x_a$ respectively.[3]

Our definition of the hypothesis space restricts the set of possible subsets of S that can be written. Since there are $8^3 = 512$ strings in S for the Marcus et al. (1999) vocabulary, the number of possible arbitrary subsets is very large. However, our notation allows us to write only $|F|^3 = 1331$ possible distinct rules in the Marcus case, of which only 758 pick out distinct subsets of S. An unconstrained version of our notation allows logically equivalent rules (e.g. $(=_2, =_3, =_1)$ and $(\cdot, =_3, =_1)$, both of which pick out strings where all three elements are equal). To avoid ambiguities of this sort, we eliminate redundant rules and assume that rules are uniquely defined by their extension.

2.2. Model 1: single rule

Model 1 begins with the framework for generalization introduced by Tenenbaum and Griffiths (2001). It uses exact Bayesian inference to calculate the posterior probability of a particular rule r given the observed set of training sentences $T = t_1 \ldots t_m$. This probability can be factored via Bayes' rule into the product of the likelihood of the training data being generated by a particular rule $p(T|r)$, and a prior probability of that rule $p(r)$, normalized by the sum of these over all rules:

$$p(r|T) = \frac{p(T|r)p(r)}{\sum_{r' \in R} p(T|r')p(r')}. \tag{1}$$

We assume a uniform prior $p(r) = 1/|R|$, meaning that no rule is *a priori* more probable than any other. For human learners the prior over rules is almost certainly not uni-

[2] Matlab code for models and simulations is available at http://langcog.stanford.edu/materials/rules.html.

[3] Due to algorithmic considerations, we assume a hypothesis space which includes only the appropriate primitives. Including a larger set of possible primitives results will have little effect except under conditions where they capture a superset of the relations captured by a current primitive (e.g. "$=_2 \vee is_{di}$"). Such cases of compositional rules could be treated appropriately using a complexity prior. We return to this issue in Section 4.

form and could contain important biases about the kinds of structures that are used preferentially in human language (whether these biases are learned or innate, domain-general or domain-specific). However, understanding the structure of this prior even for a simple hypothesis space like the one used here will take a large amount of empirical data. Since no data of this form exist, we have chosen an uninformative uniform prior.[4]

We assume that training examples are generated by sampling uniformly from the set of sentences that are congruent with one rule. This assumption is referred to as *strong sampling*, and leads to the *size principle*: the probability of a particular string being generated by a particular rule is inversely proportional to the total number of strings that are congruent with that rule (which we notate $|r|$). Under the size principle, the probability of a set of strings given a rule is

$$p(T|r) = \prod\nolimits_{t_i \in T} p(t_i|r), \tag{2}$$

where

$$p(t_i|r) = \frac{1}{|r|}. \tag{3}$$

One benefit of the simplicity of Model 1 is that we can use exact enumeration to compute the posterior probability of any particular rule given a set of training data.

2.3. Model 2: single rule under noise

Model 1 assumed that every data point must be accounted for by the learner's hypothesis. However, there are many reasons this might not hold for human learners: the learner's rules could permit exceptions, the data could be perceived noisily such that a training example might be lost or mis-heard, or data could be perceived correctly but not remembered at test. Model 2 attempts to account for these sources of uncertainty by consolidating them all within a single parameter. While future research will almost certainly differentiate these factors (for an example of this kind of work, see Frank, Goldwater, Griffiths, & Tenenbaum, 2010), here we consolidate them for simplicity.

To add noise to the input data, we add an additional step to the generative process: after strings are sampled from the set consistent with a particular rule, we flip a biased coin with weight α. With probability α, the string remains the same, while with probability $1-\alpha$, the string is replaced with another randomly chosen element.

Under Model 1, a rule had likelihood zero if any string in the set T was inconsistent with it. With any appreciable level of input uncertainty, this likelihood function would result in nearly all rules having probability zero. To deal with this issue, we assume in Model 2 that learners know that their memory is fallible, and that strings may be misremembered with probability $1-\alpha$. Eq. (3) now reflects this assumption:

$$p(t_i|r) = \begin{cases} \frac{\alpha}{|r|} + \frac{(1-\alpha)}{|S|} & \text{if } t_i \text{ consistent with } r, \\ \frac{(1-\alpha)}{|S|-|r|} & \text{otherwise,} \end{cases} \tag{4}$$

where $|S|$ is the total number of strings in the language. The top alternative—if a training string t_i is consistent with the rule r—is the sum of two terms: first, the probability $\alpha \cdot \frac{1}{|r|}$ that t_i was not corrupted and was sampled from r; and second, the probability that t was sampled uniformly from the set of all possible strings (and just happened to be consistent with r by chance). The bottom alternative—if t is not consistent with r—is the probability that t_i was sampled uniformly from the set of all possible strings and did not happen to be consistent with r by chance.

2.4. Model 3: multiple rules under noise

Model 3 loosens an additional assumption: that all the strings in the input data are the product of a single rule. Instead, it considers the possibility that there are multiple rules, each consistent with a subset of the training data. We encode a weak bias to have fewer rules via a prior probability distribution that favors more compact partitions of the input. This prior is known as a Chinese Restaurant Process (CRP) prior (Rasmussen, 2000); it introduces a second free parameter, γ, which controls the bias over clusterings. A low value of γ encodes a bias that there are likely to be many small clusters, while a high value of γ encodes a bias that there are likely to be a small number of large clusters.

The joint probability of the training data T and a partition Z of those strings into rule clusters is given by

$$P(T,Z) = P(T|Z)P(Z), \tag{5}$$

neglecting the parameters α and γ. The probability of a clustering $P(Z)$ is given by $CRP(Z,\gamma)$.[5]

Then the probability of the training data given the cluster assignments is the product of independent terms for each string:

$$P(T|Z) = \prod_{t_i \in T} P(t_i|z_i), \tag{6}$$

where z_i is the cluster assignment for each individual string. Because strings in each cluster c are generated by a rule r_c for that cluster, we group the terms in Eq. (6) into a product over clusters and then a separate product over strings in that cluster:

$$P(T|Z) = \prod_{c} \prod_{t_i \in c} \sum_{r_c} P(t_i|r_c)P(r_c). \tag{7}$$

[4] Note that we distinguish between two senses of the term *prior*. In informal use, the term often refers to modeling assumptions such as the assumption that identity is a primitive operation. Here we use it in the technical sense of a probability distribution over rules that are possible in the representation language we have chosen.

