Variability, negative evidence, and the acquisition of verb argument constructions

Amy Perfors
Department of Psychology
University of Adelaide

Joshua B. Tenenbaum
Department of Brain & Cognitive Science
Massachusetts Institute of Technology

Elizabeth Wonnacott
Department of Experimental Psychology
Oxford University

We present a hierarchical Bayesian framework for modeling the acquisition of verb argument constructions. It embodies a domain-general approach to learning higher-level knowledge in the form of inductive constraints (or overhypotheses), and has been used to explain other aspects of language development such as the shape bias in learning object names. Here we demonstrate that the same model captures several phenomena in the acquisition of verb constructions. Our model, like adults in a series of artificial language learning experiments, makes inferences about the distributional statistics of verbs on several levels of abstraction simultaneously. It also produces the qualitative learning patterns displayed by children over the time course of acquisition. These results suggest that the patterns of generalization observed in both children and adults could emerge from basic assumptions about the nature of learning. They also provide an example of a broad class of computational approaches that can resolve Baker's Paradox.

Introduction

Human learning operates on multiple levels, from the induction of information idiosyncratic to particular items to the abstraction of general patterns or rules. A classic example of this can be found in the acquisition and use of verb argument constructions. In every language, different verbs take arguments in distinct constructions; for instance, the verb *load* is associated with two distinct locative constructions, as illustrated in the sentences 'He loaded apples into the cart' and 'He loaded the cart with apples.' However, not all verbs can occur in all constructions: *'He poured the cart with apples' and *'He filled apples into the cart', while perfectly understandable, are nevertheless ungrammatical. Speakers also generalize their usage of constructions beyond the set of verbs in which they have been encountered. Gropen, Pinker, Hollander, and Goldberg (1991) demonstrated that children would spontaneously produce sentences such as 'He is mooping the cloth with marbles' when introduced to the novel verb 'mooping' in the context of an experimenter placing marbles into a sagging cloth. This ability to generalize verbs to unattested constructions suggests that verb-general knowledge about construction usage exists in tandem with knowledge about patterns specific to individual lexical items.

Further empirical support for the idea that people are sensitive to information on both the specific and the general level comes from the literature on sentence processing. Verb-specific syntactic biases about which structures are likely to follow individual verbs have strong effects on real-time processing (e.g., Trueswell, Tanenhaus, & Kello, 1993; Snedeker & Trueswell, 2004). People are sensitive to verb-general effects as well: for instance, they will sometimes interpret nouns occurring after a verb as a direct object, even if that particular verb does not take direct objects (Mitchell, 1987). This effect is more common if the verb is low in frequency. The idea that learning occurs at both the item-specific level and at varying levels of generalization is also in line with the approach taken by 'usage-based' theories of language acquisition (Tomasello, 2003; Goldberg, 2006).

How is verb knowledge on multiple levels acquired? An experiment by Wonnacott, Newport, and Tanenhaus (2008) suggests that both levels can be learned on the basis of distributional statistics about syntactic information alone. Adults were presented with novel verbs and constructions in one of two artificial languages. Those taught what Wonnacott et al. (2008) termed the Generalist language saw eight verbs, all of which occurred in both of two novel constructions. In the Lexicalist language, each of the eight verbs occurred in only one construction. People were capable of learning verb-specific information about the construction patterns of each individual verb. They also acquired verb-general information, both about the distribution of each construction across the language as a whole as well as about the variability with which individual verbs matched that distribution. Speakers of the Lexicalist language assumed that a completely novel verb could occur in only one construction, while speakers in the Generalist language were happy to generalize it to both (with a bias toward the overall most frequent construction). This suggests that not only are people capable of learning some verb alternations on the basis of syntactic information alone, but also that they make inferences about distributional statistics on several levels of abstraction simultaneously. Similar results have recently been found with 5 and 6 year olds learning a language (Wonnacott & Perfors, 2009) in which nouns co-occur with 'particles' in patterns analogous to the verb/construction co-occurrences in Wonnacott et al. (2008).

These results pose a challenge to existing accounts of verb construction learning. Most theories address the issue of learning verb-general knowledge in terms of making abstract generalizations about verb classes or construction types – corresponding, for instance, to the realization that the dative alternation is associated with certain features and not others. The issue of learning on a higher level of abstraction – making inferences about the sort of feature variability one would expect across verbs or constructions in language in general – has not been considered. The results of the experiment by Wonnacott et al. (2008) suggest that listeners are highly sensitive to distributional cues and can use them to make productive generalizations in language. Yet we do not know exactly how useful this sort of learning would be given more naturalistic data.

More generally, we do not yet have a precise, rigorous account of how learners might be able to combine abstract inferences about feature variability with verb-general insights on the level of construction types and verb-specific acquisition on the level of individual lexical items. Can this be done in a way that is consistent with decades' worth of evidence about how verb constructions are acquired by children?

In this paper we present a computational model that addresses this problem, explaining the acquisition of verb constructions as a rational statistical inference. We use a hierarchical Bayesian model that has previously been applied to other aspects of cognitive development, such as acquiring the shape bias in word learning (Kemp, Perfors, & Tenenbaum, 2007). We demonstrate that this model captures the results of the artificial language learning experiments and also that, given naturalistic data, it accounts for several other characteristic phenomena observed in the verb acquisition literature. This work demonstrates how the patterns of overgeneralization observed in both children's speech and adult behavior might emerge from basic, domain-general assumptions about the nature of learning. It also suggests a possible resolution to Baker's Paradox, and explains some of the qualitative learning patterns displayed by children over the time course of acquisition.

Before presenting the particulars of the model, we will first situate it in terms of important issues that emerge from the study of verb argument constructions.

Baker's Paradox and the puzzle of verb learning

Much of the current work on the acquisition of verb constructions can trace its motivation to a more general learning problem – being able to rule out unseen examples as ungrammatical, while still being able to productively generalize. We can illustrate this with the phenomenon of construction alternation, an example of which is shown in Table 1. Given that many verbs in English occur in both the prepositional dative (PD) and direct object dative (DOD) constructions, it may seem natural to conclude that all verbs that occur in one are also grammatical in the other. Unfortunately, this generalization, known as the dative alternation, does not always apply: for instance, *confess* is grammatical in PD syntax ('Jonathan confessed the truth to Doug') but not in DOD (*'Jonathan confessed Doug the truth'). Despite never having been explicitly taught that *confess* is ungrammatical in the double object dative – and even though a near-synonym, *tell*, is grammatical – fluent speakers of English appear to have no trouble avoiding the incorrect form. This poses a learnability puzzle, sometimes referred to as Baker's Paradox (Baker, 1979; Pinker, 1989), which is a classic negative evidence problem: how do children know that some unobserved pattern is ungrammatical, rather than simply not yet heard?

Table 1: Constructions in the dative alternation.

Construction name	Abstract form	Example
Prepositional dative (PD)	NP_1 V NP_2 to NP_3	Debbie gave a pretzel to Dean.
Double object dative (DOD)	NP_1 V NP_3 NP_2	Debbie gave Dean a pretzel.

Although there is some evidence that children receive some negative evidence in the form of statistically different reactions to ill-formed utterances (e.g., Chouinard & Clark, 2003), it is unclear how they might be able to make use of this sort of evidence to resolve Baker's Paradox; a Gold-style ideal learner would not be able to converge onto the correct language given statistical evidence of this sort (Gordon, 1990), and as yet we are unaware of any formal, concrete proposals for what sort of learner could. It is also evident that children do not solve the problem through a strategy of conservatism, as suggested by the Subset Principle (Berwick, 1985), since there is a vast quantity of evidence suggesting that human learners do overgeneralize verb argument constructions (e.g., Pinker, 1989).

One possibility is that there might be *positive* evidence about which verbs enter into an alternation, in the form of morphological, phonological, or semantic features that are correlated with the verb's syntax (e.g., Pinker, 1989; Morgan & Demuth, 1996). This 'bootstrapping' hypothesis would overcome the learnability problem by providing one or more features that the child could use to distinguish between verbs that do and do not occur in a given alternation (Mazurkewich & White, 1984; Pinker, 1984). In fact, there is a strong correlation between verb meaning and the syntactic structures in which verbs occur (Fisher, Gleitman, & Gleitman, 1991). For example, while both PO and DOD syntax are associated with the meaning of transfer, the DOD form is specifically associated with the notion of transfer of possession. Thus verbs which clearly depict transfer of possession are more likely to occur in the DOD construction than verbs which depict motion; children are indeed sensitive to these cues when generalizing with novel verbs (Gropen, Pinker, Hollander, Goldberg, & Wilson, 1989; Ambridge, Pine, Rowland, & Young, 2008). However, such cues do not provide sufficient conditions for verb usage: for example, it is not clear why *kick* but not *carry* occurs in DOD syntax, when a kicking action is not inherently more likely to result in transfer of possession than a carrying one. Although some researchers have claimed that more fine-grained semantic and/or morpho-phonological distinctions can capture verb syntax (Pinker, 1989), others have pointed out the class-inclusion criteria are inconsistent (Bowerman, 1988; Braine & Brooks, 1995; Goldberg, 1995). Thus the learner must still acquire lexically based exceptions – and thus the learning 'paradox' remains. Moreover, it is demonstrably difficult to ascertain the meaning of many verbs based simply on environmental contingencies and co-occurrences (Gleitman, 1990; Gillette, Gleitman, Gleitman, & Lederer, 1991).

Another hypothesis, originally contributed by Braine (1971), suggests that learners might be able to use indirect negative evidence, inferring that if a certain form does not occur given enough input, then it is probably ungrammatical. One way this might occur is through pre-emption: if a verb is encountered in one construction, when another, more felicitous construction would provide the same information, the pragmatic conclusion would be that the unseen construction is actually ungrammatical (Goldberg, 1995). Children 4.5 years and older appear to be receptive to this sort of information (Brooks & Tomasello, 1999), but as of yet there is scant data from younger children.

Another form of indirect negative evidence is entrenchment – the idea that the more often a verb is observed in one or more constructions, the less likely it is to be generalized to new ones (Braine & Brooks, 1995). Consistent with this notion, both children and adults are more likely to rate overgeneralizations as grammatical if they occur with low-frequency rather than high-frequency verbs (Theakston, 2004; Ambridge et al., 2008). Even children as young as two and a half to three years of age are sensitive to frequency; the less frequent a verb is, the more likely children are to produce overgeneralizations (Brooks, Tomasello, Dodson, & Lewis, 1999; Matthews, Lieven, Theakston, & Tomasello, 2005). Note that using this type of indirect evidence does *not* require the child to make pragmatic judgments involving knowledge of the felicity conditions associated with constructions. Rather, they are required to make inferences from patterns of usage (like frequent occurrence in one construction but not the construction being considered).[1]

Both entrenchment and pre-emption can be understood in terms of competition: repeated usage of a verb in some construction(s) may oust the usage of that verb in another, competing, construction (see also MacWhinney, 2004). The model presented in this paper formally instantiates a version of competition and, in addition, incorporates the ability to learn about the variability of construction usage. This higher-level learning may constrain the extent to which the usage of the verb in one construction should block its usage in the other.

