Word Learning as Bayesian Inference: Evidence from Preschoolers

Fei Xu
fei@psych.ubc.ca
Department of Psychology
University of British Columbia

Joshua B. Tenenbaum
jbt@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

Most theories of word learning fall into one of two classes: hypothesis elimination or associationist. We propose a new approach to word learning within a Bayesian framework. Tenenbaum and Xu (2000) presented a Bayesian model of adults learning words for hierarchically structured categories. We report two experiments with 3- and 4-year-old children, providing evidence that the basic principles of Bayesian inference are employed when children acquire new words at different hierarchical levels. Implications for theories of word learning are discussed.

Introduction

The problem of word learning has been a well-cited example of the general problem of induction. With each referential act, e.g., "Look, a dog!" there are an infinite number of hypotheses for the meaning of the word "dog" that is consistent with the data (Quine, 1960). The word could refer to all (and only) dogs, all mammals, all animals, all Dalmatians, this individual Max, all dogs plus the Lone Ranger's horse, all dogs except Labradors, all spotted things, the front half of a dog, undetached dog parts, things which are dogs if first observed before next Monday but cats if first observed thereafter, and on and on. Yet despite this severe under-determination, even 2- or 3-year-old children seem to be remarkably successful at learning the meanings of words from just one or a few positive examples (Bloom, 2000; Carey, 1978; Markman, 1989; Regier, 1996; among others). How do they do it?

Most theories of word learning fall under what we call the *hypothesis elimination* approach to learning. The learner effectively considers a hypothesis space of possible concepts onto which words will map, and assumes that each word maps onto exactly one of these concepts. The act of learning consists of eliminating incorrect hypotheses about word meaning, based on a combination of a priori knowledge and observations of how words are used to refer to aspects of experience, until the learner converges on a single consistent hypothesis. More precisely, some logically possible hypotheses may be ruled out a priori because they do not correspond to any psychologically natural concepts, e.g., the hypothesis that "dog" refers to things which are dogs if observed before Tuesday but cats if observed thereafter. Other hypotheses may be ruled out because they are inconsistent with examples of how the word is used, e.g., the hypotheses that "dog" refers to all and only terriers, can be ruled out upon seeing the example of Max the Dalmatian.

Settling on one hypothesis by eliminating all others as incorrect amounts to taking a deductive approach to the logical problem of word learning. The learner essentially deduces the word's meaning from a set of premises, which include the assumption that the word maps onto one of the learner's hypotheses and the a priori knowledge and observational data that rule out all but one hypothesis. The success of word learning is then explained by the deductive validity of this inference. Thus we will sometimes refer to hypothesis elimination approaches as *deductive* approaches to word learning.

To illustrate how word learning has traditionally been explained within a deductive framework, let us return to our opening question of how a child could possibly infer the meaning of the word "dog" from a typical labeling event. One influential proposal has been that children come to the task of word learning equipped with strong prior constraints on viable word meanings (e.g., Markman, 1989), allowing them to rule out a priori many logically possible but psychologically unnatural extensions of a word. Often it is most natural to view these constraints as giving structure to the learner's hypothesis space, but they could also be seen as eliminating the implausible hypotheses from a larger space of all logically possible concepts.

Two classic constraints on the meanings of common nouns include the whole object constraint and the taxonomic constraint (Markman, 1989). The whole object constraint requires words to refer to whole objects, as opposed to parts of objects or attributes of objects, thus ruling out word meanings such as the front half of a dog, or undetached dog parts. The taxonomic constraint requires words refer to taxonomic classes, typically in a tree-structured hierarchy of natural kind categories (Keil, 1979), thus ruling out word meanings such as all dogs plus the Lone Ranger's horse, all spotted things, all running things or all dogs except Labradors. Whether these constraints are specific to the domain of word meaning or reflect more general restrictions on the structure of natural kind concepts is controversial (e.g., Tomasello, 2001), but their importance in guiding the process of word learning is fairly well accepted.

In most cases, such as our example of a child learning the word "dog", these constraints are useful but not sufficient to solve the inference problem. Even after ruling out all hypotheses that are inconsistent with a typical labeled example (e.g., Max the Dalmatian), a learner will still be left with many consistent hypotheses that also correspond to possible meanings of common nouns. How are we to infer whether a word that has been perceived to refer to Max applies to all and only Dalmatians, all and only dogs, all canines, all mammals, or all animals, and so on? This problem of inference in a hierarchy is interesting in its own

right, but more importantly as a special case of the fundamental challenge faced by deductive approaches to word learning. In most interesting semantic domains, the natural concepts that can be named in the lexicon are not mutually exclusive, but overlap in some more or less structured way. In other words, most objects can be described with more than one word. Thus no example can ever rule out all but one of the a priori plausible hypotheses, as the hypothesis elimination framework requires for successful inductive inference.

