Children's Inferences in Generalizing Novel Nouns and Adjectives

Annie Gagliardi (acg39@umd.edu)
Erin Bennett (ebennet2@umd.edu)
Jeffrey Lidz (jlidz@umd.edu)
Naomi H. Feldman (nhf@umd.edu)
Department of Linguistics, University of Maryland, College Park, MD 20742 USA

Abstract

By the time children begin to rapidly acquire new word meanings they are already able to determine the grammatical category of novel words based on syntactic and morphological cues. Here we test whether children can leverage this knowledge when inferring the meaning of a novel word. Through a novel word learning experiment we determine that children can use this information, drawing different conclusions for the most likely meanings of novel words in distinct grammatical categories. We use a Bayesian model to formalize the higher level knowledge that children might have about noun and adjective meanings. Simulations show that children's behavior closely matches what we would predict on the basis of noun and adjective meanings in the English lexicon.

Keywords: language acquisition; word learning; Bayesian inference

Introduction

One of the most striking phenomena in language acquisition is children's ability to rapidly learn the meanings of novel words with only limited exposure. How exactly children do this has been researched extensively, with three lines of inquiry dominating the attempts to formalize this process: *hypothesis elimination* (Berwick, 1963; Pinker, 1989; Siskind, 1996), *associative learning* (Colunga & Smith, 2005; Regier, 2005) and *Bayesian inference* (Xu & Tenenbaum, 2007). Xu and Tenenbaum argue that Bayesian inference is superior to hypothesis elimination and associative learning because it uniquely allows the learner to take advantage of 'suspicious coincidences' when learning words for overlapping concepts. For example, in a word learning experiment they found that when children were shown three Dalmatians labeled with a novel object label, there was a strong bias for children to think that the novel word meant Dalmatian, rather than dog, or animal. This bias was not as strong when children only saw one Dalmatian labeled with the novel label. Neither hypothesis elimination nor associative learning predict the effect of the suspicious coincidence that results from the narrow distribution of exemplars on the kind hierarchy (which is in turn contingent on the number of exemplars). Xu and Tenenbaum's model does predict this effect, via the likelihood term, which takes into account both the number of exemplars and the size of the hypothesis.

One key assumption that Xu and Tenenbaum relied on was that the candidate concepts fell on a hierarchy of kinds. That is, in their model the learner does not have to determine what domain to generalize across, as this domain was given by the kind hierarchy. This assumption has two implications for their model: (1) most of the work in hypothesis selection is being done by the likelihood, as the prior probability of each hypothesis is comparatively much less variable and (2) it largely limits the model to the discussion of object label learning, as this is the domain that primarily uses the kind hierarchy.

In this paper we probe the predictions of the Bayesian model on different grammatical categories, nouns and adjectives, which tend to draw from different concept hierarchies. This allows us to better test the role of the prior probability of a concept given a grammatical category by letting us examine the link between grammatical category and concept hierarchy. Toward these goals we conducted a word learning experiment that replicates Xu and Tenenbaum's finding with learning novel nouns, and extends the paradigm to novel adjective learning. We find that children use the grammatical category of the novel word to constrain their hypotheses about the meaning of the novel word. This is demonstrated through their sensitivity to the suspicious coincidence in the distribution of exemplars on the kind hierarchy when learning nouns but not adjectives. A Bayesian model that takes into account not only conceptual similarity but also the link between grammatical category and concept provides a close fit to the children's data. Through this work we extend the Bayesian model of word learning in ways that make it more realistic with respect to both the structure of natural language and the task faced by a child acquiring novel words.

Our paper is organized as follows. We first present our word learning experiment. We then use a Bayesian model to formalize children's prior distribution over concepts. The next section presents simulations comparing the model to children's behavior. We conclude by discussing the implications that this work has for language acquisition, in particular the importance of considering how a learner's prior knowledge affect the way in which the data from the environment are used in language acquisition.

Word Learning Experiment

In a novel word learning experiment children were presented with an array of animals and vehicles and taught a novel label (noun or adjective) for a concept. Children were then asked to generalize their inferred concept to novel items. The stimuli allowed generalization along both kind and property dimensions. If children are able to use syntactic information to constrain their inference of words' meanings, then we should expect them to generalize differently when learning nouns versus adjectives.