Bringing it all together

Our goal in this paper is to present a learning mechanism which meets the following criteria:

1. It learns abstract knowledge about the variability of verb constructions, as in the Wonnacott et al. (2008) experiment.
2. It solves Baker's Paradox, learning the correct alternation pattern in the absence of explicit correction, but without requiring direct negative evidence or employing a strategy of strict conservatism.
3. It combines verb-specific information about individual lexical items with verb-general information about construction-based classes of verbs.
4. It productively overgeneralizes verb constructions, and is more likely to do so if the verb in question is low in frequency.
5. It is capable of combining information from multiple features, including semantic and syntactic cues.

[1] A closely related notion is what Goldberg (2006) calls "statistical pre-emption." This conception of pre-emption emphasizes statistical inference based on frequently encountering a verb in one construction when its use in another would satisfy the functional demands of the situation. Goldberg argues that statistical pre-emption only occurs between functionally related constructions, in contrast to some views of entrenchment, which suggest that repeatedly encountering a verb in *any* construction can lead to its being 'entrenched' in that construction, thereby preventing its extension to new structure (Braine & Brooks, 1995). Since none of the models we consider in this paper have semantics associated with constructions, we do not distinguish between the entrenchment and the statistical pre-emption hypotheses here.

We present a domain-general hierarchical Bayesian model that fulfills these desiderata. It explains how abstract inferences about feature variability may be combined with verb-general acquisition on the level of construction-based verb classes and verb-specific knowledge on the level of individual lexical items – and does so in a way that also captures the other qualitative learning patterns found in the literature. In Study 1 we address the first desideratum by presenting our model with the same artificial language input received by adult subjects in the experiments. Study 2 demonstrates that when given input representing the syntactic distribution of verbs occurring in the dative alternation in a corpus of child-directed speech, our model qualitatively captures the final three desiderata. These results are interesting in light of the fact that our model was originally developed to capture the emergence of feature biases in word learning, rather than anything particular about verb learning or verbal knowledge at all. Finally, in Study 3 we explore how learning is affected by semantic information being added to the input. Though extra-syntactic cues are not necessary to resolve Baker's Paradox, the model is capable of using semantic class information to make syntactic generalizations in a sensible manner, even without making strong assumptions about the nature of the semantic representations in question.

Study 1: Modeling adult artificial language learning

Model

We use a hierarchical Bayesian model (HBM), which supports the simultaneous acquisition of multiple levels of knowledge: both concrete and item-specific, as well as abstract and general. Goodman (1955) provided an example of this type of learning. Suppose we have many bags of colored marbles and discover that some bags have black marbles while others have white marbles. However, every bag is uniform in color; no bag contains both black and white marbles. Once we realize this, we have acquired knowledge on two levels: the item-based knowledge about the color of marbles in each particular bag, but also the higher-level knowledge (called, following Goodman, an *overhypothesis*) that bags tend to be uniform in color. This higher-level knowledge allows us to make inferences given very small amounts of data: for instance, given a new bag from which one black marble has been drawn, we can infer that all of the other marbles in the bag are probably black, too.

This schematic example is analogous to the situation confronted by the verb learner, where 'bags' become 'verbs' and 'marbles' become 'constructions.' A learner might acquire verb-specific knowledge about which constructions are associated with which specific lexical items, but she might also learn verb-general knowledge about how uniformly constructions are spread over verbs in general. Just as each bag was associated with one color of marble, does each verb tend to be associated with one constructions? Or do verbs tend to be alternating – grammatical in more than one construction? Learning overhypotheses about verbs and their constructions can enable a learner to answer these questions, and to constrain generalization of new verbs in just the same way that learning overhypotheses about bags of marbles constrains generalizations when presented with a new bag.

We depict this type of learning graphically in Figure 1(a) and formalize it as follows. Level 1 knowledge about how often each verb occurs with each construction (or marbles of each color were drawn from each bag) is represented by $\boldsymbol{\theta}$, and is acquired with respect to a more abstract both of knowledge, Level 2 knowledge, which in this case is knowledge about the distribution of verb constructions. It is represented in our model by two parameters, α and $\boldsymbol{\beta}$: roughly speaking, α captures the extent to which each individual verb occurs uniformly in one construction (or not), and $\boldsymbol{\beta}$ captures the overall frequency of each construction, independent of any particular verb.[2]

Level 2 knowledge depends on knowledge at a higher level, Level 3, which is represented in our model by two (hyper-)parameters λ and μ. They capture prior knowledge about α and $\boldsymbol{\beta}$, respectively: the range of values expected about the uniformity of constructions within the verb (λ), and the range of values of the expected distribution of verb constructions across the language (μ). In principle, we could extend the model to contain arbitrarily many levels, not just two or three; each subsequent level encodes ever-more abstract knowledge consisting of (generally weak) expectations about the nature of the knowledge on the next lowest level. Thus, Level 4 knowledge would capture prior knowledge about λ and μ: expectations about the expectations about uniformity of verbs and the distribution of verb constructions.

The idea of wiring in abstract knowledge at higher levels of hierarchical Bayesian models may seem reminiscent of nativist approaches to cognitive development, but several key features fit well with empiricist intuitions about learning. In HBMs, the top level of knowledge is always prespecified implicitly or explicitly, but every level beneath that can be learned. As one moves up the hierarchy, knowledge becomes increasingly abstract and imposes increasingly weak constraints on the learner's specific beliefs at the bottom, most concrete level of observable data. Thus, a version of the model that learns at higher levels (e.g., acquiring hyperparameters λ and μ as well as α and β) builds in weaker constraints than a version that learns only at lower levels (acquiring only α and β). In general, we can come ever-closer to the empiricist ideal of bottom-up, data-driven learning by adding levels to an HBM and incorporating pre-specified knowledge only at the highest, most abstract level. Almost everyone along the empiricist-nativist continuum agrees that there must be some innate constraints; the question is how strong and how domain-specific those constraints are. HBMs provide a means for learning relatively strong, domain-specific knowledge given only weakly specified, more domain-general prior knowledge.

Learning in an HBM corresponds to making inferences about higher-level parameters based on data observed from verbs occurring in the input (the constructions that verb i occurs in are denoted $\boldsymbol{y^i}$). Generalization corresponds to making predictions about the parameters of novel verbs; for instance, given a new verb $\boldsymbol{y^{new}}$, the model makes inferences about the most likely verb-specific distribution over constructions $\boldsymbol{\theta^{new}}$ based on the combination of the observations and the inferred higher-level knowledge about verbs in general. Verbs are generalized assuming that new instances will match the inferred construction distribution: if the model infers that $\boldsymbol{\theta^{new}} = [0.6\ 0.4]$ (that is, that the new verb will occur 60% of the time in construction 1 and 40% of the time in construction 2), then we say that 60% of the tokens it 'produces' will occur in construction 1 and 40% of them will occur in construction 2. Generalization is calculated by performing a stochastic search over the space of parameter values. Appendix 1 contains further technical details.

[2] One way of thinking about the relationship between θ, α, and $\boldsymbol{\beta}$ is that α captures how close, on average, each individual θ is to $\boldsymbol{\beta}$ (i.e., how close each individual verb's construction distribution is to the overall distribution across all verbs). Thus, in the Generalist language, where every θ is equal to $\boldsymbol{\beta}$, α is very high; in the Lexicalist language, where each θ is very different from $\boldsymbol{\beta}$, α is very low.

In this section of the paper (Study 1) we consider two closely related models, which will serve as the basis for the expanded models introduced in Study 2. The models are somewhat simplistic, in that they can acquire knowledge and perform inferences on multiple levels but do not have the capability to group verbs into classes (as the models in Study 2 do). Nevertheless, we begin with these in order to illustrate the utility of learning on multiple levels, and to provide a point of comparison to models that can learn verb classes.

Model L2 is equivalent to that specified in Kemp et al. (2007); it assumes that the Level 3 knowledge (λ and μ) is already known, and learns the α and $\boldsymbol{\beta}$ values that maximize posterior probability for the given data. Model L3, by contrast, learns λ and μ in addition to α and $\boldsymbol{\beta}$, and assumes that knowledge at higher levels (above λ and μ) is given. Model L3 is apt to be useful in those situations in which the inferred values at Level 2, α and $\boldsymbol{\beta}$, tend to be unlikely given the built-in prior knowledge at Level 3 (λ and μ). If Model L2 contains built-in knowledge that weakly favors the conclusion that constructions tend to be uniform within verbs, but is presented with a dataset in which they are not, it will learn that dataset relatively poorly; if the built-in knowledge favors the conclusion that constructions are evenly spread within verbs, it will be slower to learn datasets in which each verb is associated with only one construction. Model L3 is therefore more flexible than Model L2, since it can 'tune' the higher order parameters at Level 3, and therefore will be able to learn both datasets equally well. Both models acquire Level 1 knowledge about the expected constructions found in specific individual verbs.

Data

The purpose of the artificial language learning experiment[3] of Wonnacott et al. (2008) was to determine whether adults exposed to a novel language could acquire distributional knowledge at both the verb-specific and verb-general levels. Over the course of five days, subjects were taught a language with five novel nouns and eight novel verbs occurring in one of two possible novel constructions: *VAP* (verb agent patient) and *VPA-ka* (verb patient agent particle (*ka*)). During training, participants were presented with a set of scene/sentence pairs, hearing sentences corresponding to video clips of scenes in which puppets acted out the sentence meaning. Because part of the purpose of the experiment was to explore performance given only syntactic information, both constructions had the same meaning.

Subjects were divided into two conditions. In the Generalist condition, each of the eight verbs occurred in both constructions, but seven times as often in the *VPA-ka* construction as in the *VAP*. In the Lexicalist condition, seven verbs occurred in the *VPA-ka* construction only, and one verb occurred in the *VAP* only. In both conditions, the absolute and relative frequencies of the *VAP* and *VPA-ka* constructions are the same, but the conditions differ widely in terms of the distribution of those constructions across individual verbs; this allows for the evaluation of whether learners can acquire and use verb-general as well as verb-specific statistical information. When presented with a novel verb in one construction, would participants in the Generalist condition be apt to infer that it can occur

[3] For simplicity of presentation, we focus on Experiment 3 in their paper, although the model captures the results of all three experiments, each of which focuses on the acquisition of knowledge about variability across individual verbs.

in both? And would participants in the Lexicalist condition tend to think that it occurs only in one?

To test this, participants' productive language use was evaluated in a procedure in which subjects viewed a scene, heard the verb corresponding to the action in the scene, and were asked to complete the sentence aloud. Each of the four novel verbs – which had not been heard during training at all — occurred four times: two of the verbs only in the *VAP* construction, and the other two only in *VPA-ka*. The subjects' performance is shown in Figure 2a. People in the Generalist condition were likely to produce a novel verb in both constructions, matching the overall frequency of each in the language, rather than the single construction it was heard in. People in the Lexicalist condition, whose previous input consisted of verbs that occurred in only one construction, tended to produce the novel verb only in the single construction in which it was observed.

Results

We present Models L2 and L3 with data corresponding to the input given to adult subjects over the course of the five days of training in the Generalist and Lexicalist conditions. To evaluate generalization beyond learned lexical items, the models are also presented with a single exemplar of a completely novel verb occurring either in the *VAP* or *VPA-ka* construction.[4] Results are shown in Figure 2. Both models replicate the difference between conditions, demonstrating that the model makes inferences much as humans do. Novel verbs in the Generalist condition are assumed to occur in both constructions, while the same novel verbs in the Lexicalist condition are assumed to occur in only one. The models have abstracted information about verb-general variability, and have used that as the basis of productive generalization of novel input.