[5] The form of the Chinese Restaurant Process (CRP) is

$$CRP(Z,\gamma) = \frac{\Gamma(\gamma) \cdot \gamma^{|Z|}}{\Gamma(\gamma+n)} \prod_{z \in Z} \Gamma(|z|),$$

where Γ is the gamma (generalized factorial) function, $|Z|$ is the number of clusters, $|z|$ is the size of each cluster, and n is the number of total training examples.

Because r_c is not known, the inner sum integrates the predictions of all rules congruent with the strings in the cluster, weighted by their prior $P(r_c)$. As in Models 1 and 2 we assume a uniform prior $P(r_c)$. We use Eq. (4) (the noise likelihood function) from Model 2 to give us the probability of a particular test string given a rule.

Unlike in Models 1 and 2, inference by exact enumeration is not possible and so we are not able to compute the normalizing constant. But we are still able to compute the relative posterior probability of a partition of strings into clusters (and hence the posterior probability distribution over rules for that cluster). Thus, we can use a Markov-chain Monte Carlo (MCMC) scheme to find the posterior distribution over partitions. In practice we use a Gibbs sampler, an MCMC method for drawing repeated samples from the posterior probability distribution via iteratively testing all possible cluster assignments for each string (MacKay, 2003).

2.5. Input data and linking hypotheses

In all simulations we calculate the posterior probability distribution over rules given the set of unique string *types* used in the experimental stimuli. We use types rather than rather than individual string *tokens* because a number of computational and experimental investigations have suggested that types rather than tokens may be a psychologically natural unit for generalization (Gerken & Bollt, 2008; Goldwater, Griffiths, & Johnson, 2006; Richtsmeier, Gerken, & Ohala, in press).[6]

To assess the probability of a set of test items $E = e_1 \ldots e_n$ (again computed over types rather than tokens) after a particular training sequence, we calculate the total probability that those items would be generated under a particular posterior distribution over hypotheses. This probability is

$$p(E|T) = \sum_{r_j \in R} \prod_{e_k \in E} p(e_k|r_j)p(r_j|T), \tag{8}$$

which is the product over examples of the probability of a particular example, summed across the posterior distribution over rules $p(R|T)$. For Model 1 we compute $p(e_k|r_j)$ using Eq. (2); for Models 2 and 3 we use Eq. (4).

We use *surprisal* as our main measure linking posterior probabilities to the results of looking time studies. Surprisal (negative log probability) is an information-theoretic measure of how unlikely a particular outcome is. It has been used previously to model adult reaction time data in sentence processing tasks (Hale, 2001; Levy, 2008) as well as infant looking times (Frank, Goodman, & Tenenbaum, 2009). For the studies that used two-alternative forced-choice measures, we used a Luce choice rule (Luce, 1963) to compute the probability of choosing one alternative over the other. Since both of these studies were fit using Model 3, the score for each alternative was the non-normalized posterior probability of the appropriate clustering.

3. Results

We report simulation results for each of the three models across a variety of experiments in the literature on rule learning. Results are ordered in terms of which models adequately capture the pattern of results. Table 1 gives a summary of model coverage.

3.1. Marcus et al. (1999)

Marcus et al. (1999) exposed infants to strings of syllables of the form *ABA* or *ABB* and then evaluated whether infants had learned the appropriate rule by exposing them to alternating sets of strings made up of novel syllables but conforming to either the same regularity they had heard during training or another (e.g. for *ABB* training, test was novel strings of forms *ABA* and *ABB*). As a group, infants listened longer to the strings instantiating the rule they had not learned, despite the matched novelty of the individual syllables.

All three models were able to learn the correct rules in these experiments. When trained on 16 *ABA* training strings, Model 1 identified two hypotheses with non-zero posterior probability: $(\cdot, \cdot, \cdot)$ and $(\cdot, \cdot, =_1)$, but the more specific identity rule received far higher posterior probability; the same was true for *ABB* (Table 2). Model 1 also showed far higher surprisal to rule-incongruent strings (Table 3). Results of simulations with Models 2 and 3 confirmed that the posterior distribution over rules for both models very strongly supported the correct generalizations.

Table 1
Summary of results fit by Models 1–3.

Paper	Result modeled	Model 1	Model 2	Model 3
Marcus et al. (1999)	Rule learning	x	x	x
Endress et al. (2007)	Ordinal vs. identity rules	x	x	x
Frank et al. (2009)	Uni- vs. multi-modal learning	x	x	x
Gerken (2006)	Breadth of generalization	x	x	x
Gerken (2010)	Breadth of generalization		x	x
Marcus et al. (2007)	Asymmetric cross-modal transfer		x	x
Saffran et al. (2007)	Individual differences in learning		x	x
Gómez (2002)	Multiple non-adjacent dependencies			x
Kovács and Mehler (2009)	Bilingual rule learning			x

[6] In the models as formulated we have assumed that strings in the language are sampled in sequence with replacement from all grammatical strings. When using type-level data, sampling without replacement and without regard for the sequence of observations may be more natural. This produces similar model predictions but makes the mathematics more complex. For instance, the likelihood for Model 1 (Eqs. (2) and (3)) would become $\frac{m!(|r|-m)!}{|r|!}$, representing the probability of drawing a subset of m types from a language with $|r|$ types in total. When r is large relative to m (as it is in nearly all of our simulations), this formulation is closely approximated by the likelihood we use here.

Table 2
Log posterior probabilities over rules for simulations using Model 1. All rules with non-zero probability are listed. The highest probability rule for each experiment is shown in bold. Note: −0.00 represents very small numbers, corresponding to probabilities very slightly less than 1.