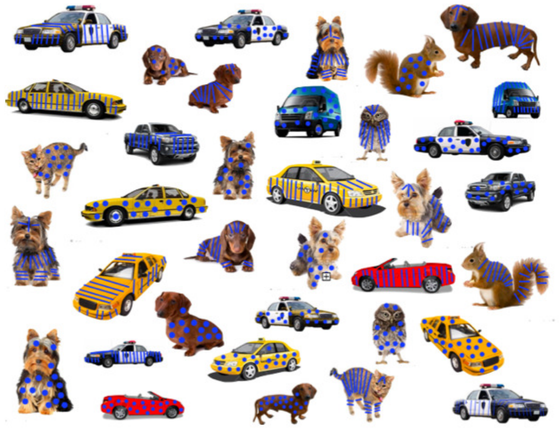

Figure 1: The stimuli for our experiment included 36 objects in subordinate, basic, and superordinate vehicle and animal categories. Half the items were striped and half spotted.

Methods

Our experiment tested two groups of children using a between subjects design. The noun group learned two novel nouns, and the adjective group learned two novel adjectives.

Participants Participants were 24 children (mean age = 4;0, range = 3;6-5;0) recruited from the greater College Park area as well as an on campus preschool. Children either visited the lab with their parents or were visited by researchers at their preschool. Four children were excluded from the final analysis for the following reasons. One was too shy to interact with the snail and three said they didn't know when they were asked to perform the generalization task outlined below.

Stimuli All children were presented with an array of pictures (Figure 1) that included 36 items from two superordinate categories on the kind hierarchy (18 vehicles and 18 animals). Each category had items from several basic levels (animals: 12 dogs, 2 cats, 2 squirrels, 2 owls; vehicles: 12 roofed cars, 2 convertibles, 2 vans, 2 trucks). One basic level from each superordinate category had items from two subordinate level categories (dogs: 6 Dachshunds and 6 Yorkshire terriers, roofed cars: 6 taxis and 6 police cars). There were both striped and spotted items of each item type.

Procedure A snail puppet was introduced to the child, and the child was told that the snail spoke a funny snail language that was mostly like English but included some new words. The experimenter explained to the child that they would try to figure out the snail's words by listening to him talk about some of the pictures. Before proceeding further, the experimenter checked that both the snail and the child could see all of the pictures in the array. This ensured that participants were aware of the range of items in the experimental world.

During the **word learning phase** the snail looked at the pictures and pointed out an item from one of the subordinate level categories (e.g. a striped dachshund). In the noun condition he described it as *a blick*, and in the adjective condition he described it as *a blicky one*. This happened 3 times, with the snail pointing to a different striped dachshund each time. Then the snail would get tired and retire to his shell for a nap.

Speaker	**Utterance**	**Action**
Snail	'This is a *blicky one*'	points to striped Dachshund 1
Snail	'Look, another *blicky one*'	points to striped Dachshund 2
Snail	'Here's another *blicky one*'	points to striped Dachshund 3
Snail	'I'm going to go have a rest in my shell'	retreats to shell
Experimenter	'Here are some more pictures, can you put circles on all the *blicky ones* to surprise the snail when he wakes up?'	lays out new array of pictures and gives the child a set of rings
Child	—	puts rings on items that match child's hypothesis for the meaning of *blicky*

Table 1: Sample adjective trial - Noun trials are identical with *blick* substituted for *blicky one*

While the snail slept, the experimenter initiated the **test phase**, during which the child was presented with another array of pictures and asked to place circles on the other *blicks* (noun condition) or *blicky ones* (adjective condition). A single trial is schematized in Table 1. The entire procedure was repeated for a second novel word used to describe another item from a different subordinate level (e.g. a spotted taxi). Order of item (dog before vehicle or vice versa), described subordinate level item (dachshund vs yorkie and taxi vs police car), and described pattern order (striped before spotted and vice versa) were all counterbalanced across subjects.

Results

Children's choices were coded as follows, with one response recorded per trial. Subordinate responses were recorded if children chose only animals/vehicles from the same subordinate level as the example (e.g. only more dachshunds after being presented with dachshunds). Basic responses were recorded if children chose from only the basic level (i.e. either dog type after being presented with dachshunds) or from the basic and subordinate levels. A superordinate response was recorded if children chose only from the superordinate level (e.g. any animal after being presented with dachshunds) or from the superordinate level with any combination of the lower levels. Finally, neutral responses were recorded if children chose from anywhere on the kind hierarchy (e.g. chose anything from the vehicle hierarchy after being shown a dachshund).