While deduction or hypothesis elimination may be the dominant framework in which researchers have sought to understand the inferential processes underlying word learning, it is not the only standing candidate. Connectionist or neural network models (e.g., Regier, 1996; Gasser & Smith, 1998) treat word learning as a kind of associative learning process. Similarity-based models treat word learning as a process of exemplar memorization and generalization by graded matching (Landau, Smith, & Jones, 1988). By using internal layers of "hidden" units and appropriately designed input and output representations, or appropriately tuned similarity metrics, these models are able to produce abstract generalizations of word meaning that go beyond what one might expect from their roots as models of the simplest animal learning or memory processes. They also have the potential to capture some aspects of word learning that are not easily explained within the deductive framework, such as the graded nature of many generalizations, the noise tolerance of learning, or the varying degrees of confidence that word learners may have in their inferences. However, associative or similarity-based models have not replaced hypothesis elimination as the dominant way of thinking about word learning (Bloom, 2000). In large part, this is probably because they have not yet exhibited the essential capacities of fast mapping. The term "fast mapping" has come to mean different things for different researchers. We emphasize children's ability not only to form a link between a phonological form and a meaning, but also to generalize to novel instances based on one or just a few positive examples (see Carey & Bartlett, 1978; Markman, 1989; Markson & Bloom, 1997; Waxman & Booth; 2001 for other definitions).

While both deductive and associationist models offer certain insights into the processes of word learning, we believe that neither approach provides an adequate framework in which to explain how people actually learn the meanings of words. Here we propose a novel approach based on principles of rational statistical inference. Our framework combines the principal advantages of both deductive and associationist frameworks: it supports the rational inferences underlying generalization in fast mapping, but it also exhibits a graded sensitivity to uncertainty in prior knowledge and the input. It can be viewed as a natural extension of the hypothesis elimination approach, in which hypotheses are evaluated not by deductive logic but by the machinery of Bayesian probability theory. Thus hypotheses are not simply ruled in or out, but scored according to their probability of being correct. The result is a much wider spectrum of inferences, ranging from complete certainty to complete uncertainty, and including both logical deductions and true inductive leaps based only on suspicious coincidences in the observed data. This allows the Bayesian framework to explain crucial fast mapping phenomena and other word learning behaviors that neither previous framework can make sense of.

We will study a phenomenon in the context of learning words for taxonomic categories, which strongly suggest that a statistical inference mechanism is at work. Suppose that after observing Max the Dalmatian labeled a "fep", and inferring (based on a basic-level preference within a taxonomic hypothesis space) that "fep" refers to all and only dogs, we then see three more objects labeled as feps, each of which is also a Dalmatian. These additional examples are consistent with exactly the same set of taxonomic hypotheses that were consistent with the first example; no potential meanings can be ruled out as inconsistent that were not already inconsistent after seeing one Dalmatian called a "fep". Yet after seeing these additional examples, the word "fep" seems relatively more likely to refer to just Dalmatians than to all dogs. Intuitively, this inference appears to be based on a suspicious coincidence: it would be quite surprising to observe only Dalmatians called "fep" if in fact the word referred to all dogs (and if the first four examples were a random sample of "fep" in the world).

In previous research we presented evidence that adults' performance in a word learning task accords with these intuitions and we presented a Bayesian model of adults' generalization behavior (Tenenbaum & Xu, 2000). Given a hypothesis space and one or more examples of a novel word's referents, the learner evaluates all hypotheses for candidate word meanings by computing their posterior probabilities, proportional to the product of prior probabilities and likelihoods. The prior probabilities embody all of the learner's beliefs about which hypotheses are more or less plausible, independent of the observed examples. Constraints on word meaning (e.g., the whole object assumption or the taxonomic assumption) are part of the prior. Prior probabilities may be innate or learned, and they may change over time as the lexicon grows. The likelihood captures the statistical information in the examples. It reflects the learner's expectations about which examples are likely to be observed given a particular hypothesis about word meaning, e.g., the learner assumes that the observed examples are a representative sample of the word/concept. The posterior captures the learner's degree of belief that the hypothesized word meaning is the true meaning of the word given the examples.