Paper	Condition	Rule	Log P
Marcus (1999)	*ABA*	$(\cdot,\cdot,\cdot)$	−33.27
		$(\cdot,\cdot,=_1)$	**−0.00**
	ABB	$(\cdot,\cdot,\cdot)$	−33.27
		$(\cdot,\cdot,=_2)$	**−0.00**
Endress et al. (2007)	*ABB*	$(\cdot,\cdot,\cdot)$	−8.32
		$(\cdot,\cdot,=_2)$	**−0.00**
	LHM	$(\cdot,\cdot,\cdot)$	−8.33
		$(\cdot,\cdot,>_1)$	−5.21
		$(\cdot,\cdot,<_2)$	−5.21
		$(\cdot,>_1,\cdot)$	−5.21
		$(\cdot,>_1,>_1)$	−3.42
		$(\cdot,>_1,<_2)$	−3.42
		$(\cdot,>_3,>_1)$	**−0.09**
Frank et al. (2009)	*ABB*-uni	$(\cdot,\cdot,\cdot)$	−5.38
		$(\cdot,\cdot,=_2)$	**−0.00**
	ABB-multi	$(\cdot,\cdot,\cdot)$	−10.75
		$(\cdot,\cdot,=_2)$	**−0.00**
Gerken (2006)	*AAB*	$(\cdot,\cdot,\cdot)$	−8.32
		$(\cdot,=_1,\cdot)$	**−0.00**
	AAx	$(\cdot,\cdot,\cdot)$	−16.64
		$(\cdot,\cdot,is_x)$	−8.32
		$(\cdot,=_1,\cdot)$	−8.32
		$(\cdot,=_1,is_x)$	**−0.00**
Gerken (2010)	column + 5	$(\cdot,\cdot,\cdot)$	−18.71
		$(\cdot,=_1,\cdot)$	**−0.00**
	music + 5	$(\cdot,\cdot,\cdot)$	−10.40
		$(\cdot,=_1,\cdot)$	**−0.00**

Table 3
Surprisal (negative log probability) for a single test item for simulations with Model 1. The rule stimulus inferred by the experimenters to have been learned (if there was evidence of learning) is shown in bold. Due to differences in the age of participants, levels of surprisal necessary for success are not comparable across experiments.

Paper	Condition	Test	Surprisal
Marcus (1999)	*ABA*	*ABB*	39.51
		ABA	4.16
	ABB	**ABB**	4.16
		ABA	39.51
Endress et al. (2007)	*ABB*	**ABB**	4.97
		ABA	17.39
	LHM	*LHM*	5.44
		MHL	9.45
Frank et al. (2009)	*ABB*-uni	*ABB*	3.59
		ABA	10.76
	ABB-multi	**ABB**	7.17
		ABA	21.50
Gerken (2006)	*AAB*	**AAB**	4.16
		ABA	14.56
	AAx	*AAB*	12.48
		ABA	22.87
	AAx E2	**AAx**	2.08
		AxA	22.87
Gerken (2010)	column + 5	**AAB**	4.16
		ABA	24.95
	music + 5	*AAB*	4.16
		ABA	16.64

3.2. Endress, Dehaene-Lambertz, and Mehler (2007)

Endress et al. (2007) investigated differences between identity functions and what they called "ordinal" functions: higher or lower in musical pitch. They presented adult participants with sets of four three-tone strings instantiating either identity rules like *ABB* or *ABA* or else ordinal rules like *low–high–middle* (*LHM*) or *middle–high–low* (*MHL*). In their experiments they found that while participants were able to distinguish test sequences in the identity condition (*ABB* vs. *ABA*), they never distinguished the ordinal rules (*LHM* vs. *MHL*) at greater than chance levels. This result was interpreted as evidence for the view that identity functions but not ordinal functions are "perceptual primitives" which are accessible for the construction of symbolic abstractions.

Our results suggest that their experimental stimuli show structural differences which confound the issue of the functions they used with the complexity of the rules constructed from those functions. While the correct hypothesis in the identity case had much higher posterior probability under Model 1 than did the only competitor (the null rule), the same was not true in the case of the ordinal rules, where a number of possible rules were consistent with the training stimuli (Table 2). Surprisal values showed that the test stimuli were more different from one another in the identity condition than the ordinal condition (Table 3). Thus, the striking difference in the performance of adult participants observed by Endress et al. could have been due to the complexity of the rules being learned, as well as to the kinds of functions in those rules.[7]

3.3. Frank et al. (2009)

Experiments by Frank, Slemmer, Marcus, and Johnson (2009) suggested that five-month-old infants did not show evidence of learning *ABA* or *ABB* rules when they were presented unimodally using either auditory syllables or sequentially looming visual stimuli. When training examples were presented using coordinated multi-modal stimuli (a looming shape accompanied by a syllable), however, five-month-olds showed evidence of discrimination at test.

This effect may be captured in a number of ways by Models 1 and 2. Under Model 1, these results can be described via differences in the dimensionality of S (the set of possible training sequences) for uni- and multi-modal stimuli: with eight elements arranged into three-item strings, there are $8^3 = 512$ unimodal strings possible; with eight syllables which each can be accompanied by one of eight shapes, there are $8^2 = 64$ primitive units and $64^3 = 262{,}144$ possible strings. Tables 2 and 3 show results using these different dimensionalities.

Alternatively, this result can be fit in Model 2 simply by assuming different values of α for uni- and multi-modal stimuli. This account is in alignment with the predictions the Intersensory Redundancy Hypothesis (Bahrick &

[7] A simpler comparison of ordinal and identity relations would be a comparison of $(\cdot,\cdot,=_2)$ and $(\cdot,\cdot,>_2)$.