Model L3 outperforms L2 in the Generalist condition, more accurately predicting human production for the novel verb occurring in the less-frequent construction in the language (*VAP*). Even though that verb was heard in the *VAP* construction only, both humans and Model L3 predict that it will occur in the other (*VPA-ka*) construction nearly 87.5% of the time (which is the base rate of *VPA-ka* in the language as a whole). Although Model L2 qualitatively captures the difference between the Generalist and Lexicalist conditions, it is quantitatively less accurate, extending the *VAP* form to *VPA-ka* 60% rather than 87.5% of the time. The reason for this difference is that the hyperparameters (λ and μ) over α and $\boldsymbol{\beta}$, which are 'built in' for Model L2, weakly restrict the range of α and $\boldsymbol{\beta}$ to avoid extreme values. In this case, the built-in hyperparameters for Model L2 implement a slight bias for values of α and $\boldsymbol{\beta}$ that correspond to the assumption that verbs are more likely to occur in only one construction. The Generalist condition, in which each of the individual verbs' distribution of constructions precisely mirrors the distribution in the language as a whole, is thus best captured by values of α and $\boldsymbol{\beta}$ that happen to be dispreferred by the hyperparameters of Model L2. One might select different hyperparameters, but that would be arbitrary and *post hoc*; it could also cause the model to be less accurate at capturing the Lexicalist language. By contrast, because Model L3 can learn the appropriate bias about the uniformity of constructions across verbs (i.e., it can learn hyperparameters λ and μ), it

[4] Humans were presented with four tokens of a single novel verb, but because the model does not have the memory limitations that humans do, it is more appropriate to evaluate its generalization given only one token.

infers that the more extreme values are more appropriate for this dataset. In other words, Model L3 (which builds less in, since it learns the hyperparameters that are 'innate' for Model L2) does better, precisely because it has more flexibility for capturing the data.

In the all of the subsequent sections, both Model L2 and L3 were analyzed, and results were qualitatively similar for both. For space and readability reasons, we report only the results from Model L3, which usually slightly outperforms Model L2.

Study 2: Modeling the dative alternation

Model

We have demonstrated that both models can acquire verb-general variability information, just as adults do when presented with artificial language data. However, in natural language, verb-general statistics may be shared among only a subset of verbs rather than over all the verbs in the language. For instance, some verbs occur in both constructions in the dative alternation, but others occur in only one. A learner that could only make inferences about verb-general statistics across the language as a whole would not be capable of realizing that there were these two types of verbs. Presented with a novel verb occurring only once in one construction, such a learner might be more likely to generalize it to both than one who realized that it might belong to a non-alternating class. Would a model that can make inferences about classes of verbs perform significantly better on real-world data than a model without classes?

An intuitive way to address this possibility is to add the ability to discover verb classes[5] to both Models L2 and L3. We denote this extension with the prefix K (i.e., Model K-L2 and K-L3) and depict it graphically in Figure 1(b). As in Kemp et al. (2007), this results in a model that assumes that verbs may be grouped into several classes, where each class is associated with its own hyperparameters. The model is not told how many classes there are, nor which verbs occur in which class; instead, it forms the classes based on the data, in combination with a prior under which all class assignments are possible but fewer classes are favored. The goal of learning is to simultaneously infer how verbs are assigned to classes, along with the values of the hyperparameters that describe each class. The model extension is described more fully in Appendix 1.

Data

We explore the acquisition of verb constructions by presenting the model with real-world data taken from a corpus of child-directed speech. Because the dative alternation is a central, well-studied example relevant to Baker's Paradox, we choose to focus on verbs that occur in it. The data is collected from the sentences spoken by adults in the Adam corpus (Brown, 1973) of the CHILDES database (MacWhinney, 2000), and consists of the counts of each construction (PD and POD) for each of the dative verbs (as listed in Levin (1993)) that occur in the corpus. Note that we do not consider all of the constructions in the language – only these two alternating ones. We return to this point later.

An additional variable of interest is what sort of evidence may be available to the child at different ages. This can be loosely approximated by tallying the number of occurrences of

[5] Because 'class' is more sensible nomenclature for verbs, we will refer to them as classes throughout this paper, but these are the same entities referred to as 'kinds' in the Kemp et al. (2007) work.

each verb in each construction in subsets of the corpus split by age (see Table 3 in Appendix 2). The Adam corpus has 55 files, so the first segment, Epoch 1, contains the verbs in the first 11 files. The Epoch 2 corpus corresponds to the cumulative input from the first 22 files, and so on up until the full corpus at Epoch 5. The dative verbs in the first file only, corresponding to approximately one hour of input, constitute Epoch 0.

Results

Figure 3 shows class assignments predicted by Model K-L3. It captures the intuitively sensible pattern: those verbs that occur in one construction tend to be found in a separate class from verbs that occur in both. When there is less data (i.e., at the earlier Epochs), the model is less certain: the class assignments for subsets of the full corpus are generally less sharp than they are for the entire corpus. Frequency also plays a role; the model is more certain about the class assignments of high-frequency verbs like *give* and *make*, and much less confident about the class assignments of low-frequency verbs like *pay.* In part because of this lack of certainty, we would expect the model to be more likely to overgeneralize the low-frequency verbs beyond the constructions in which they occur in the input.

There are two ways of testing this prediction. First, we can examine model predictions for how to produce novel instances for each of the input verbs. These results are shown in Figure 4. It is evident the model overgeneralizes more often for the low-frequency verbs. The predicted construction distribution for high-frequency verbs like *give* or *make* is very similar to the observed distribution (shown in the associated pie chart). But low-frequency verbs like *explain* or *sing*, which only occur in one construction in the input, are somewhat likely to be produced in the other construction. This is because there is still some possibility that they are actually members of the alternating class; as more and more verb tokens are heard and these verbs are still only heard in one construction, this becomes less and less likely.

Second, instead of exploring generalization on all verbs, we can also focus on the subset of non-alternating verbs to explore *over*-generalization – the degree to which the model generalizes each verb to a construction in which it has never occurred – as a function of verb frequency. Figure 5 shows the degree of overgeneralization for each of the verbs that occurred in just one construction at each Epoch in Model K-L3. This is calculated by finding the difference between the proportion of times the verb is observed vs. predicted in the DOD construction (although, since there are only two constructions, one could equivalently calculate this for the PD construction). If this difference is zero then it means the model produces the verb constructions precisely at the same frequency as they occurred in the corpus. The larger this difference is, the more the model has 'smoothed', or overgeneralized away from, the observed data.

The results indicate that, as the frequency of the verb increases, overgeneralization decreases: the difference between observed and predicted approaches zero. There is also an interaction with Epoch: verbs of equivalent frequencies are overgeneralized more in earlier Epochs. For instance, verbs that occur once in the full corpus are overgeneralized less than one-third as often as verbs that occur once at Epoch 2. The reason for this is that there is more data in the corpus at later Epochs, and the model is therefore more certain about the probable constructions it infers for even the low-frequency verbs.

The model appears to be learning in the absence of negative evidence: without receiving any correction or being explicitly told that some verbs are non-alternating, the model eventually forms alternating and non-alternating classes. This qualitatively captures two of the major phenomena found in the acquisition of verb argument constructions: more frequent verbs being overgeneralized more rarely, and a general decrease of overgeneralization with age.

Is the success of this model due to the fact that it can group verbs into classes? We address this question by comparing the performance of the class-learning model K-L3 with Model L3 from Study 1, the results of which are shown in Figure 6. Removing the ability to learn verb classes, so that the model does not discover the classes of non-alternating verbs, does increase overgeneralization overall. This is because when there are no classes verbs are more strongly influenced by the behavior of *all* of the other verbs in the language, including those that alternate. Nevertheless, both models differentiate between verbs based on their frequency and age: verbs encountered more often in only one construction are less likely to be overgeneralized to a different construction. The class-learning model performs better because a novel verb which has been encountered only in one construction is with some probability assigned to the same class as the more frequent non-alternating verbs; this increases the probability that the novel verb is believed by the model to be non-alternating itself.

Which of the two models captures human learning better is currently unknown, although the class-learning model does more strongly limit overgeneralization and changes more across EPOCHS. For the present, we emphasize that many of the qualitative effects are similar for both models. Each captures three of the major empirical phenomena: learning in the absence of overt negative evidence, and decreasing overgeneralization with increasing age as well as verb frequency. This is because these aspects of model performance result from general characteristics of Bayesian learning, rather than particular assumptions made by any specific analysis. We address this point in more detail in the Discussion section.

Study 3: Exploring the role of semantics

Many approaches to resolving Baker's Paradox have relied on the presence of semantic or morpho-phonological feature(s) to distinguish verbs that take part in an alternation from verbs that don't. However, as discussed in the introduction, it is unclear whether such features are always available: some arbitrary verb-specific alternations may always need to be acquired, and so the logic underlying Baker's Paradox remains. The models presented thus far demonstrate that Baker's Paradox may be resolvable on the basis of the syntactic distribution of constructions and verbs only.[6] Nevertheless, there are strong correlations between verb semantics and verb syntax; in particular, verbs that undergo an alternation may have different semantic properties from verbs that do not. Many computational models that address Baker's Paradox, while not endorsing the claim that verb semantics fully determines the syntactic distribution of verbs, have made use of these correlations in learning.

[6] Note that we are not claiming that the syntactic constructions *themselves* are necessarily learnable in the absence of semantics. Acquiring the DOD construction must involve learning the mapping between NP_1 V NP_3 NP_2 syntax and thematic role assignment (i.e., semantics). As we elaborate more fully in the discussion, our work does not address this type of learning and assumes a pre-existing knowledge of constructions. Our point is simply that Baker's Paradox – the problem of overgeneralization *given* a set of constructions – can be resolved without assuming that verbs that do not enter into the alternation share extra-syntactic features.

It is therefore important to determine what such cues could contribute in the current model. Moreover, since there is also clear evidence that human generalization can be affected by semantics, it is interesting to evaluate whether our model can use semantic information in a sensible way. Although it is beyond the scope of our simple model to address all of the subtleties of natural language semantics, we can still explore the effect of adding extra-syntactic cues in a manner which captures the overall pattern envisaged by Pinker and colleagues. Given that Study 1 and 2 have suggested that at least some aspects of the no negative evidence problem are in principle solvable based on syntactic input alone, what is the impact of learning when verbs with the same syntactic distribution share semantic features? To what extent does our model capture human generalization based on verb semantics?

Model

The models in Study 1 and Study 2 incorporate only a single feature, syntax, but can be trivially extended to include multiple features. We assume that each feature is independently generated, which allows inference to proceed exactly as before except that each α and $\boldsymbol{\beta}$ is learned separately for each feature. The posterior probability for the full model is therefore the product of the probabilities along each feature. Appendix 1 contains additional details about this extension.