Tenenbaum and Xu (2000) specified how these various terms could be instantiated in the Bayesian model. For the case of learning words for kinds, hypotheses were assumed to correspond to classes in a hierarchical taxonomy of kinds. A representation of subjects' taxonomies was obtained by hierarchical clustering of similarity judgments. The clusters included subordinate, basic-level, and superordinate categories, as well as others (see T&X for details). The more distinctive a cluster was in terms of similarity, the higher the prior probability was that a word would map onto that cluster. The likelihood reflects a size principle: Assuming that the examples are randomly sampled from the word's extension, hypotheses corresponding to smaller

extensions are preferred relative to larger extensions, and the preference increases exponentially as the number of consistent examples increases. This captures the intuition of "suspicious coincidence": If the first example of "fep" is a Dalmatian, either all Dalmatians or all dogs may be plausible hypotheses. But if the first three examples of "fep" are all Dalmatians, the word seems more likely to refer to just the Dalmatians than to all dogs. Lastly, generalization to new objects is determined by averaging the predictions of all hypotheses weighted by their posterior probabilities.

Previous studies showed that adults' data fit well with this model. What about children who are in the midst of rapidly learning new words?

Experiment 1

Experiment 1 investigated how 4-year-old children learn words for subordinate, basic-level and superordinate categories. Children were taught novel words for object categories and were asked to generalize these words to new objects. As in T&X (2000), two factors were manipulated: the number of examples labeled (1 vs. 3) and the range spanned by the examples (e.g., three Dalmatians, three kinds of dogs, or three kinds of animals).

Method

Participants

Participants were thirty-six 4-year-old children (mean age 4 years 1 month, ranged from 3 years 6 months to 5 years 0 months). All participants were recruited from the Greater Boston area by mail and subsequent phone calls. English was the primary language for all children.

Materials

The stimuli were 45 objects. They were distributed across three different superordinate categories (animals, vegetables, and vehicles) and within those, many different basic-level and subordinate-level categories (e.g., Dalmatians/terriers/hush puppies, pelicans, cats, etc.). These stimuli were divided into a training set of 21 stimuli and a test set of 24 stimuli. The test stimuli included subordinate matches (e.g., other terriers), basic-level matches (e.g., other dogs), and superordinate matches (e.g., other animals).

Design and Procedure

Each child was randomly assigned to one of two conditions: the One-example Condition and the Three-example Condition. Each child received a total of 3 trials. In the One-example Condition, every child received 3 trials, one from each domain. In the Three-example Condition, because it might be too demanding to ask children for generalizations at all three levels within a single domain, each child received one trial in each of the three domains.

Children were introduced to a puppet, Mr. Frog, and were told that they were helping the puppet who speaks a different language to pick out the objects he wants. The test array of 24 objects was randomly laid out in front of the child and the experimenter. The experimenter held the puppet and said to the child, "This is my friend Mr. Frog. Can you say "hello" to Mr. Frog?" [Child says "Hello."] These are all of Mr. Frog's toys, and he would like you to play a game with him. Would you like to play a game with Mr. Frog?" [Child says "yes."] Then the experimenter says, "Good! Now, Mr. Frog speaks a different language and he has different names than we do for his toys. He is going to pick out some of them and he would like you to help him pick out the others like the ones he has picked out, okay?" [Child says 'ok.'] Three novel words were used: blick, fep, and dax.

One-example Condition.

On each trial, the experimenter picked out an object from the array, e.g., a green pepper, and labeled it, "See? A blick." Then the child was told that Mr. Frog is very picky. The experimenter said to the child, "Now, Mr. Frog wants you to pick out all the blicks from his toys, but he doesn't want anything that is not a blick. Remember that Mr. Frog wants all the blicks and nothing else. Can you pick out the other blicks from his toys?" The child was then allowed to choose among the 24 test objects to find the blicks and put them in front of Mr. Frog. If a child only picks out one toy, the experimenter reminded him/her, "Remember Mr. Frog wants all the blicks. Are there more blicks?" If a child picks out more than one object, nothing more was said to encourage him/her to pick out more toys. At the end of each trial, the experimenter said to the child, "Now, let's put all the blicks back and play the game again. Mr. Frog is going to pick out some more toys and he would like you to help him pick out others like the ones he picks, okay?" Then another novel word was introduced as before.