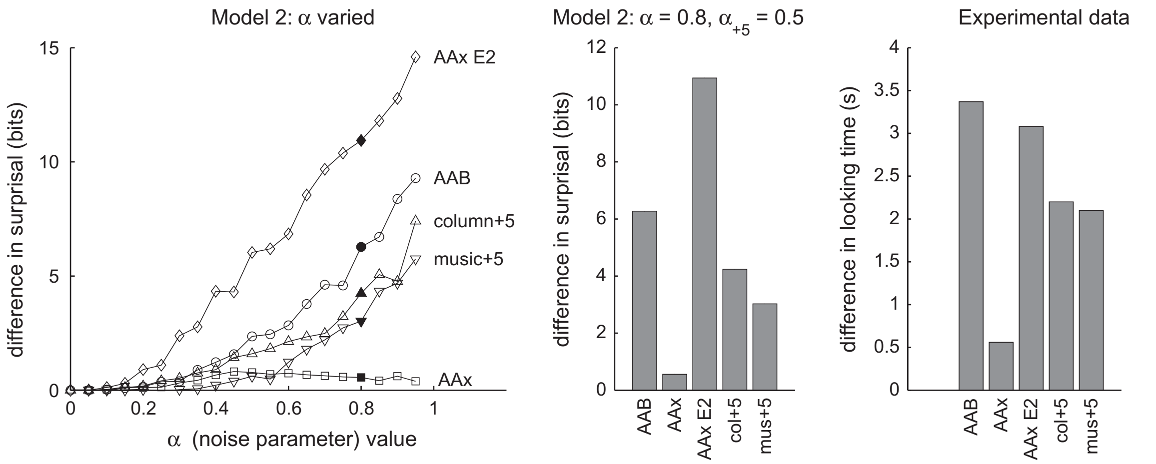

Fig. 1. Simulation results for Model 2 on the results of Gerken (2006, 2010). Left side: horizontal axis shows values of α (memory noise parameter). Vertical axis shows the difference in surprisal values (in bits) for rule-incongruent stimuli relative to rule-congruent stimuli, an index of learning. Middle: difference in surprisal is plotted across the five conditions in the two experiments for the parameter shown by filled markers on the left axis. Right side: differences in looking times from Gerken (2006, 2010). Note that the music + 5 condition produced a familiarity, rather than novelty, preference.

Lickliter, 2000; Bahrick, Flom, & Lickliter, 2002), which assumes greater salience—rather than greater informational content—for intermodal stimuli (for more discussion of the differences between these accounts, see Frank et al., 2009).

3.4. Gerken (2006)

Gerken (2006) investigated the breadth of the generalizations drawn by learners when exposed to different input corpora, testing learners on strings either of the form *AAB* or of the form *AAx* (where *x* represented a single syllable). Model 1 correctly identified the rule $(\cdot, =_1, \cdot)$ for *AAB* and $(\cdot, =_1, is_x)$ for *AAx* (Table 2). Unlike human infants, who did not discriminate examples from the more general rules (*AAB* vs. *ABA*) when trained on specific stimuli (*AAx*), Model 1 showed differences in surprisal between all three conditions (Table 3). The absolute magnitude of the probability of the congruent (*AAB*) test items in the **AAx**-training condition was extremely low, however.

Model 2 produced a similar pattern of results to those observed by Gerken (Fig. 1), with the majority of α values producing a qualitatively similar picture. With *AAx* training and testing on *AAx* and *AxA* strings (notated in the figure as *AAx* E2), there was a large difference in surprisal; because of the specificity of the *AAx* rule, $(\cdot, =_1, is_{di})$ was highly favored in the posterior and the incongruent strings were highly surprising relative to the congruent strings. *AAB* training with *AAB* vs. *ABA* test also produced differences. Model 2 showed no difference between congruent and incongruent test items for the condition in which infants failed, however, suggesting that the probability of memory noise in Model 2 swamped the low absolute probabilities of the test items.

3.5. Gerken (2010)

Gerken (2010) investigated the flexibility of infants' generalizations by testing whether they were able to switch between a narrow and a broad generalization with only a small amount of evidence. Seven and a half month-olds were either trained on *AAx* stimuli for the majority of the familiarization with three of the last five strings consistent with *AAB* ("column + 5" condition), or played music for the majority of the familiarization and then the same five strings ("music + 5" condition). At test, infants familiarized to the column + 5 condition discriminated *AAB* from *ABA* stimuli, while those in the music + 5 condition showed a comparably large but non-significant familiarity preference. Under Model 1, differences between the column + 5 and music + 5 condition were apparent but relatively slight, since all of the three *AAB*-consistent strings supported the broader generalization in both conditions.

These results do not take into account the much greater exposure of the infants to the narrow-generalization strings (those consistent with *AAx*). To capture this differential we conducted simulations with Model 2 where we assumed a lower level of α (which we denote by α_{+5}) for the three new string types that were introduced at the end of exposure (Fig. 1).[8] Across a range of α and α_{+5} values, although the model did not reproduce the familiarity/novelty reversal seen in the experimental data, there was a significant difference between the column + 5 condition and the music + 5 condition. This difference was due to the extra support for the broad generalization given by the well-remembered familiarization strings in the column + 5 condition.

3.6. Interim discussion

In the preceding results, differences across conditions and experiments produced a range of differences in surprisal based in Model 1. Correct and incorrect test items varied in both their relative and absolute probabilities,

[8] Although these strings could also be argued to be more salient because of their recency, this effect is likely to be small relative to the dozens of repetitions of the *AAx* strings heard during familiarization.

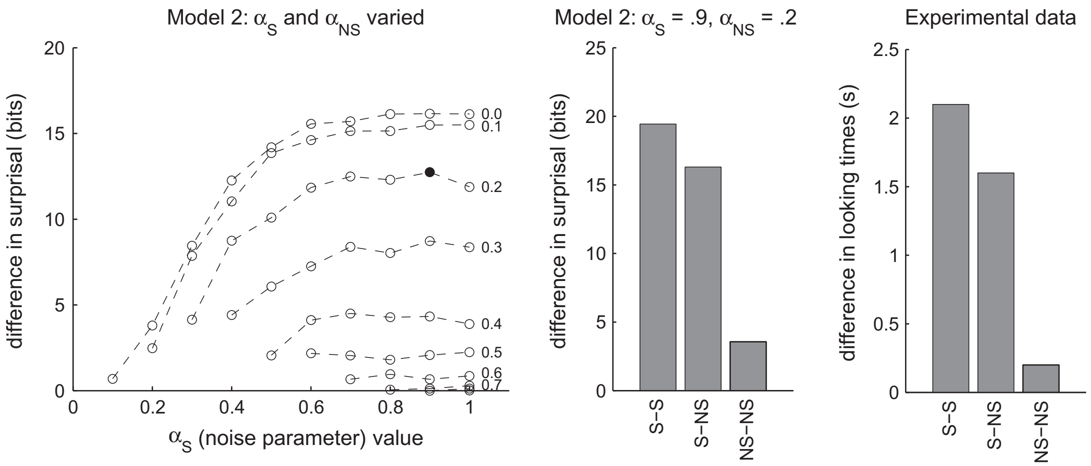

Fig. 2. Simulation results for Model 2 on the stimuli used by Marcus et al. (2007). Left side: difference in surprisal between S–NS and NS–NS conditions across a range of values of α_S and α_{NS} (note that all differences are positive, indicating that the S–NS condition always had higher surprisal). Middle: difference in surprisal on all conditions for a single parameter set, marked with a filled circle on the left axis. Right side: experimental data from Marcus et al. (2007), replotted by difference in looking time.

and understanding when these differences were likely to predict a difference in looking time or responding was not always straightforward. We introduced Model 2 in part to quantify the intuition that, in cases where absolute probabilities were very small (as in the *AAx* condition of Gerken, 2006), the differences between conditions would be overwhelmed by even a low level of memory noise.