Data

We evaluate how semantics may be incorporated into the model by adding to our corpus of dative verbs a semantic feature which we associate with the different classes of verbs. Note that the constructions themselves therefore have no semantics *per se*, which is a simplification.[7] The feature precisely parallels the semantics of each of the three classes and therefore has three possible values: one corresponding to the class of alternating verbs (which we will call semantic class A), one to those verbs occurring only in PD syntax (semantic class P), and one to verbs occurring only in DOD syntax (semantic class D). For instance, *give*, which occurs 106 times in DOD syntax and 44 times in PD syntax, has a semantic feature occurring 150 times in semantic class A; *make*, which occurs 11 times in DOD syntax, occurs 11 times in semantic class D; and *say*, which occurs 6 times in PD syntax, has a semantic feature occurring 6 times in semantic class P.

Though this instantiation of semantics is highly simplistic, it has the merit of allowing for a clear exploration of precisely how one might generalize when some features associated with verbs are (and are not) correlated with verb syntax. Essentially, we begin by exploring a best-case scenario in which the feature in question is perfectly correlated with the syntactic verb types; later in this section we relax this assumption.

We predict that adding semantic features will not qualitatively change learning but will result in less overgeneralization of existing verbs, since the model can use the additional semantic information to make more accurate inferences about how each verb should be

[7] Researchers acknowledge that constructions have some semantics independent of the verbs they contain, so that the usage of a verb in a construction partly depends on the fit of that verb's semantics with the construction semantics. Pinker (1989) also points to semantic features of verbs which correlate with syntax but are independent of construction semantics (e.g., a class of verbs (including *throw* or *kick*) that involve 'instantaneous imparting of force in some manner causing ballistic motion' and that all take part in the PD and DOD alternation). Our model follows Pinker in associating features directly with the verb.

Table 2: Conditions based on semantic and syntactic features of novel verb.

	Syntactic form	
Semantic form	PD	DOD
D	Other Non-Alternating	Same Non-Alternating
A	Alternating	Alternating
P	Same Non-Alternating	Other Non-Alternating

classified. We also wish to test whether the model generalizes new verbs in a way that is qualitatively similar to children's performance in experiments such as Gropen et al. (1989), even given the simplistic semantic representations we used. To test this, we added one additional novel verb to the input for the model. The six conditions differ in the nature of that novel verb, as shown in Table 2. In three conditions the syntax of the novel verb is DOD, and in the other three it is PD. Each syntactic form is paired with each semantic form. When the syntactic form corresponds to the same semantic form it matches in the input data, we refer to that as the SAME NON-ALTERNATING FORM condition (i.e., PD syntax, semantic class P; and DOD syntax, semantic class D). If the semantic form occurs in the alternating class in the input, that is the ALTERNATING FORM condition (i.e., both PD and DOD syntax, semantic class A). And if the semantic form and syntactic form conflict based on the input corpus, we refer to that as the OTHER NON-ALTERNATING FORM condition (PD syntax, semantic class D; DOD syntax, semantic class P). As a baseline, we compare these six conditions to the sort of generalization that occurs when there are no semantic features at all.

Results

As predicted, the model shows less overgeneralization than the equivalent model with no access to semantic features. In addition, as illustrated in Figure 7(a), Model K-L3 generalizes based on semantic features in a sensible way. Both the model and the children in the Gropen et al. (1989) experiment are more likely to produce the construction they were not presented with if its semantic feature is associated with a different class of verbs. Sensibly, if the semantic feature matches the syntax of the same non-alternating form, then the model rarely produces the unattested construction.

The fact that the success of the model depends on its ability to form separate verb classes is apparent when we examine the performance of Model L3, shown in Figure 7(b). This model, which cannot form separate classes, generalizes each novel verb identically, regardless of its semantic features. In essence, the class information is the vehicle for capturing the relationship between semantic features and construction usage. A more sophisticated model – for instance, one that can associate semantic features directly with constructions – might allow for the correct semantic generalizations even without class information, but our work demonstrates that semantic effects can arise given very simple assumptions about semantics as long as there is some mechanism for relating the verbs that share a similar construction distribution.

It is unrealistic to assume the existence of one semantic feature that precisely follows the correct verb classes: our purpose was simply to evaluate the extent to which the presence of such a feature would affect learning. It is probably more accurate to assume that there are many different features, each somewhat noisy, that are only statistically associated with the correct classes. When this is the case, do the semantics continue to aid in learning? Does our model produce sensible generalizations even if the semantic features are less clean?

To evaluate this we present the model with the same corpus, but this time associate each verb with three semantic features rather than one. In addition, each feature is associated with the correct verb class 60% (rather than 100%) of the time. The results in the same generalization task as before, shown in Figure 8, are qualitatively identical to the previous results, although noticeably noisier. As before, there is less generalization than the equivalent model with no access to semantic features, though there is more generalization than for the previous semantically augmented model. It is evident that it is not necessary for the semantic feature(s) to be perfectly clean in order to qualitatively capture the same generalization patterns.

Even this version of the model is a gross simplification: certainly, one of the things that makes verb learning difficult is that there are many features (semantic and otherwise) that are simply irrelevant, uncorrelated with the correct class assignments, or perhaps even correlated with other regularities among the verbs. In terms of our model, it is always possible to identify an extreme at which the additional features are noisy enough – or pick out other categories strongly enough – to completely eliminate any effects of the semantic and syntactic features. For instance, if there are 13 semantic features, six of which are consistent with an interpretation in which the verbs all belong to the same class and seven of which vary randomly, model performance is ruined. In general, correct generalization is a function of the coherence and number of features that pick out the correct class assignments; the model is somewhat robust to noise and error, but is not infinitely so (nor should it be). Further work is necessary to understand precisely how and to what extent additional features matter – to flesh out what the shape and nature of that ‘generalization function’ is. We have demonstrated here that in at least some situations, the model is capable of qualitatively capturing human generalization patterns on the basis of semantic features.

Discussion

In this paper we have presented a domain-general hierarchical Bayesian model that addresses how abstract learning about feature variability can be combined with verb-general learning on the level of construction-based verb classes and verb-specific learning on the level of individual lexical items. It captures the qualitative patterns exhibited by adults in an artificial language learning task, as well as those exhibited by children over the course of acquisition. Our model suggests that Baker’s Paradox can be resolved by a certain kind of learner, even based on syntactic-only input. Furthermore, it does so in a (largely) domain-general way, without making strong language-specific representational assumptions about verb constructions.

In the following section we evaluate our conclusions in more detail. We take special care to orient this research with respect to other computational models of verb argument construction learning in order to highlight the contributions of our approach, as well as some of its limitations.

A solution to Baker's Paradox?

One implication of our work is that it is may not be necessary to rely on non-syntactic features in order to solve Baker's Paradox (although such information might still be valuable for other important aspects of verb learning, like identifying the alternations in the first place - a point to which we return below). Our Bayesian learner, given the syntactic information from a corpus of dative verbs used in child-directed speech, resolves the negative evidence problem: it correctly realizes that that verbs that have been observed often in one construction but never in another probably are not grammatical in both, but that verbs that have been observed rarely in one construction and never in another might be. In essence, our learner takes indirect negative evidence into account by formally instantiating the notions of entrenchment/pre-emption (or, more broadly, competition), as suggested by other researchers (Braine, 1971; Braine & Brooks, 1995; Goldberg, 1995; MacWhinney, 2004). Each time a verb is encountered in one of the two competing structures, it is *not* encountered in the other, and this provides cumulative evidence against a grammar that allows this usage. Consistent with this, our model – like people – is more apt to overgeneralize lower-frequency verbs, and more likely to overgeneralize all verbs earlier in the process of acquisition. Adding correlated semantic features boosts learning but does not qualitatively alter its pattern.

This performance is not an idiosyncratic property of specific choices made in setting up our model, but is rather the result of a general property of optimal inference. We can illustrate this abstractly using schematic dot diagrams as in Figure 9, where each datapoint (e.g., a verb usage) is represented by a dot generated by some underlying process (i.e., underlying verb knowledge, perhaps instantiated by a rule of some sort). The job of the learner is to evaluate hypotheses about which process best describes the observed data, and we can represent different processes as different subsets of space. To perform this evaluation, a rational learner should trade off the complexity of the hypothesis (captured via the prior in Bayesian learning) with how well it predicts the observed data (captured by the likelihood). Bayesian inference recognizes that a hypothesis that is too complex for the observed data will overfit, missing important generalizations, while one that is insufficiently complex will not be explanatory enough. As shown in Figure 9, this will result in a preference for a hypothesis that is neither too simple (Hypothesis A) nor too complex (Hypothesis C), but 'just right' (Hypothesis B).

One implication is that a distinctive pattern of reasoning naturally emerges as the amount of data changes. When there are few datapoints, the simpler theories are favored, resulting in a tendency toward overgeneralization, as we saw in Study 2. As the number of datapoints increases, the likelihood increasingly favors the theory that most closely matches the observed data, and overgeneralization decreases. This captures the notion of a suspicious coincidence, since hypotheses that predict the observation of datapoints that in fact never occur tend to be increasingly disfavored. It also provides a natural solution to the problem of deciding among hypotheses given positive-only examples. As the size of the dataset approaches infinity, a Bayesian learner rejects larger or more overgeneral hypotheses in favor of more precise ones. But with limited amounts of data, the Bayesian approach can make more subtle predictions, as the graded size-based likelihood trades off against the preference for simplicity in the prior. The likelihood in Bayesian learning can thus be seen as a principled quantitative measure of the weight of implicit negative evidence – one that explains both how and when overgeneralization should occur.

Because this pattern is a general property of Bayesian inference, other computational approaches to the acquisition of verb argument constructions provide the same natural solution to Baker's Paradox; indeed, our model is an instance of a class of computational models which explicitly explain the acquisition of verb knowledge in terms of rational statistical inference, using either the Bayesian or the closely related minimum description length (MDL) framework (Dowman, 2000; Onnis, Roberts, & Chater, 2002; Chater & Vitànyi, 2007; Alishahi & Stevenson, 2008; Hsu, 2009). For instance, Dowman (2000) compares toy grammars with and without subclasses of non-alternating verbs and finds that as the amount of data increases, a more complex grammar is preferred and overgeneralization disappears. This work involves a simplistic toy grammar segment and an idealized artificial corpus rather than the more naturalistic child-directed data considered in our work, but both models show the same ability to deal sensibly with the problem of negative evidence. More similarly to our work, Onnis et al. (2002) use a Bayesian model to demonstrate the learnability of an alternation based on statistics from corpora of child-directed speech. Their model succeeds in this for the same reason ours does. Our model makes different (in many ways more domain-general) representational assumptions, and is in other ways more flexible and powerful, with the ability to learn on multiple levels of abstraction, and the ability to determine flexibly how many classes of verbs there are. But in terms of the problem of negative evidence, all of these models – ours included – solve it in the same way. In fact, even connectionist models (e.g., Allen & Seidenberg, 1999) implicitly incorporate a sort of tradeoff between complexity and goodness-of-fit. Often the tradeoff is non-optimal, since the preference for simplicity emerges out of choices about network architecture, number of training epochs, and other modelling choices rather than the mathematics of probability theory, but as long as any tradeoff is being made, overgeneralization will decrease with increasing amounts of data; the difference is that in Bayesian models that tradeoff is statistically optimal.