Each child received three trials, one from each of the three superordinate categories, e.g., a Dalmatian (animal), a green pepper (vegetable), and a yellow truck (vehicle). The order of the trials and the novel words used (blick, fep, and dax) were counterbalanced across participants.

Three-example Condition.

On each trial, the procedure was the same as in the one-example trial with one important difference: The experimenter picked out one object, labeled it for the child, e.g., "See? A fep." Then she picked out two more objects, one at a time, and labeled each one for the child, e.g., "Look, another fep!" The order of the superordinate category (animal, vegetable, and vehicle), the range spanned by the examples (subordinate, e.g., three very similar Dalmatians; basic, e.g., three different dogs; superordinate, e.g., three different animals), and the novel words were counterbalanced across participants.

Results

Since no child chose any of the distracters in this experiment, all analyses excluded the distracter scores. Figure 1 shows the percentage of responses at the various hierarchical levels.

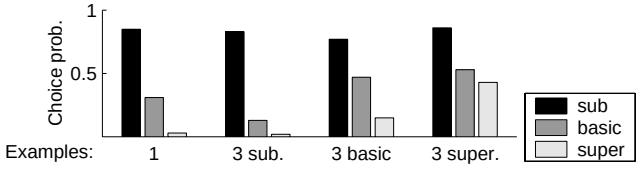

Figure 1: Generalization data from Experiment 1

Two questions are addressed with planned t-tests. First, did children behave differently in the 1-example trials compared with the 3-example subordinate trials when they were given 1 vs. 3 virtually identical exemplars? More specifically, did they show a falling off at the basic level in the 1-example trials and did they restrict their generalization to the subordinate level in the 3-example trials? Second, did the 3-example trials differ from each other depending on the range spanned by the examples? More specifically, did children modify their generalization to the most specific level that was consistent with the set of exemplars?

To investigate the first question, we compared the percentages of responses that matched the example(s) at the subordinate, basic-level, or the superordinate level. On the one-example trials, participants chose more subordinate (85%) than basic-level matches (31%), and more basic-level than superordinate matches (3%) ($p < .0001$ for both comparisons). When presented with three very similar exemplars from the same subordinate category, participants chose more subordinate matches (83%) than both basic-level (13%) and superordinate matches (3%) ($p < .0001$ for both comparisons). Furthermore there was a reliable difference in basic-level matches (31% vs. 13%, $p < .01$).

To investigate the second question, we tested a series of predictions based on our model. A set of planned comparisons address this question by comparing the percentages of response at each level. Given 3 examples from the same subordinate level category, the model predicts a sharp drop between subordinate level generalization and basic-level generalization (83% vs. 13%, $p < .0001$). Given 3 examples from the same basic-level category, the model predicts a sharp drop between basic-level generalization and superordinate level generalization (47% vs. 15%, $p < .0001$). Given 3 examples from the same superordinate category, the model predicts that generalization should include all exemplars from that superordinate category (86%, 53%, and 43%). Children's performance is in broad agreement with the predictions. With three examples, children generalized to the appropriate level consistent with the examples and their generalizations were sharper than with one example.

Discussion

Four-year-old children's performance was in broad agreement with our predictions. On the 1-example trials, they showed graded generalization. Interestingly, they did not show a strong basic-level bias. On the 3-example trials, the children modified their generalizations depending on the span of the examples. Their generalizations were consistent with the most specific category that included all the examples. However, the children's data were much noisier than those of the adults in T&X. Several methodological reasons may account for these differences. The overall level of response was much lower for children. Perhaps the task of freely choosing among 24 objects was too demanding for children of this age and some of them may be reluctant to choose more than a few objects.

In the next experiment, we presented children with each of 10 objects and ask for a yes/no response for each. This modification ensured that all children provide us with judgment on each of the test objects.

The critical prediction made by our Bayesian framework was whether the learner's generalization function differed when labeling a single example vs. three independent examples. However, given that each object was labeled once, the three-example trials contained three times as many labeling events as the one-example trials. Thus we are not able to tell if the learner kept track of the number of examples labeled or simply the number of labeling events (i.e., word-object pairings). This is particularly important since some associative word-learning models (e.g., Colunga & Smith, 2001) claim that keeping track of word-object pairings is the very mechanism of children's word learning. To distinguish the Bayesian approach from associative approaches, it is important to tease apart these possibilities. In Experiment 2 we equated the number of labeling events between the 1- and 3-example trials by labeling the single object three times.