The α parameter in Model 2 is useful for more than just explaining these situations, however. In the following sections we turn to two results where there may be intrinsic differences in the representation of stimuli between modalities or between individuals. We then end by considering two results that can only be fit by Model 3, a model that can learn different rules to explain different subsets of the input data.

3.7. Marcus, Fernandes, and Johnson (2007)

Marcus et al. (2007) reported that while 7.5 month-olds showed evidence of learning rules in sung speech stimuli (with different syllables corresponding to each tone), they did not appear to have learned the same rules when the training stimuli were presented in pure tones instead of sung speech. In addition, children trained with speech stimuli seemed to be able to discriminate rule-congruent and rule-incongruent stimuli in other modalities—tones, musical instrument timbres, and animal sounds—at test. Marcus and colleagues interpreted this evidence for cross-modal transfer as suggesting that infants may analyze speech more deeply than stimuli from other modalities.

Model 2 allows a test of a possible alternative explanation, inspired by the robust effects of prior knowledge on the recognition of stimuli in noise found in object perception (e.g. Biederman, 1972; Eger, Henson, Driver, & Dolan, 2007; Gregory, 1970; Sadr & Sinha, 2004): knowing what object you are looking for allows recognition under a higher level of noise than when the object is unknown. If non-speech domains are "noisier" (encoded or remembered with lower fidelity) than speech stimuli, rules may be easier to recognize in noisy, non-speech stimuli than they are to extract from those same stimuli.

Model 2 reproduces the cross-modal transfer asymmetry reported by Marcus et al. (2007) although it assumes only differences in memory—rather than structural differences in the kinds of patterns that are easy to learn—across domains. To capture the hypothesis of differential familiarity with speech, we assumed that whatever the value of α_S for speech, the value of α_{NS} for non-speech stimuli would be lower. Fig. 2, left, plots the difference in surprisal between the speech/non-speech (S–NS) and non-speech/non-speech (NS–NS) conditions while varying α_S and α_{NS}. Surprisal was higher in the S–NS condition than in the NS–NS condition, suggesting that a basic difference in memory noise could have led to the asymmetry that Marcus et al. reported.[9]

3.8. Saffran, Pollak, Seibel, and Shkolnik (2007)

Saffran et al. (2007) showed that infants succeeded in learning rules from simultaneously-presented pictures of dogs of different breeds (e.g. *malamute-cattle dog-cattle dog*). The authors reported a correlation between parents' ratings of how interested the infant was in dogs and the size of the rule-congruent/rule-incongruent looking time difference at test, suggesting that individual differences

[9] Although test stimuli were scored using Eq. (4), in these simulations (as in others) they were not themselves corrupted. This manipulation reflects the differences between training (in which it is necessary to remember multiple strings to make a generalization) and test (in which posterior surprisal can be calculated for individual strings). We conducted an identical set of simulations for the Marcus et al. (2007) data where test sequences *were* corrupted and again found a large range of α values under which the cross-modal transfer condition produced significantly higher surprisal values than the non-speech condition (although there was now an appreciable gap in performance between speech and cross-modal transfer conditions as well).

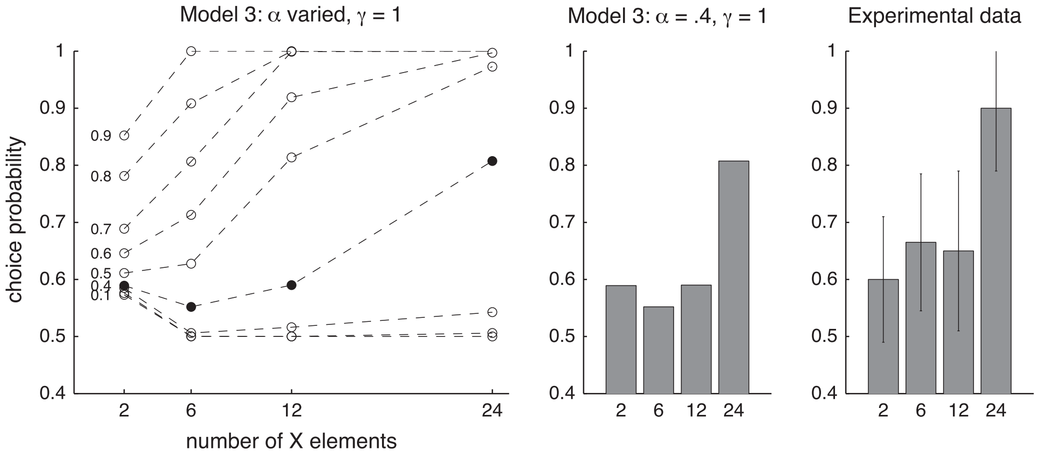

Fig. 3. Model 3 simulation results on the experimental stimuli of Gómez (2002). Left side: choice probability for two rules (correct) vs. one rule (incorrect) for a range of α values at $\gamma = 1$, plotted by the number of X elements (see text). Middle: results from a single parameter set, marked with filled circles in the left axis. Right side: adult experimental data from Gómez (2002), replotted as proportion correct. Error bars show standard error of the mean.

in experience with a particular stimulus item might aid in rule learning. Using Model 2, we simulated rule learning across a range of α values and tested whether there was a correlation between the strength of encoding and the resulting surprisal values. Higher values of α led to higher surprisal at test ($r = .88$), indicating a relationship between familiarity/encoding fidelity and learning similar to that observed by Saffran et al.