More generally, all of these models are examples of the notion of competition, which was first incorporated into the Competition Model (MacWhinney, 1987, 2004), though the same notion is also implicit in the 'sieve' model of Braine (1971). Competition results from the two opposing pressures of episodic and analogical support (MacWhinney, 1987, 2004). Episodic support consists of item-specific knowledge (which pressures a learner towards conservativism and lack of generalization), and analogical support consists of the tendency to form generalizations based on analogies to other verbs (which pressures a learner towards overgeneralization). The preference for simplicity serves the same functional role as analogical support, and the preference for goodness-of-fit to the data serves the same functional role as episodic support. Bayesian and MDL approaches quantify the statistical tradeoff between the two preferences in a precise way, but the idea of competition provides a unifying framework for conceptualizing all of the major approaches.

Abstract learning about feature variability

The primary feature distinguishing our model from other approaches, and from previous Bayesian/MDL-based work in particular, is our focus on learning at multiple levels of abstraction simultaneously (but see also Hsu (2009) for another application of this model to the task of learning verb alternations). In particular, our model can learn abstract knowledge about variability across verb types, just as adult subjects do in the experiment performed by Wonnacott et al. (2008). Both humans and our model acquire different generalizations based on whether the input comes from a language in which all verbs occurred in both constructions or a language in which each verb occurred in only one construction. In essence, the model is able to quantify the extent to which encountering a verb in one structure should 'count against' its potential future occurrence in the other. This sort of statistical learning is part of a larger process of balancing information on multiple levels of abstraction, and our work demonstrates how this balance may be achieved.

Although the importance of tracking both item-specific and verb-general information is acknowledged by many researchers (e.g., Braine, 1971; MacWhinney, 1987, 2004, among many others), many previous models address the issue of learning lexically-specific and detailed verb information (e.g., Dominey, 2003), and several models are capable of learning verb-general information about classes, features, or construction types (e.g., MacWhinney, 1987; Dowman, 2000; Onnis et al., 2002; Alishahi & Stevenson, 2008), previous models have not attempted to explain the acquisition of knowledge of *variability* at the level of verb classes or constructions. For example, the work by Dowman (2000) and Onnis et al. (2002) is focused mainly on the problem of negative evidence, and compares segments of toy grammars of varying degrees of complexity. The work by Alishahi and Stevenson (2008), which is an impressive model capable of capturing many aspects of semantic and syntactic acquisition, nevertheless has not attempted to account for higher-order knowledge about feature variability (necessary for explaining human behavior in the Wonnacott et al. (2008) task). That model has numerous similarities with ours: both are Bayesian, both can flexibly determine how many constructions or verb classes are appropriate given the data, and both capture similar qualitative patterns in acquisition. However, because their model does not do inference on the highest levels (Level 2 and Level 3, corresponding to hyperparameters α, $\boldsymbol{\beta}$, λ, and $\boldsymbol{\mu}$ in our model), it learns only about which semantic and syntactic features to expect for each construction; it does not make more abstract judgments about how variable they are expected to be.

Is the ability to learn on this level important in order to explain the acquisition of verb constructions in natural language? Because many of the qualitative aspects of acquisition can be captured by models that cannot learn on that level, it remains possible that this sort of learning is not necessary. It may be that the multi-level learning observed in adults and children in the experiments of Wonnacott et al. (2008) and Wonnacott and Perfors (2009) is just a byproduct of some more general human ability to acquire overhypotheses and is not naturally used by people when learning verbs. Nevertheless, the results of these experiments suggest that humans are at least *capable* of this sort of learning. Given that these learning abilities are also useful in making more accurate generalizations about appropriate verb usage from realistic natural language input, as we have shown here, it would be surprising if people did not sometimes use them in natural language learning.

There may be conditions under which the ability to learn about variability is particularly useful. For instance, our model is capable of learning the distinction between alternating and non-alternating verb classes on the basis of syntactic input alone. The model of Alishahi and Stevenson (2008), which learns constructions rather than classes, forms constructions only on the basis of differences in features rather than on patterns of feature variability. As a result, it would be unable to form the distinction between non-alternating on alternating verb classes without additional semantic features to assist. Each individual verb usage would occur only in PD or DOD syntax, and without semantic information differentiating alternating from non-alternating verbs, the model would tend to infer a maximum of two constructions (DOD and PD). It would thus be able to learn that individual verbs are alternating or non-alternating (and therefore to resolve the negative evidence problem for those verbs). However, because it would have no reason to form a single alternating construction, it would not – as our model does – be able to infer that a completely novel verb appearing once in each construction is alternating, but that a verb appearing twice in one may not be. This is the sort of generalization made by adult subjects in the artificial language of Wonnacott et al. (2008). In future work we aim to address the empirical question of whether children or adults make such inferences in more naturalistic circumstances.

Current limitations and future directions

One aspect of the problem that few models address – ours included – is the question of how the child knows which sort of evidence is important. Pinker raised this point about indirect negative evidence, arguing that the problem of deciding which of the infinite number of sentences one hasn't heard are ungrammatical (rather than simply unattested) is "virtually a restatement of the original learning problem." (Pinker (1989), p. 14). How does the child know that those particular syntactic forms are the interesting and relevant ones? This knowledge has just been given to our model, and our work makes no particular claims about how it comes about. However, we have not simply restated the learning problem, as Pinker suggests: rather, we have suggested an answer to one problem (how to rule out logically possible alternatives without negative evidence), leaving another still unsolved (how to know which of a potentially infinite number of dimensions one should generalize along). The logic of Baker's Paradox would be the same whether there is one possible dimension of generalization, or an infinite number: the dilemma comes because one can never be certain that an unobserved datapoint (along that dimension) is truly ungrammatical, or simply unobserved. By converting this logical problem to a probabilistic one and demonstrating formally that the unobserved ones simply become increasingly unlikely, we have shown how a learner might be able to constrain their generalizations appropriately. How the learner knows which dimensions to pay attention to is a different issue.

Our model was originally developed and applied to a question in a different domain: the acquisition of the shape bias in word learning (Kemp et al., 2007). It incorporates relatively little domain-specific built-in knowledge – just the highest-level information governing the expectations about the sort of constructions, and their uniformity, that are likely to occur. Because the priors we built into the model were extremely weak, this amounted to the expectation that verbs within a class can either be uniform within a construction *or* precisely alternating *or* something in between. Thus, although any model (of any sort, not just a Bayesian model) must make some assumptions about what is built-in, we have endeavored to make them minimal. We can identify only two major assumptions:

1. Constructions have no internal representation: syntactic information is represented as a vector of features. We thus assume a learner who has already learned that, say,

NP_1 V NP_3 NP_2 is a construction. This is itself a complex learning problem and not one that this work bears upon.

2. The dataset presented to the model only includes (a) two constructions considered to be in alternation; and (b) the set of verbs that occur in at least one of these two constructions.

Regarding the first assumption, and as noted earlier, Baker's Paradox is a problem of knowing how to generalize appropriately over constructions in the input *once it is clear* what those constructions are. An advantage of our simpler representation is that it clarifies which phenomena emerge due to the nature of the data and the characteristics of Bayesian (optimal) inference, rather than because of the domain-specific representation. Furthermore, it allows us to explore the contribution of additional semantic features in the abstract, without worrying about their accessibility or precisely what they are. The tradeoff, of course, is that we are therefore abstracting over many details that might be critical for understanding the acquisition of particular verbs or asking the question of what precise semantic, conceptual, or syntactic knowledge must be built in in order for the child to perceive which features are relevant.

The second assumption mimics the input given to learners in the Wonnacott et al. (2008) experiments, but natural languages are more complex. A question for future research is whether our model could scale up to deal with a complex dataset involving all verbs and constructions. We think it possible that, due to its ability to learn about higher-level variability, our model could identify alternating verb classes even without prior information about which pairs of constructions potentially alternate. (Note that such learning could only work for the version of the model which can learn classes). Another possibility is that this learning relies on knowledge of construction semantics. For example, Goldberg (2006) has argued that repeatedly encountering a verb in one construction only pre-empts its usage in a non-occurring construction if that usage would satisfy the funcitonal demands of the context at least equally well. This suggest that only functionally related constructions may pre-empt each other and points to a potential constraint in the learning system, although which types of constructions are considered to be 'in competition' is still an open empirical question. Again, however, this is a question focused on how the learner knows which dimensions to pay attention to when using implicit negative evidence, rather than how implicit negative evidence constrains generalization; the latter is the question we address here.

Although Study 2 may appear to imply that some aspects of verb construction learning could be accomplished without semantic information, we are not suggesting that semantic knowledge is not important when learning verbs. Indeed, Study 3 was motivated by the fact that verb learning must include semantic as well as syntactic knowledge. Our work suggests the possibility that although semantic information is ultimately used, syntactic information may be more important initially – a suggestion that is consistent with the claims of syntactic bootstrapping (e.g., Gleitman, 1990), as well as a study by Ambridge et al. (2008), which found that for younger children (age 5-6) there was only a small effect of semantic class on generalization, whereas for older children and adults, there was a larger one. Similarly, Brooks et al. (1999) found that entrenchment effects in syntax emerged before semantic class effects.

It is also possible that syntactic information was so effective precisely because we gave our model clean features that already picked out precisely the constructions of interest. If both the semantic and syntactic features available to the child are far more noisy – and hidden amongst many irrelevant features in the environment – then it may be that semantic and syntactic features only become accessible through a process of mutual bootstrapping. Addressing this question, or others that explore the role of semantics more fully, will probably require a richer semantic representation than the current model instantiates. Future work we will explore this idea in more detail.

We would like to close by considering the question of convergence, which has been a central consideration for theories of acquisition. How do all children end up with the same grammar – one that accepts the same set of sentences as grammatical? In part, the focus on this question has arisen from the assumption of a deterministic end-state grammar that produces absolute yes-no grammaticality judgments for any sentence. However, there is evidence that judgments of overgeneralizations are variable even in adult native speakers, and continue to be influenced by lexical frequency (e.g., Theakston, 2004); as a result, many researchers have rejected this formulation of the end state. Despite this, our model does show high levels of convergence and systematically less generalization as it acquires more data, as do adults and children. Models such as ours also provide a method for exploring how different assumptions about learning and the data lead to different results. How our models link to the underlying psychological processes remains a fundamental question for further research.

Acknowledgements

We would like to thank Charles Kemp, Lila Gleitman, Brian MacWhinney, Nick Chater, and an anonymous reviewer for helpful comments. A version of this work made up a portion of the PhD thesis of the first author. This work was supported by an NDSEG graduate fellowship (AP), an NSF graduate fellowship (AP), the Paul E. Newton Career Development Chair (JBT) the James S. McDonnell Foundation Causal Learning Collaborative Initiative (JBT), and AFOSR grant FA9550-1-0075 (JBT).

Appendix 1: Models

The two models, referred to as L2 and L3, perform inference over different levels of abstraction. Both models learn at Level 1, the level of specific knowledge about individual verbs. Model L2 is specified in Kemp et al. (2007) and performs inference about the variability of Level 2 features α and $\boldsymbol{\beta}$ as well. Model L3 performs inference on an even high level, learning also about the expected range that that variability can take (denoted via parameters λ and μ).