Experiment 2

Method

Participants

Participants were thirty-six 4-year-old children (mean age 4 years 0 months, ranged from 3 years 6 months to 5 years 0 months). Participants were recruited as in Experiment 1.

Materials

The stimuli were the same 45 objects as in Experiment 1, except that the five Dalmatians were replaced by five slightly different terriers.

Design and Procedure

The procedure was identical to that of Experiment 1, except for the following. In the One-example Condition, each object was labeled three times. For example, the experimenter may pick out a green pepper, show it to the child, and say, "See? A fep." She put the pepper down on the floor, then picked it up again, and said, "Look, a fep." She put down and picked up the pepper the third time and said, "It's a fep." The experimenter made sure that the child was following her actions so it was clear that the same pepper had been labeled three times.

In the Three-example Condition, each object was labeled exactly once. Again, the experimenter monitored the child's attention to ensure that joint attention was established before the labeling event for each object.

Although all 24 objects were laid out in front of the child, the experimenter chose 10 of these objects as target objects. The experimenter picked up each of the 10 objects and asked the child, "Is this a fep?" The target set included 2 subordinate matches, 2 basic-level matches, 4 superordinate-level matches, and 2 distracters.

Results

Figure 2 shows the percentage of responses at the various hierarchical levels.

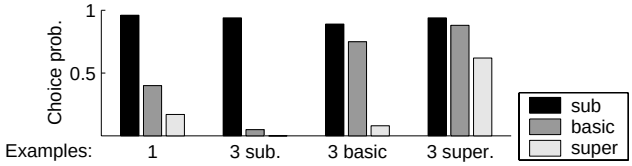

Figure 2: Generalization data from Experiment 2

The same two questions are addressed as in Experiment 1. First, did children behave differently in the 1-example trials compared with the 3-example subordinate trials when they were given 1 vs. 3 virtually identical exemplars? Second, did the 3-example trials differ from each other depending on the span of the examples?

To investigate the first question, we compared the percentages of responses that matched the example(s) at the subordinate, basic-level, or superordinate level. On the one-example trials, participants chose more subordinate (96%) than basic-level matches (40%), and more basic-level than superordinate matches (17%) ($p < .001$ for both comparisons). In contrast, when presented with three very similar exemplars from the same subordinate category, participants chose more subordinate matches (94%) than both basic-level (6%) and superordinate matches (0%) ($p < .0001$ for both comparisons). Furthermore, there was a reliable difference between the basic-level matches (40% vs. 6%, $p < .01$).

To investigate the second question, we tested a series of predictions based on our model. With the modifications on methodology, children's performance is very consistent with our predictions. Given 3 examples from the same subordinate level category, the model predicts a sharp drop between subordinate level and basic-level generalization (94% vs. 5%, $p < .0001$). Given 3 examples from the same basic-level category, the model predicts a sharp drop between basic-level and superordinate level generalization (75% vs. 8%, $p < .0001$). Given 3 examples from the same superordinate category, the model predicts that generalization should include all exemplars from that superordinate category (94%, 88%, and 62%).

Discussion

With the modifications on the experimental design, preschool children showed evidence of computing over the number of examples labeled (not just the number of word-object pairings) and computing over the span of the examples. These results replicated and extended those of Experiment 1, providing stronger evidence for our model.

General Discussion

In order to test specific predictions of the Bayesian framework, our experiments investigated the effects of number of examples (1 vs. 3), span of examples presented to our participants (subordinate, basic, vs. superordinate levels), and number of labeling events (one object labeled three times vs. three objects labeled once each). Each of these experimental design features sheds new light onto the process of word learning.

By varying the number of examples, we were able to examine the effects of multiple examples on generalization. We found that word learning displays the characteristics of a statistical inference, with both adult and child learners becoming more accurate and more confident in their generalizations as the number of examples increased. This effect was not the typical gradual learning curve that is typically associated with "statistical learning"; rather, there was a strong shift in generalization behavior from one to three examples, reflecting the statistical intuition that the span of three independent, randomly sampled examples warrants a sharp increase in confidence for particular hypotheses. These results suggest that both adult and child learners are very sensitive to the "suspicious coincidence" in the input.