3.9. Gómez (2002)

Gómez (2002) investigated the learning of non-adjacent dependencies by adults and 18-month-olds. Eighteen-month-olds were trained on a language that contained sentences of the form *aXb* and *cXd* where *a*, *b*, *c*, and *d* represented specific words while *X* represented a class of words whose membership was manipulated across participants. When participants were trained on sentences generated with 2, 6, or 12 *X* elements, they were not able to distinguish *aXb* elements from *aXc* elements at test, suggesting that they had not learned the non-adjacent dependency between *a* and *b*; when they were trained on sentences with 24 *X* elements, they learned the non-adjacent dependency.[10]

Models 1 and 2 both fail in this task because both only have the capacity to encode a single rule. Under Model 1, all training stimuli are only consistent with the null rule $(\cdot, \cdot, \cdot)$; under Model 2, at some levels of α, a single rule like $(is_a, \cdot, is_b)$ is learned while strings from the other rule are attributed to noise.

In contrast, Model 3 with $\alpha = 1$ successfully learns both rules—$(is_a, \cdot, is_b)$ and $(is_b, \cdot, is_c)$—in all conditions, including when *X* has only two elements. To test whether memory noise might impede performance, we calculated the choice probability for the correct, two-rule hypothesis compared with the incorrect, one-rule hypothesis across a range of values of α and γ. Results for $\gamma = 1$ are shown in Fig. 3; in general, γ values did not strongly affect performance. For high levels of noise, one rule was equally likely as the variability of *X* increased; in contrast, for low levels of noise, two rules were more likely as the variability of *X* increased. At a moderate-to-high level of noise ($\alpha = .4$), Model 3 matched human performance, switching from preferring one rule at lower levels of variability to two rules at the highest level of variability. These results suggest that, although variability increases generalization (as claimed by Gomez), memory constraints may work against variability by increasing the number of examples necessary for successful generalization (for a more detailed discussion and some experimental results in support of this hypothesis, see Frank & Gibson, in press).

3.10. Kovács and Mehler (2009)

Work by Kovács and Mehler (2009b) investigated the joint learning of two rules at once. They found that while bilinguals were able to learn that two different rules (*ABA* and *AAB*) cued events at two different screen locations, monolinguals only learned to anticipate events cued by one of the two rules. They interpreted these results in terms of early gains in executive function by the bilingual infants (Kovács & Mehler, 2009a).

Under Model 3, Kovács & Mehler's hypothesis can be encoded via differences in the γ parameter across "bilingual" and "monolingual" simulations. The γ parameter controls the concentration of the prior distribution on rules: when γ is high, Model 3 is more likely to posit many rules to account for different regularities; when γ is low, a single, broader rule is more likely. We conducted a series of simulations, in which we assumed that bilingual learners had a higher value of γ than did monolingual learners.

[10] Gómez's experiment differs from the other experiments described here in that a learner could succeed simply by memorizing the training stimuli, since the test was a familiarity judgment rather than a generalization test. Nevertheless, the results from Gómez's experiments—and from a replication in which novel and familiar test items were rated similarly (Frank & Gibson, in press)—both suggest that memorization is not the strategy taken by learners in languages of this type.

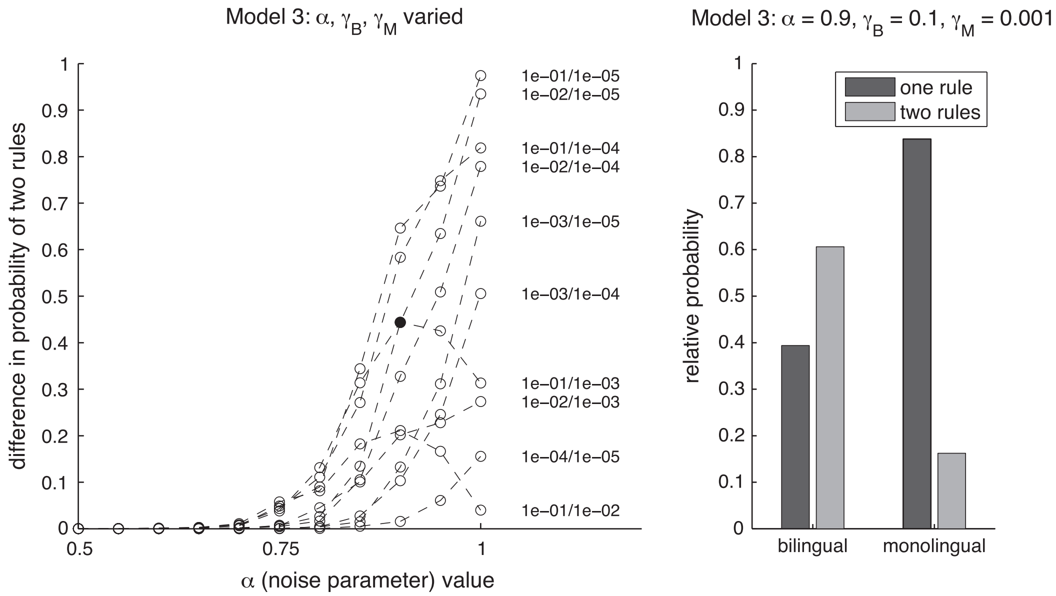

Fig. 4. Model 3 simulation results on the experimental stimuli of Kovács and Mehler (2009). Left axis plots difference in probability between two rules and one rule by α (noise parameter) across a range of parameter values. Each line shows this probability difference for a pair of γ_B (bilingual) and γ_M (monolingual) values, shown on the right side. Right axis shows relative rule probability for bilingual and monolingual simulations for one set of parameters, marked with a filled circle in the left axis.

Results are shown in Fig. 4. Across a range of parameter values, simulations with the "bilingual" parameter set assigned a higher choice probability to learning two rules, while simulations with the "monolingual" parameter set assigned more probability to learning a single rule.

Model 3's success suggests that the empirical results can be encoded as a more permissive prior on the number of regularities infants assume to be present in a particular stimulus. In practice this may be manifest via better executive control, as hypothesized by Kovács & Mehler. Although in our current simulations we varied γ by hand, under a hierarchical Bayesian framework it should be possible to learn appropriate values for parameters on the basis of their fit to the data, corresponding to the kind of inference that infant learners might make in deciding that the language they are hearing comes from two languages rather than one.