Model L2: Learning at Level 2

Model L2, specified in Kemp et al. (2007), is known to statisticians as a Dirichlet-Multinomial model, and can be written as:

$$\begin{aligned} \alpha \quad & \sim \text{Exponential}(\lambda) \\ \boldsymbol{\beta} \quad & \sim \text{Dirichlet}(\boldsymbol{\mu}) \\ \boldsymbol{\theta^i} \quad & \sim \text{Dirichlet}(\alpha\boldsymbol{\beta}) \\ \boldsymbol{y^i} \,|\, n^i & \sim \text{Multinomial}(\boldsymbol{\theta^i}) \end{aligned}$$

where n^i is the number of observations for verb i. Because the model makes inferences on Level 2, we specify the Level 3 knowledge by setting parameters $\lambda = 1$ and $\boldsymbol{\mu} = \mathbf{1}$, which indicates weak prior knowledge that the expected range of α and $\boldsymbol{\beta}$ does not contain extreme values.

Inference is performed by computing posterior distributions over the unknown knowledge at the higher levels. For instance, the posterior distribution $P(\alpha, \boldsymbol{\beta}|\boldsymbol{y})$ represents a belief given the data $\boldsymbol{y}$ (the verbs heard so far). We formally instantiate this model by denoting the true distribution over constructions of verb i as $\boldsymbol{\theta^i}$; thus, if the true distribution of constructions means the verb occurs 70% of the time in the prepositional dative (PD) construction and 30% of the time in the direct object dative (DOD), then $\boldsymbol{\theta^i} = [0.3 \; 0.7]$. If we have observed that verb once in DOD and four in PD, then $\boldsymbol{y^i} = [1 \; 4]$. We assume that $\boldsymbol{y^i}$ is drawn from a multinomial distribution with parameter $\boldsymbol{\theta^i}$, which means that the observations of the verbs are drawn independently at random from the true distribution of verb i. The vectors $\boldsymbol{\theta^i}$ are drawn from a Dirichlet distribution parameterized by a scalar α and a vector $\boldsymbol{\beta}$: α determines the extent to which each verb tends to be associated with only one construction, and $\boldsymbol{\beta}$ represents the distribution of constructions across all verbs in the language.

To fit the model to data we assume that counts $\boldsymbol{y}$ are observed for each verb in the input set. Our goal is to compute the posterior distribution $P(\alpha, \boldsymbol{\beta}, \{\boldsymbol{\theta^i}\}|\boldsymbol{y})$. Inferences about α and $\boldsymbol{\beta}$ can be made by drawing a sample from $P(\alpha, \boldsymbol{\beta}|\boldsymbol{y})$ – the posterior distribution on $(\alpha, \boldsymbol{\beta})$ given the observed verbs. Inferences about $\boldsymbol{\theta^i}$, the distribution of constructions for verb i, can be made by integrating out α and $\boldsymbol{\beta}$:

$$P(\boldsymbol{\theta^i}|\boldsymbol{y}) = \int_{\alpha,\boldsymbol{\beta}} p(\boldsymbol{\theta^i}|\alpha, \boldsymbol{\beta}, \boldsymbol{y}) p(\alpha, \boldsymbol{\beta}|\boldsymbol{y}) d\alpha d\boldsymbol{\beta} \tag{1}$$

We estimate this using numerical integration via a Markov Chain Monte Carlo (MCMC) scheme. Our sampler uses Gaussian proposals on log(α), and proposals for $\boldsymbol{\beta}$ are drawn from a Dirichlet distribution with the current $\boldsymbol{\beta}$ as its mean.

Model L3: Learning at Level 2 and Level 3

Model L3 model is quite similar to Model L2, except that instead of assuming that λ and μ are known, we learn those as well. (nb: μ is a scalar, but in order for it to be a proper hyperparameter for the vector $\boldsymbol{\beta}$, it is vectorized: $\boldsymbol{\mu} = \mu\mathbf{1}$.) In statistical notation, this model can be written:

$$\begin{aligned}
\lambda &\sim \text{Exponential}(1) \\
\mu &\sim \text{Exponential}(1) \\
\alpha &\sim \text{Exponential}(\lambda) \\
\boldsymbol{\beta} &\sim \text{Dirichlet}(\boldsymbol{\mu}) \\
\boldsymbol{\theta^i} &\sim \text{Dirichlet}(\alpha\boldsymbol{\beta}) \\
\boldsymbol{y^i} \,|\, n^i &\sim \text{Multinomial}(\boldsymbol{\theta^i})
\end{aligned}$$

where n^i is the number of observations for verb i.

As before, inference is performed by computing posterior distributions over the unknown knowledge at the higher levels. The only difference is that the posterior distribution is now given by $P(\lambda, \mu, \alpha, \boldsymbol{\beta}, \{\boldsymbol{\theta^i}\}|\boldsymbol{y})$. Inferences about λ, μ, α, and $\boldsymbol{\beta}$ can be made by drawing a sample from $P(\alpha, \boldsymbol{\beta}, \lambda, \mu|\boldsymbol{y})$, which is given by:

$$P(\alpha, \boldsymbol{\beta}, \lambda, \mu|\boldsymbol{y}) \propto P(\boldsymbol{y}|\alpha, \boldsymbol{\beta})P(\alpha|\lambda)P(\boldsymbol{\beta}|\mu)P(\lambda)P(\mu) \tag{2}$$

Inferences about $\boldsymbol{\theta^i}$, the distribution of constructions for verb i, can be made by integrating out α, $\boldsymbol{\beta}$, λ, and μ:

$$P(\boldsymbol{\theta^i}|\boldsymbol{y}) = \int_{\alpha,\boldsymbol{\beta},\lambda,\mu} P(\boldsymbol{\theta^i}|\alpha, \boldsymbol{\beta}, \boldsymbol{y})P(\alpha, \boldsymbol{\beta}, \lambda, \mu|\boldsymbol{y})d\alpha d\boldsymbol{\beta} d\lambda d\mu \tag{3}$$

Again, we estimate this using numerical integration via a Markov Chain Monte Carlo (MCMC) scheme. As before, our sampler uses Gaussian proposals on $\log(\alpha)$, $\log(\lambda)$, and $\log(\mu)$, and proposals for $\boldsymbol{\beta}$ are drawn from a Dirichlet distribution with the current $\boldsymbol{\beta}$ as its mean.

Model extension: Learning verb classes

To add the ability to discover verb classes to both Models L2 and L3, we assume that verbs may be grouped into classes, each of which is associated with its own hyperparameters. For Model L2, this means that there is a separate α^c and $\boldsymbol{\beta^c}$ for each class c inferred by the model; for Model L3, there is a separate α^c, $\boldsymbol{\beta^c}$, λ^c, and μ^c for each class c. In both cases, the model partitions the verbs into one or more classes, where each possible partition is represented by a vector $\boldsymbol{z}$. Thus, a partition of six verbs in which the first three verbs are in one class and the last three were in another can be represented by the vector [1 1 1 2 2 2]. The prior distribution on $\boldsymbol{z}$ is induced by the Chinese Restaurant Process:

$$P(z_i = c|z_1, \ldots, z_{i-1}) = \begin{cases} \frac{n_c}{i-1+\gamma} & n_c > 0 \\ \frac{\gamma}{i-1+\gamma} & c \text{ is a new class} \end{cases} \tag{4}$$

where z_i is the class assignment for verb i, n_c is the number of verbs previously assigned to class c, and γ is a hyperparameter which captures the degree to which the process favors simpler class assignments (we set $\gamma = 1$). The Chinese Restaurant Process prefers to assign verbs to classes that already have many members, and therefore tends to prefer partitions with fewer classes.

The extension for Model L3 can now be written as follows:

$$\begin{aligned}
\boldsymbol{z} &\sim \text{CRP}(\gamma)\\
\lambda^c &\sim \text{Exponential}(1)\\
\mu^c &\sim \text{Exponential}(1)\\
\alpha^c &\sim \text{Exponential}(\lambda^c)\\
\boldsymbol{\beta^c} &\sim \text{Dirichlet}(\boldsymbol{\mu^c})\\
\boldsymbol{\theta^i} &\sim \text{Dirichlet}(\alpha^{c_i}\boldsymbol{\beta^{c_i}})\\
\boldsymbol{y^i} \,|\, n^i &\sim \text{Multinomial}(\boldsymbol{\theta^i})
\end{aligned}$$

The equivalent extension for Model L2 is trivially extendable from this, and is also described in Kemp et al. (2007).

If $\boldsymbol{z}$ is known, the extended model reduces to several independent versions of the basic (L2 or L3) model, and predictions can be computed using the techniques described earlier. Since $\boldsymbol{z}$ is unknown, we must integrate over each of the possible class partitions $\boldsymbol{z}$:

$$P(\boldsymbol{\theta^i}|\boldsymbol{y}) = \sum_{\boldsymbol{z}} P(\boldsymbol{\theta^i}|\boldsymbol{y},\boldsymbol{z})P(\boldsymbol{z}|\boldsymbol{y}) \tag{5}$$

where $P(\boldsymbol{z}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{z})P(\boldsymbol{z})$ and $P(\boldsymbol{z})$ is the prior induced by the CRP process. For Model L2, computing $P(\boldsymbol{y}|\boldsymbol{z})$ reduces to the problem of computing several marginal likelihoods

$$P(\boldsymbol{y\prime}) = \int_{\alpha,\boldsymbol{\beta}} P(\boldsymbol{y\prime}|\alpha,\boldsymbol{\beta})P(\alpha,\boldsymbol{\beta})d\alpha d\boldsymbol{\beta} \tag{6}$$

which we estimate by drawing 10,000 samples from the prior $P(\alpha,\boldsymbol{\beta})$.

For model L3, computing $P(\boldsymbol{y}|\boldsymbol{z})$ reduces to computing

$$P(\boldsymbol{y\prime}) = \int_{\alpha,\boldsymbol{\beta},\lambda,\mu} P(\boldsymbol{y\prime}|\alpha,\boldsymbol{\beta})P(\alpha,\boldsymbol{\beta}|\lambda,\mu)P(\lambda,\mu)d\alpha d\boldsymbol{\beta} d\lambda d\mu \tag{7}$$

which is also estimated by drawing 10,000 samples, this time from the joint prior $P(\alpha,\boldsymbol{\beta}|\lambda,\mu)P(\lambda,\mu)$.

Appendix 2: Corpus

The data of verb counts is collected from the sentences spoken by adults in the Adam corpus (Brown, 1973) of the CHILDES database (MacWhinney, 2000), and consists of all instances of each of the dative verbs listed in Levin (1993), including the number of occurrences in each construction (PD and DOD). EPOCHS correspond to the counts for verbs in subsections of the corpus of 55 files, split by age (EPOCH 1 is the first 11 files, EPOCH 2 is the first 22, and so on). The counts are shown in Table 3.

Table 3: Number of times each verb appears in each construction (Adam corpus).