Results are shown in Figure 2(a). In the noun condition, we replicated Xu & Tenenbaum's finding, uncovering a bias for the subordinate level meaning when all observations fall into the same subordinate level. In the adjective condition however, we see a different pattern. The placement of the item on the kind hierarchy had no bearing on children's choices, with the overwhelming majority choosing the neutral interpretation, indicating their belief that the novel adjective's meaning referred just to the most salient property (striped versus spotted) rather than the kind. Planned comparisons revealed that the proportion of trials that children chose the subordinate and neutral meanings differed significantly by condition (subordinate: $t(33) = 3.49, p < 0.002$, neutral: $t(26) = 3.39, p < 0.003$).

Discussion

These results demonstrate that children use their knowledge of grammatical categories, and the associated kinds of meanings that correlate with these categories, when inferring the meanings of novel words. In particular, they favor concepts from a kind hierarchy for novel nouns, and from a property hierarchy for novel adjectives. In one respect this result is not new, as infants as young as 14 months have been shown to know the mapping between grammatical and conceptual categories (Waxman & Markow, 1998; Booth & Waxman, 2003, 2009). Instead, the novelty is in showing that this mapping constrains children's inferences. A very low prior probability for a hypothesis on the kind hierarchy blocks it from being determined the most likely for a novel adjective meaning, despite it being the narrowest possible hypothesis.

This finding emphasizes the role of the hypothesis space, as the most likely hypothesis differs depending on the grammatical category of the word being learned. In order to determine whether children are behaving optimally with respect to a specific hypothesis space (conditioned by grammatical category and the information available to them in the English lexicon), we used a Bayesian model to predict generalization behavior from the nouns and adjectives that are likely to be present in the children's early lexicons.

Model

We assume the generative model shown in Figure 3. Our model assumes that the snail in our experiment, having chosen a syntactic category for the word he will teach the children, chooses a concept to teach (such as *dog*, *striped*, or *dachshund*), and then independently chooses three objects as examples of that concept.

The children in our experiment inferred what concept a new word referred to based on the lexical category of the novel word (noun or adjective) and the objects the snail identified as examples of that word. Our model therefore computes the probability of each concept C for a given syntactic category P and set of objects X,

$$\mathbb{P}(C|X,P) \tag{1}$$

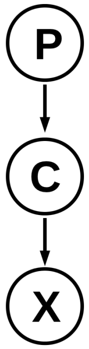

Figure 3: Syntactic categories P determine the parameters for our prior over concepts C. Specific objects X are sampled from the set of items that exemplify a concept.

Concept	$\rightarrow$	Kind
	$\rightarrow$	Property
	$\rightarrow$	Kind $\wedge$ Property
Kind	$\rightarrow$	animal
	$\rightarrow$	dog
	$\rightarrow$	dachshund
	$\rightarrow$	$\cdots$
	$\rightarrow$	vehicle
	$\rightarrow$	car
	$\rightarrow$	taxi
Property	$\rightarrow$	spotted
	$\rightarrow$	striped

Figure 4: A probabilistic context-free grammar for concepts. Probabilities for each expansion rule are discussed in the Concept Prior section.

We can use Bayes' rule to compute the posterior probability over concepts given a set of examples and a word's syntactic category,

$$\mathbb{P}(C_i|X,P) = \frac{\mathbb{P}(X|C_i)\cdot\mathbb{P}(C_i|P)}{\sum\limits_{C_j\in\{\text{all concepts}\}}\mathbb{P}(X|C_j)\cdot\mathbb{P}(C_j|P)} \tag{2}$$

We assume that the probability of the data X depends only on the concept C and is independent of the syntactic category, given the concept. Since the normalizing constant in the denominator will be the same for all candidate concepts, we only need to find the values of $\mathbb{P}(X|C_i)$ and $\mathbb{P}(C_i|P)$ for the concepts we are considering.