4. General discussion

The infant language learning literature has often been framed around the question "rules or statistics?" We suggest that this is the wrong question. Even if infants represent symbolic rules with relations like identity—and there is every reason to believe they do—there is still the question of how they learn these rules, and how they converge on the correct rule so quickly in a large hypothesis space. This challenge requires statistics for guiding generalization from sparse data.

In our work here we have shown how domain-general statistical inference principles operating over minimal rule-like representations can explain a broad set of results in the rule learning literature. We created ideal observer models of rule learning that incorporated a simple view of human performance limitations, seeking to capture the largest possible set of results while making the fewest possible assumptions. Rather than providing models of a specific developmental stage or cognitive process, our goal was instead to create a baseline for future work that can be modified and enriched as that work describes the effects of processing limitations and developmental change on rule learning. Our work contrasts with previous modeling work on rule learning that has been primarily concerned with representational issues, rather than broad coverage.

The inferential principles encoded in our models—the size principle (or in its more general form, Bayesian Occam's razor) and the non-parametric tradeoff between complexity and fit to data encoded in the Chinese Restaurant Process—are not only useful in modeling rule learning within simple artificial languages. They are also the same principles that are used in computational systems for natural language processing that are engineered to scale to large datasets. These principle have been applied to tasks as varied as unsupervised word segmentation (Brent, 1999; Goldwater, Griffiths, & Johnson, 2009), morphology learning (Albright & Hayes, 2003; Goldwater et al., 2006; Goldsmith, 2001), and grammar induction (Bannard, Lieven, & Tomasello, 2009; Klein & Manning, 2005; Perfors, Tenenbaum, & Regier, 2006). Our work suggests that although the representations used in artificial language learning experiments may be too simple to compare with the structures found in natural languages, the inferential principles that are revealed by these studies may still be applicable to the problems of language acquisition.

Despite the broad coverage of the simple models described here, there are a substantial number of results in the broader literature on rule learning that they cannot capture. Just as the failures of Model 1 pointed towards Models 2 and 3, phenomena that are not fit by our models point the way towards aspects of rule learning that need to be better understood. To conclude, we review three sets of issues—hypothesis space complexity, type/token effects in memory, and domain-specific priors—that suggest other interesting computational and empirical directions for future work.

First, our models assumed the minimal machinery needed to capture a range of findings. Rather than making a realistic guess about the structure of the hypothesis space for rule learning, where evidence was limited we assumed the simplest possible structure. For example, although there is some evidence that infants may not always encode absolute positions (Lewkowicz & Berent, 2009), there have been few rule learning studies that go beyond three-element strings. We therefore defined our rules based on absolute positions in fixed-length strings. For the same reason, although previous work on adult concept learning has used infinitely expressive hypothesis spaces with prior distributions that penalize complexity (e.g. Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Kemp, Goodman, & Tenenbaum, 2008), we chose a simple uniform prior over rules instead. With the collection of more data from infants, however, we expect that both more complex hypothesis spaces and priors that prefer simpler hypotheses will become necessary.

Second, our models operated over unique string types as input rather than individual tokens. This assumption highlights an issue in interpreting the α parameter of Models 2 and 3: there are likely different processes of forgetting that happen over types and tokens. While individual tokens are likely to be forgotten or misperceived with constant probability, the probability of a type being misremembered or corrupted will grow smaller as more tokens of that type are observed (Frank et al., 2010). An interacting issue concerns serial position effects. Depending on the location of identity regularities within sequences, rules vary in the ease with which they can be learned (Endress, Scholl, & Mehler, 2005; Johnson et al., 2009). Both of these sets of effects could likely be captured by a better understanding of how limits on memory interact with the principles underlying rule learning. Although a model that operates only over types may be appropriate for experiments in which each type is nearly always heard the same number of times, models that deal with linguistic data must include processes that operate over both types and tokens (Goldwater et al., 2006; Johnson, Griffiths, & Goldwater, 2007).

Finally, though the domain-general principles we have identified here do capture many results, there is some additional evidence for domain-specific effects. Learners may acquire expectations for the kinds of regularities that appear in domains like music compared with those that appear in speech (Dawson & Gerken, 2009); in addition, a number of papers have described a striking dissociation between the kinds of regularities that can be learned from vowels and those that can be learned from consonants (Bonatti, Peña, Nespor, & Mehler, 2005; Toro, Nespor, Mehler, & Bonatti, 2008). Both sets of results point to a need for a hierarchical approach to rule learning, in which knowledge of what kinds of regularities are possible in a domain can itself be learned from the evidence. Only through further empirical and computational work can we understand which of these effects can be explained through acquired domain expectations and which are best explained as innate domain-specific biases or constraints.

Acknowledgements

Thanks to Denise Ichinco for her work on an earlier version of these models, reported in Frank, Ichinco, and Tenenbaum (2008), and to Richard Aslin, Noah Goodman, Steve Piantadosi, and the members of Tedlab and Cocosci for valuable discussion. Thanks to AFOSR Grant FA9550-07-1-0075 for support. The first author was additionally supported by a Jacob Javits Graduate Fellowship and NSF DDRIG #0746251.

References

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

Altmann, G. (2002). Learning and development in neural networks—The importance of prior experience. *Cognition, 85*(2), 43–50.

Bahrick, L., Flom, R., & Lickliter, R. (2002). Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Developmental Psychobiology, 41*(4), 352–363.

Bahrick, L., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology, 36*(2), 190–201.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences, 106*(41), 17284.

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177*(43), 77–80.

Bonatti, L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations. *Psychological Science, 16*(6), 451–459.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning, 34*(1), 71–105.

Chater, N., & Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences, 10*(7), 335–344.

Chomsky, N. (1981). Principles and parameters in syntactic theory. *Explanation in Linguistics: The Logical Problem of Language Acquisition*, 32–75.

Christiansen, M. H., & Curtin, S. L. (1999). The power of statistical learning: No need for algebraic rules. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the 21st annual conference of cognitive science society*. Mahwah, NJ: Erlbaum.