	Epoch 0		Epoch 1		Epoch 2		Epoch 3		Epoch 4		Full corpus	
Verb	DOD	PD	DOD	PD	DOD	PD	DOD	PD	DOD	PD	DOD	PD
take			0	5	0	9	0	11	0	16	0	16
say			0	3	0	4	0	6	0	6	0	6
explain					0	1	0	1	0	1	0	1
send							0	1	0	1	0	2
sell							0	1	0	1	0	1
mail							0	1	0	1	0	1
throw			1	0	1	2	1	2	1	2	1	2
read	1	1	2	5	2	11	3	12	3	13	3	16
give	2	1	15	18	39	27	62	31	82	33	106	44
show	2	1	10	5	23	9	27	11	31	15	36	17
bring			2	1	4	3	6	3	9	4	11	5
tell			1	1	8	1	14	1	17	1	22	1
sing					1	0	1	0	1	0	1	0
pay									2	0	2	0
serve							2	0	2	0	2	0
find											2	0
ask					2	0	3	0	3	0	4	0
make					1	0	5	0	6	0	11	0

References

Alishahi, A., & Stevenson, S. (2008). A computational model for early argument structure acquisition. *Cognitive Science*, *32*.

Allen, J., & Seidenberg, M. (1999). The emergence of grammaticality in connectionist networks. In B. MacWhinney (Ed.), *Emergence of language.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Ambridge, B., Pine, J., Rowland, C., & Young, C. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition*, *106*, 87–129.

Baker, C. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, *10*, 533–581.

Berwick, R. (1985). *The acquisition of syntactic knowledge.* Cambridge, MA: MIT Press.

Bowerman, M. (1988). The no negative evidence problem: How do children avoid constructing an overly general grammar? In J. Hawkins (Ed.), *Explaining language universals.* Oxford: Basil Blackwell.

Braine, M. (1971). On two types of models of the internalization of grammars. In D. Slobin (Ed.), *The ontogenesis of grammar: A theoretical symposium.* New York, NY: Academic Press.

Braine, M., & Brooks, P. (1995). Verb argument structure and the problem of avoiding an overgeneral grammar. In M. Tomasello & W. Merriman (Eds.), *Beyond names of things: Young children's acquisition of verbs* (pp. 353–376). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Brooks, P., & Tomasello, M. (1999). How children constrain their argument structure constructions. *Language*, *75*, 720–781.

Brooks, P., Tomasello, M., Dodson, K., & Lewis, L. (1999). Young children's overgeneralizations with fixed transitivity verbs. *Child Development*, *70*, 1325–1337.

Brown, R. (1973). *A first language: The early stages.* Harvard University Press.

Chater, N., & Vitànyi, P. (2007). 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, *51*(3), 135–163.

Chouinard, M., & Clark, E. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, *30*, 637-669.

Dominey, P. (2003). Learning grammatical constructions in a miniature language from narrated video events. *Proceedings of the 25th Annual Conference of the Cognitive Science Society.*

Dowman, M. (2000). Addressing the learnability of verb subcategorizations with Bayesian inference. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society.*

Fisher, C., Gleitman, H., & Gleitman, L. (1991). On the semantic content of subcategorization frames. *Cognitive Psychology*, *23*, 331–392.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1991). Human simulations of vocabulary learning. *Cognition*, *73*, 153–176.

Gleitman, L. (1990). The structural sources of word learning. *Language Acquisition*, *1*, 3–55.

Goldberg, A. (1995). *A construction grammar approach to argument structure.* Chicago, IL: University of Chicago Press.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language.* Oxford: Oxford University Press.

Goodman, N. (1955). *Fact, fiction, and forecast.* Cambridge, MA: Harvard University Press.

Gordon, P. (1990). Learnability and feedback. *Developmental Psychology*, *26*(2), 217–220.

Gropen, J., Pinker, S., Hollander, M., & Goldberg, R. (1991). Syntax and semantics in the acquisition of locative verbs. *Journal of Child Language*, *18*(1), 115–151.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R., & Wilson, R. (1989). The learnability and acquisition of the dative alternation in English. *Language*, *65.*

Hsu, A. (2009). Differential use of implicit negative evidence in generative and discriminative language learning. *Advances in Neural Information Processing Systems*, *22.*

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation.* University of Chicago Press.

MacWhinney, B. (1987). The competition model. In B. MacWhinney (Ed.), *Mechanisms of language acquisition.* Hillsdale, NJ: Erlbaum.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (Third ed.). Lawrence Erlbaum Associates.

MacWhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. *Journal of Child Language*, *31*, 883–914.

Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2005). The role of frequency in the acquisition of the English word. *Cognitive Development*, *201*, 121–136.

Mazurkewich, I., & White, L. (1984). The acquisition of the dative alternation: Unlearning overgeneralizations. *Cognition*, *16*, 261–283.

Mitchell, D. (1987). Lexical guidance in human parsing: Locus and processing characteristics. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 601–681). Hillsdale, NJ: Erlbaum.

Morgan, J., & Demuth, K. (1996). *Signal to syntax: Bootstrapping from speech to grammar in early acquistition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Onnis, L., Roberts, M., & Chater, N. (2002). Simplicity: A cure for overregularizations in language acquisition? *Proceedings of the 24th Annual Conference of the Cognitive Science Society.*

Pinker, S. (1984). *Language learnability and language development.* Cambridge, MA: Harvard University Press.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure.* Cambridge, MA: MIT Press.

Snedeker, J., & Trueswell, J. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, *49*, 238–299.

Theakston, A. (2004). The role of entrenchment in childrens and adults performance limitations on grammaticality judgment tasks. *Cognitive Development*, *19*, 15–34.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Harvard University Press.

Trueswell, J., Tanenhaus, M., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 528–553.

Wonnacott, E., Newport, E., & Tanenhaus, M. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*, 165–209.

Wonnacott, E., & Perfors, A. (2009). Constraining generalisation in artificial language learning: Children are rational too. *Poster presented at 22nd Annual CUNY conference on human sentence processing.*

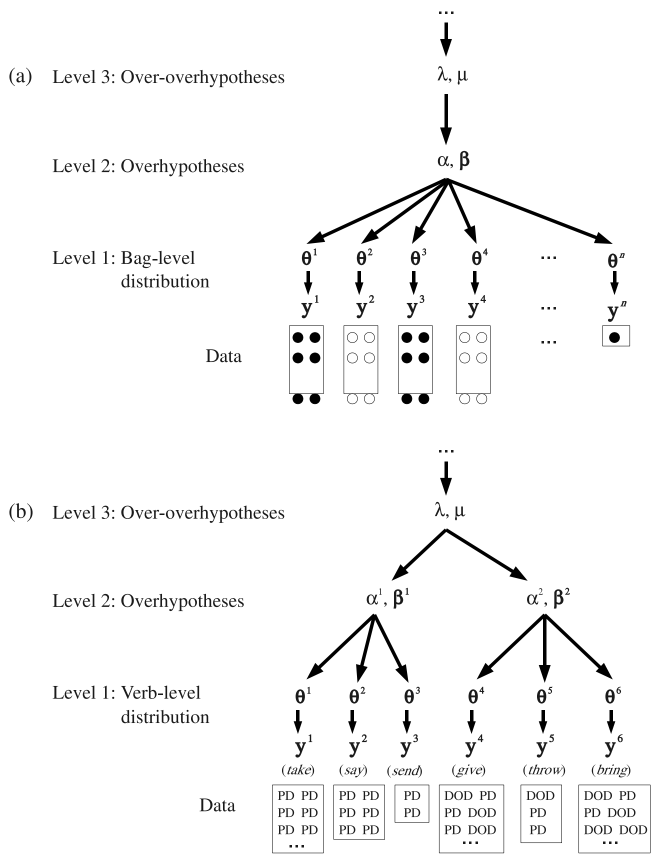


Figure 1. (a) A hierarchical Bayesian model (HBM). Each setting of $(\alpha, \boldsymbol{\beta})$ is an overhypothesis: $\boldsymbol{\beta}$ represents the color distribution of marbles across all bags (or, equivalently, the distribution of constructions across all verbs in a language), and α represents the variability/uniformity of colored marbles within each bag (or, equivalently, the degree to which each verb tends to be alternating or non-alternating). In principle HBMs can be extended to arbitrarily high levels of increasingly abstract knowledge and need not 'ground out' at Level 3. (b) A model with separate overhypotheses for two verb classes, loosely corresponding to a non-alternating class of verbs that occur exlusively in the PD construction and an alternating class of verbs that occur in both PD and DOD constructions. α^1 represents knowledge about the uniformity of constructions within the non-alternating class (i.e., that it is non-alternating) and $\boldsymbol{\beta}^1$ captures the characteristic constructions of the verbs in that class (i.e., that they occur in the PD construction). This figure is adapted from Figure 1 in Kemp et al. (2007).

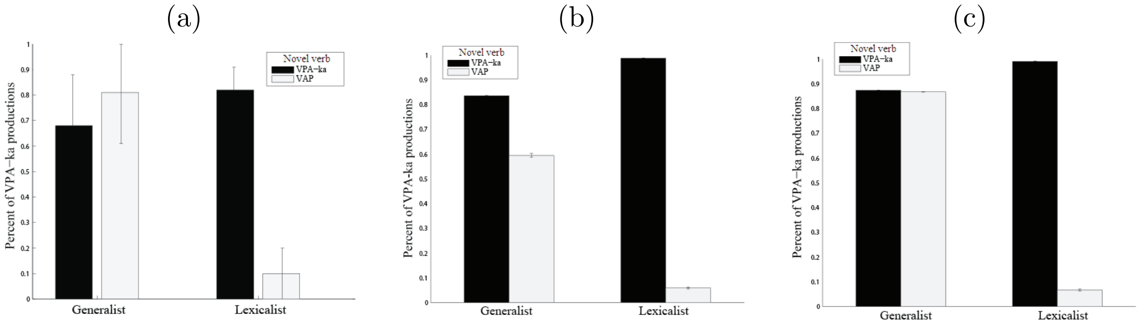
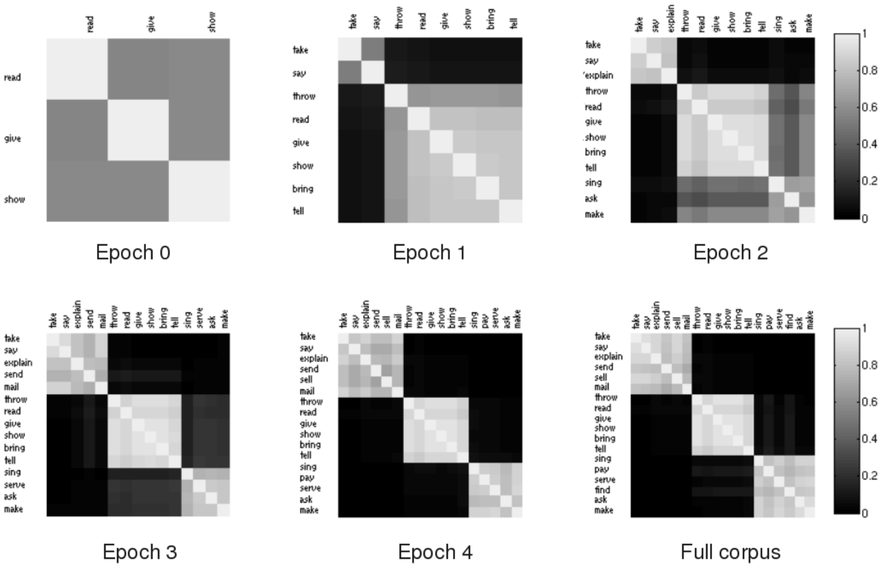


Figure 2. Comparison of model performance with human production for novel verbs in an artificial language. (a) Adult performance. Subjects in the Generalist condition were likely to produce a novel verb in both constructions, matching the overall frequency of each in the language, rather than the single construction it was heard in. Subjects in the Lexicalist condition, whose previous input consisted of verbs that occurred in only one construction at a time, tended to produce the novel verb only in the single construction it occurred in. (b) Model L2. (c) Model L3. Both models qualitatively replicate the difference in human performance in the each condition. Model L3, which can learn at a higher level of abstraction, matches human performance more closely.