Concept Prior: $\mathbb{P}(C|P)$

Following Goodman, Tenenbaum, Feldman and Griffiths (2008) (cf. Austerweil & Griffiths, 2010), we represent con-

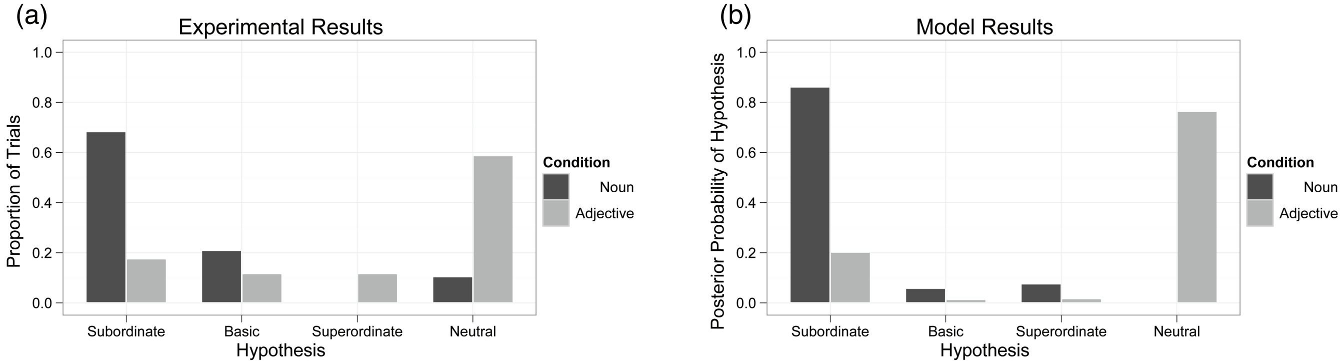

Figure 2: (a) Results of word learning experiment and (b) results of modeling

cepts according to the concept grammar in Figure 4, with nonterminal nodes *Kind* and *Property* representing the dimensions a concept is defined along. Words like *dog* and *striped* are defined along only one of these dimensions (*Kind* and *Property*, respectively). Words like *kitten*, which must describe a young cat, are defined along both dimensions ($Kind \wedge Property$). The derivation of each concept involves first applying a rule determining the dimension of the concept and then applying the dimension-specific rules until all terminal nodes have been identified. For example, in our concept language, the concept *dog* is formed by first applying the rule $Concept \rightarrow Kind$ and then applying the rule $Kind \rightarrow dog$.

If we assign probabilities to each of the rules in this concept grammar and assume that the rules are applied independently of one another, then the resulting PCFG will determine the probabilities of all the concepts in our experiment. The probability of each concept would be the product of the probabilities of the rules applied to form it,

$$\mathbb{P}(C) = \prod_{R \in \{\text{rules to form C}\}} \mathbb{P}(R) \tag{3}$$

The differences in the types of concepts represented by nouns and adjectives are represented in our model through differences in the probability distributions over the set of rules that expand *Concept* to particular dimensions. We assume children are computing this prior distribution separately for each part of speech, keeping track of the number of nouns or adjectives whose meanings denote a kind, a property, or both a kind and a property. They can estimate the rule probabilities from these counts using a Dirichlet-multinomial model. Under this model, the prior over dimension expansions based on the counts $p_{d_i,P}$ of the productions seen by the learner of a particular dimension d_i for that lexical category P is:

$$\mathbb{P}(d_i|P) = \frac{p_{d_i,P}+1}{\sum_{d_j \in \{\text{all dims}\}} p_{d_j,P}+3} \tag{4}$$

We approximated these production counts from a Mechanical Turk survey where for each word in a vocabulary list of

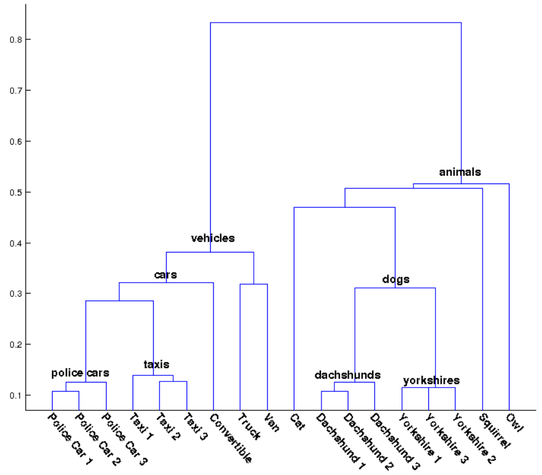

Figure 5: Hierarchical Clustering of Experimental Item Similarity

430 words (364 nouns and 66 adjectives) that 30-month-old children likely know (Dale & Fenson, 1996) we asked adult English speaking participants to judge whether the word was best described as a kind, a property, or both. Different but often overlapping sets of 10 people were asked to respond to each word, and so we had a total of 22 participants in our study. Two participants' judgments were excluded due to an extraordinarily high proportion of *Both* responses (proportion Both$>$ 0.36, over two standard deviations outside the mean proportion of *Both* responses). While the children in our experiment (3-5 year-olds) were much older than 30 months, we believe that the 30-month-old children's vocabulary list is appropriate for our purposes, since the children in our experiments are almost certainly familiar with these words and differ only in additional words they might know. We assume that the distribution of noun and adjective dimensions in this set of words is representative of that of the larger and more varied set of words that our 3-5 year-old participants are familiar with,