Dawson, C., & Gerken, L. (2009). From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition, 111*(3), 378–382.

Dominey, P., & Ramus, F. (2000). Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes, 15*(1), 87–127.

Eger, E., Henson, R., Driver, J., & Dolan, R. (2007). Mechanisms of top-down facilitation in perception of visual objects studied by fMRI. *Cerebral Cortex, 17*(9), 2123.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Endress, A., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition, 105*(3), 577–614.

Endress, A., Scholl, B., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General, 134*, 406–419.

Frank, M. C., Gibson, E. (in press). Overcoming memory limitations in rule learning. *Language Learning and Development*.

Frank, M. C., Goldwater, S., Griffiths, T., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition, 117*, 107–125.

Frank, M. C., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20*(5), 578–585.

Frank, M. C., Ichinco, D., & Tenenbaum, J. (2008). Principles of generalization for learning sequential structure in language. In B. C. Love, K. McRae, & V. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.

Frank, M. C., Slemmer, J. A., Marcus, G. F., & Johnson, S. P. (2009). Information from multiple modalities helps five-month-olds learn abstract rules. *Developmental Science, 12*(4), 504.

Geisler, W. (2003). Ideal observer analysis. In *The visual neurosciences* (pp. 825–837). Cambridge, MA: MIT Press.

Gerken, L. A. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition, 98*(3), 67–74.

Gerken, L. A. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition, 115*(2), 362–366.

Gerken, L. A., & Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Language Learning and Development, 3*, 228–248.

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences, 105*(37), 14222.

Giurfa, M., Zhang, S., Jenett, A., Menzel, R., & Srinivasan, M. (2001). The concepts of 'sameness' and 'difference' in an insect. *Nature, 410*(6831), 930–933.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics, 27*(2), 153–198.

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

Goldwater, S., Griffiths, T., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112*, 21–54.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, 431–436.

Gómez, R., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*(2), 109–135.

Gomez, R. L., & Gerken, L. A. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*, 178–186.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32*(1), 108–154.

Gregory, R. L. (1970). *The intelligent eye*. New York, NY: McGraw-Hill.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the North American association for computational linguistics* (Vol. 2, pp. 159–166).

Johnson, M., Griffiths, T., & Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. *Advances in Neural Information Processing Systems, 19*.

Johnson, S. P., Fernandas, K. J., Frank, M. C., Kirkham, N., Marcus, G. F., Rabagliati, H., et al. (2009). Abstract rule learning for visual sequences in 8- and 11-month-olds. *Infancy, 14*(1), 2.

Kemp, C., Goodman, N., & Tenenbaum, J. (2008). Learning and using relational theories. *Advances in Neural Information Processing Systems, 20*.

Kiorpes, L., Tang, C., Hawken, M., & Movshon, J. (2003). Ideal observer analysis of the development of spatial contrast sensitivity in macaque monkeys. *Journal of Vision, 3*(10).

Klein, D., & Manning, C. D. (2005). Natural language grammar induction with a generative constituent-context model. *Pattern Recognition, 38*, 1407–1419.

Kovács, A. M., & Mehler, J. (2009a). Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences, 106*, 6556–6560.

Kovács, A. M., & Mehler, J. (2009b). Flexible learning of multiple speech structures in bilingual infants. *Science, 325*(5940), 611.

Kuehne, S. E., Gentner, D., & Forbus, K. D. (2000). Modeling infant learning via symbolic structural alignment. In *Proceedings of the twenty-second annual conference of the cognitive science society* (pp. 286–291).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177.

Lewkowicz, D., & Berent, I. (2009). Sequence learning in 4-month-old infants: Do infants represent ordinal information? *Child Development, 80*(6), 1811–1823.

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.

Manning, C. D., & Schütze, H. (2000). *Foundations of statistical natural language processing*. MIT Press.

Marcus, G. F. (1999). Response to "Do infants learn grammar with algebra or statistics?" *Science, 284*(5413), 433.

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. *Psychological Science, 18*(5), 387.

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 Company.

Murphy, R. A., Mondragón, E., & Murphy, V. A. (2008). Rule learning by rats. *Science, 319*, 1849–1851.

Negishi, M. (1999). Response to "Rule learning by seven-month-old infants". *Science, 284*, 435.

Perfors, A., Tenenbaum, J., & Regier, T. (2006). Poverty of the stimulus? a rational approach. In *Proceedings of the 28th annual conference of the cognitive science society* (pp. 663–668). Austin, TX: Cognitive Science Society.

Pinker, S. (1991). Rules of language. *Science, 253*(5019), 530.

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

Richtsmeier, P. T., Gerken, L., Ohala, D. K. (in press). Contributions of phonetic token variability and word-type frequency to children's phonological representations. *Journal of Child Language*.

Sadr, J., & Sinha, P. (2004). Object recognition and random image structure evolution. *Cognitive Science, 28*(2), 259–287.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926.

Saffran, J., Newport, E., & Aslin, R. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*(4), 606–621.

Saffran, J., Pollak, S., Seibel, R., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition, 105*(3), 669–680.

Seidenberg, M., & Elman, J. (1999). Do infants learn grammar with algebra or statistics? *Science, 284*(5413), 433.

Shastri, L. (1999). Infants learning algebraic rules. *Science, 285*(5434), 1673.

Shultz, T. R. (1999). Rule learning by habituation can be simulated in neural networks. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the 21st annual conference of cognitive science society*. Mahwah, NJ: Erlbaum.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3), 1558–1568.

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

Tomasello, M. (2003). *Constructing a language*. Cambridge, MA: Harvard University Press.

Toro, J., Nespor, M., Mehler, J., & Bonatti, L. (2008). Finding words and rules in a speech stream. *Psychological Science, 19*(2), 137.

Tyrell, D. J., Stauffer, L. B., & Snowman, L. G. (1991). Perception of abstract identity/difference relationships by infants. *Infant Behavior and Development, 14*, 125–129.

Tyrell, D. J., Zingaro, M. C., & Minard, K. L. (1993). Learning and transfer of identity–difference relationships by infants. *Infant Behavior and Development, 16*, 43–52.

Wallis, J., Anderson, K., & Miller, E. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature, 411*(6840), 953–956.