Epoch 0

Epoch 1

Epoch 2

Epoch 3

Epoch 4

Full corpus

Figure 3. Class assignments given by Model K-L3. Lighter colors indicate increasing probability that the verbs in that row and column are assigned to the same class. The diagonal is always white because each verb is always in the same class as itself.

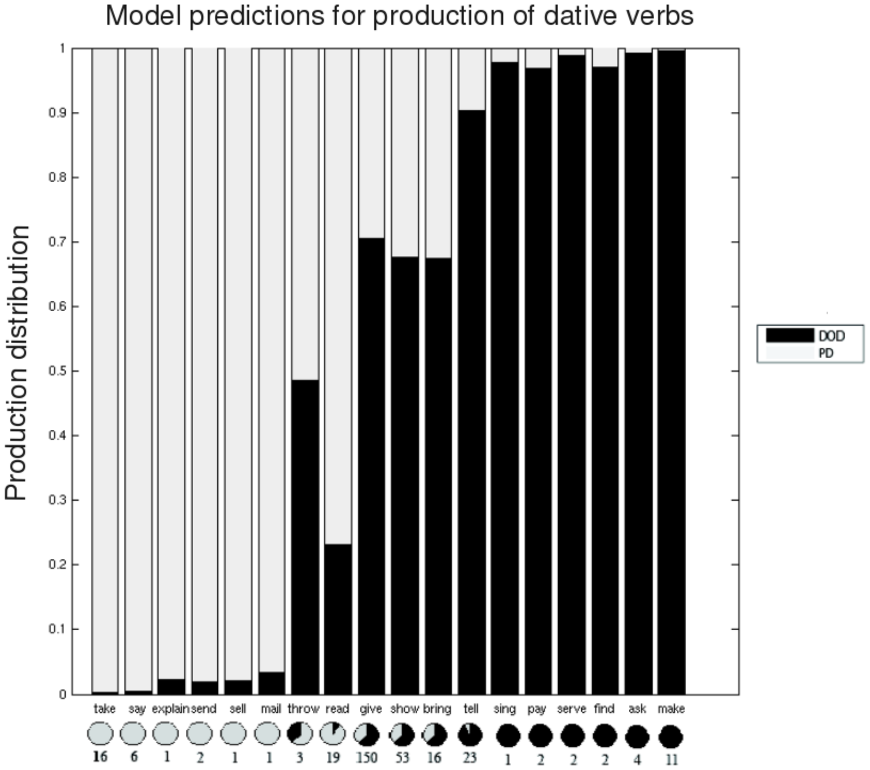


Figure 4. Production predictions of Model K-L3 for each verb in the full corpus. High-frequency verbs' constructions are produced at a distribution close to their empirical distribution, while low-frequency verbs are more likely to be overgeneralized (produced in a construction that they did not occur in in the input). The production distribution is denoted with the stacked bars; the associated pie chart depicts each verb's observed distribution, and its empirical frequency is the number under the pie chart.

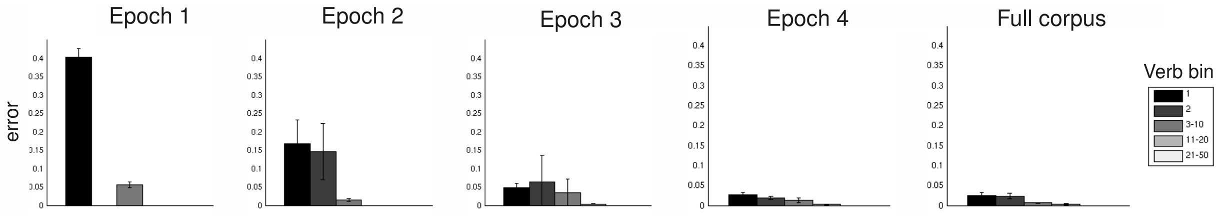


Figure 5. Degree of overgeneralization of non-alternating verbs by Epoch for Model K-L3. The y axis reflects the degree of overgeneralization, calculated by taking the absolute value of the difference between the proportion of the time the verb is observed vs. produced by the model in the DOD construction. Verbs of different frequencies are grouped in bins along the x axis: thus bin 1 contains all of the verbs that occurred only once in the corpus, bin 2 contains verbs occurring twice, and so on.

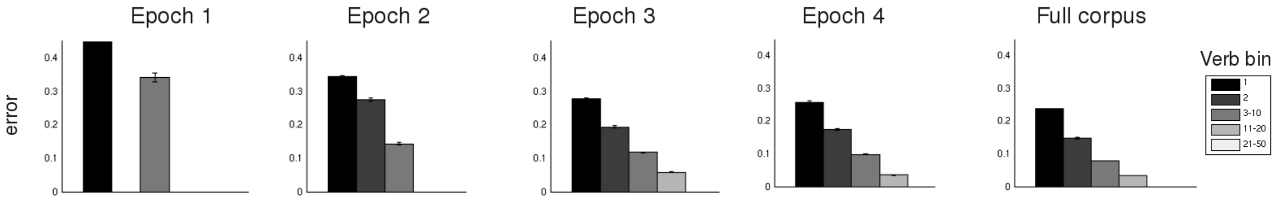


Figure 6. Degree of overgeneralization of non-alternating verbs by Epoch for Model L3.

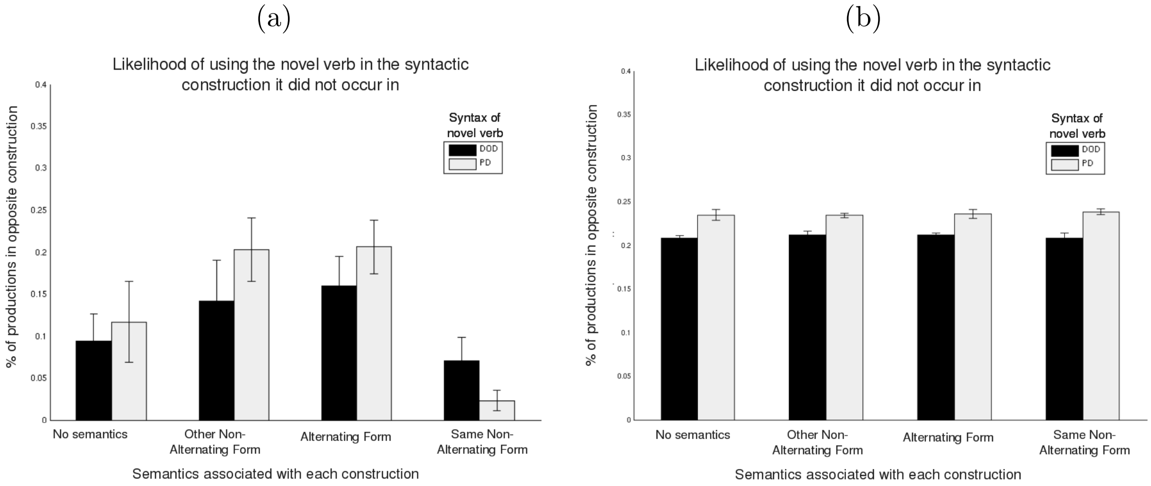


Figure 7. Percent of generalization of novel verb to the syntactic construction it was not previously observed in. (a) Model K-L3; (b) Model L3. Model K-L3 behaves qualitatively like human responses, generalizing most to the other construction when the semantic feature is consistent with the non-alternating class, and least when it is consistent with the same alternating class. Model L3 does not, suggesting that the ability to group verbs into appropriate classes might be necessary to capture this aspect of human behavior.

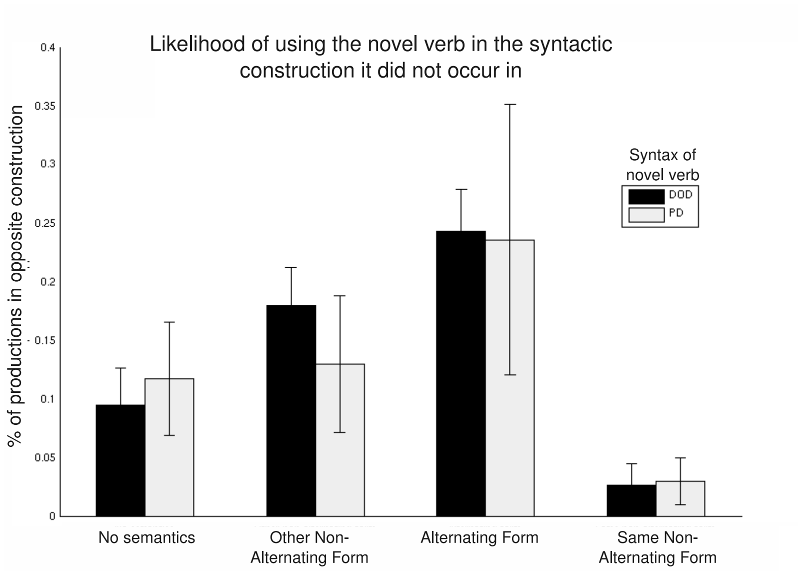


Figure 8. Percent of generalization of novel verb to the syntactic construction it was not previously observed in. As before, the model behaves qualitatively like human responses, generalizing most to the other construction when the semantic features are roughly consistent with the non-alternating class, and least when it they are more consistent with the same alternating class.

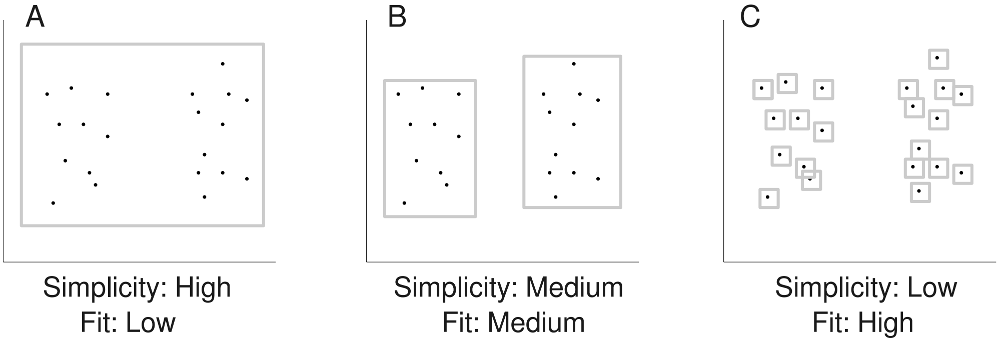


Figure 9. Hypothesis *A* is too simple, since it fits the observed data poorly; *C* fits closely but is too complex; and *B* is 'just right.' The best description of the data should optimize a tradeoff between complexity and fit, as in *B*.