For kinds, we assume a structure like Xu and Tenenbaum

(2007) where the probability of a concept depends on its distinctiveness. For these measures we use a hierarchical cluster tree as in Figure 5. To make this tree, we conducted a similarity judgment study, similar to Xu and Tenenbaum's using the items that the snail had labeled in our experiment. Our participants, 26 students from the University of Maryland who received course credit for their participation, rated the similarity of all possible pairs of the 36 pictures on a scale from 1 (not similar at all) to 9 (very similar).

To incorporate cluster distinctiveness, Xu and Tenenbaum measure the branch length (which represents the Euclidean distance) between the concept node and its parent node. By this measure, the further a particular node is from its parent, the more distinct it is considered to be. Where $\mathcal{K}$ is the set of all *Kind* concepts, the probability of a concept C_i given that it is defined over the *Kind* dimension is the branch length normed over all *Kind* concepts,

$$\mathbb{P}(C_i|\text{Kind}) = \frac{height(parent(C_i)) - height(C_i)}{\sum\limits_{C_j \in \mathcal{K}} height(parent(C_j) - height(C_j)} \tag{5}$$

For properties, we assume that in our experiment they are chosen from a Multinomial Distribution with each property equally likely to be selected. Since there were only two very salient properties in our experiment, we give each property the probability of $\frac{1}{2}$,

$$\mathbb{P}(C|\text{Property}) = \frac{1}{2} \tag{6}$$

Example Derivation of a Concept Prior Under this model of the concept prior, the prior probability that the noun *blick* refers to the concept *Dachshund* will have the following derivation. First, we have production counts for nouns that describe kinds $p_{Kind,Noun}$ that were found in our Mechanical Turk study (we found that on average 308 out of 336 nouns were categorized as kinds). From this production count and the total production counts for nouns, we derive the probability of expanding *Concept* to *Kind*.

$$\begin{aligned}\mathbb{P}(Kind|Noun) &= \frac{p_{Kind,Noun} + 1}{\sum\limits_{d \in \{Kind,\ Property,\ Both\}} p_{d,Noun} + 3} \\ &= \frac{308+1}{336+3} = 0.91\end{aligned} \tag{7}$$

Then we find the probability of the concept being *Dachshund* given that it is defined only along the *Kind* dimension, using the height of the branch *Dachshund* and its immediate parent *dog*. These heights were 0.1259 and 0.3115, respectively.

$$\begin{aligned}\mathbb{P}(dachshund|Kind) &= \frac{height(parent(dog)) - height(dog)}{\sum\limits_{C \in \mathcal{K}} height(parent(C) - height(C)} \\ &= \frac{0.1856}{1.7576} = 0.1056\end{aligned} \tag{8}$$

Finally, to compute the prior probability of the concept *Dachshund* given that it is a noun, we multiply the probability of expanding *Concept* to *Kind* by the probability of the concept being *Dachshund*.

$$\begin{aligned}\mathbb{P}(Dachshund|Noun) &= \mathbb{P}(Kind|Noun) \cdot \mathbb{P}(Dachshund|Kind) \\ &= 0.91 \cdot 0.1056 = 0.09696\end{aligned} \tag{9}$$

Concept Likelihood: $\mathbb{P}(X|C)$

We assume that, given a set of objects that are examples of a concept C, each object is equally likely to be chosen by the snail[1]. Therefore, the probability of the data given a concept is proportional to the size of the set of things matching that concept. For example, for the concept *dog*, the probability of picking a particular dog, Fido, is inversely proportional to the number of dogs there are in the scene. So if n objects are chosen by the snail as examples of a concept C, and these objects are plausible examples of the concept,

$$\mathbb{P}(X|C) = \left(\frac{1}{|C|}\right)^n \tag{10}$$

Simulations

For each experimental trial we computed the posterior probability over concepts using both the noun and adjective priors. We assumed that on each trial children were sampling a concept from the posterior distribution over concepts given the syntactic category of the novel word. Thus the posterior probability over concepts as generated by the model should give us the frequency with which a child should show any given behavior. In order to be able to compare the model to the experimental data, we sorted the concepts into the same categories that we used for analyzing the experimental data: subordinate, basic, superordinate and neutral. For example, given the data *striped Dachshund*, the candidate concepts are *striped Dachshund*, *Dachshund*, *striped dog*, *dog*, *striped animal*, *animal*, or *striped*. From this set of candidates, *striped Dachshund* and *Dachshund* mapped onto the subordinate level, *striped dog* and *dog* mapped onto the basic

[1] Xu and Tenenbaum use a different estimate of category sizes for kinds, which is based on the same heirarchy as their concept prior. We found little difference when we compared the our own likelihood distributions with those computed by Xu and Tenenbaum's methods on our experimental items. A very similar ordering applied over concepts, and each item was on the same order of magnitude for both measures of the likelihood.

level, *striped animal* and *animal* mapped onto the superordinate level and *striped* mapped on the neutral level.

The results of our model are shown in Figure 2(b). Overall the model appears to provide a very close fit to the experimental data, with a much higher posterior probability for the subordinate level given a noun, and a much higher probability for the neutral level given an adjective.

Discussion

In this paper we have shown that while children tend to map novel nouns onto a kind hierarchy, they prefer to map novel adjectives onto a property hierarchy. This behavior is predicted if children use their knowledge of grammatical categories and the distributions of different concept types within these categories to constrain the space of hypothesized meanings when learning novel words. A Bayesian model trained on the distribution of concepts across grammatical categories in the English lexicon predicts the same generalization pattern. Together these results suggest that not only are children able to use what they know about grammatical categories when inferring the meanings of novel words, the way they do this is predicted by the distributions of concept types across gramamtical categories in English. Moreover, the constraints imposed on inference by grammatical category are powerful enough to overcome the effect of the size principle on the likelihood.

These findings have several implications for language acquisition and models of language acquisition. First, while the 'size principle' has received considerable attention as a solution to the word learning problem, this work demonstrates that the beliefs children bring to the word learning task also play a key role in word learning. Second, we can ask how children behave with respect to concept hierarchies in languages that collapse the distinction between nouns and adjectives (e.g. Georgian). Does the size principle play a role only to the extent that nouns are likely to draw from the kind hierarchy? Third, as these beliefs are attributable to the distribution of concept types across grammatical categories in the children's own lexicons, there are obvious extensions of this work to modeling the infant word learning by weakening (or making nonexistent or unavailable) the link between between grammatical category and concept hierarchy. There are several findings that would be interesting to model this way, including (1) that 11-month-olds make the same generalizations for words presented as nouns and adjectives and these generalizations are neutral with respect to kind vs. property meanings (Waxman & Booth, 2003), or (2) that the noun-kind link is established earlier than the adj-property link (Booth & Waxman, 2003, 2009). Finally, we can ask to what degree a group of exemplars' distribution on a given concept hierarchy is used in acquiring linguistic phenomena that extend beyond word meanings (e.g. word classes).

Acknowledgments. This research was supported by NSF IGERT 0801465, a NSF GRF to Gagliardi, and a Baggett Fellowship to Bennett. We would like to thank the UMD Cognitive Neuroscience of Language Lab, the UMD Project on Childrens Language Learning and the UMD Computational Psycholinguistics group for helpful discussion and assistance.

References

Austerweil, J. L., & Griffiths, T. L. (2010). Learning hypothesis spaces and dimensions through concept learning. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 73–78). Austin, TX: Cognitive Science Society.

Berwick, R. C. (1963). Learning from positive-only examples: The subset principle and three case studies. In R. S. M. J. G. Carbonell & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach (vol. 2).* Los Altos, CA: Morgan Kauffman.

Booth, A. E., & Waxman, S. R. (2003). Mapping words to the world in infancy: Infants expectations for count nouns and adjectives. *Journal of Cognition & Development*, *4*, 357–381.

Booth, A. E., & Waxman, S. R. (2009). A horse of a different color: Specifying with precision infants' mappings of novel nouns and adjectives. *Child Development*, *80*.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, *112*, 347–382.

Dale, P., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, *28*, 125–127.

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

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

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *29*, 819–865.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.

Waxman, S. R., & Booth, A. E. (2003). The origins and evolution of links between word learning and conceptual organization: New evidence from 11-month-olds. *Developmental Science*, *6*.

Waxman, S. R., & Markow, D. B. (1998). Object properties and object kind: Twenty-one-month-old infants extension of novel adjectives. *Child Development*, *69*, 1313–1329.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*.