By varying the span of examples, we found that labels for subordinate and superordinate categories may not be as difficult to learn as suggested by previous studies. When given multiple examples, preschool children are able to learn words that refer to different levels of the taxonomic hierarchy, at least in the domains of animal, vehicle, and vegetable. Special linguistics cues or negative examples are not necessary for learning these words.

By varying the number of labeling events independent of the number of examples, we were able to explore the ontological underpinning of children's word learning. We found evidence that preschool children are keeping track of the number of instances labeled and not simply the number of co-occurrences between object-percepts and labels. Word learning appears to be fundamentally a statistical inference, but unlike standard associative models, the statistics are computed over an ontology of objects and classes, rather than over surface perceptual features.

Any theory of word learning needs three components: First, what is the body of prior knowledge the learner has coming into the task of learning new words at any given point in time? Second, what are the data required by the learner and what are the data actually available to the learner in order to succeed in acquiring word meanings? Third, what engine of inference is employed by the learner? Furthermore, how these three components interact is crucial for the success of any theory of word learning or inductive inference in general (Tenenbaum & Griffiths, 2001).

We make two main points in adopting a Bayesian inference framework in studying word learning. First, previous theories of word learning (both from the hypothesis elimination tradition and the associative learning tradition) have not endowed the learner with a powerful enough inference engine. Second, some researchers may suggest that what is most important is to characterize the first and the second components above, prior knowledge and input. We argue otherwise. If the inference engine were incorrectly characterized, one would necessarily err in charactering either prior knowledge or the data necessary for successful learning. If the inference engine were too weak, one would need to posit either a great deal of prior knowledge or a lot of input data. By adopting a stronger inference engine than other approaches to word learning, we are able to place stronger constraints on prior knowledge and also on the necessary input data.

The research presented here sheds light on both of these points. What is the right inference engine? Previous literature suggests two candidates: associative learning or hypothesis elimination, neither of which can easily explain our findings here. One major issue with typical associative learning rules is that they are sensitive only to the statistical relations between features, regardless of the nature of those features (e.g., Regier, 1996; Gasser & Smith, 1998). Given multiple features that are all present in all examples of a new word to be learned, standard associative models raise the weights of all these features equally; they do not recognize

that some highly correlated features are more diagnostic than others. A priori biases can be incorporated by adopting different initial values for the weights of different features, but unless one introduces an attentional mechanism, it is difficult to develop a posteriori preferences that discriminate among two features which are equally natural a priori and equally well-correlated with the observation of the word to be learned. In the context of learning words for nested categories, associative learning approaches have difficulty explaining why generalization sharpens up from 1 to 3 virtually identical examples since both sets of examples are consistent with multiple hypotheses represented in terms of correlated features; it also has difficulty choosing the right level of generalization for the same reason.

In contrast, hypothesis elimination approaches run into a different sort of problem. In order to explain the sharpening from 1 to 3 examples, one would have to posit a basic-level bias just for the 1-example case and some version of a bias towards "the smallest category consistent with the examples" just for the 3-example case. Presumably we do not want to have to posit a specific selection principle for each particular case. In addition, positing a basic-level bias makes subordinate and superordinate nouns difficult to learn. Since children do eventually learn these nouns, hypothesis elimination approaches have to posit further constraints to override the basic-level bias.

Although it may be possible to modify existing models to account for these results, the advantage of the Bayesian inference framework is that it explains both the transition from 1 to 3 examples and the appropriate level of generalization without having to posit somewhat post hoc constraints. The graded generalization with 1 example follows straightforwardly from the mechanism of hypothesis averaging. The sharpening from 1 to 3 examples follows straightforwardly from the size principle. For the problem of learning words for kinds, at least, the Bayesian framework provides the most principled and parsimonious account.

Choosing the right inference engine also has implications for the information needed for successful learning, both in terms of prior knowledge and input data. Associative learning requires many examples and sometimes both positive and negative examples. At least in certain domains of word learning (e.g., count nouns for kinds), a small number of examples are sufficient for generalization in both children and adults, and vast majority of the words are learned through positive examples alone (Bloom, 2000). Similarly, hypothesis elimination approaches have to posit specific prior constraints (e.g., the basic-level bias) in order to explain fast mapping. These constraints often have to be overridden by other constraints (e.g., mutual exclusivity so that each category can only have one basic-level label). The advantage of the Bayesian framework is that it arrives at the right generalization pattern from just a few positive examples, and it does not need special linguistic cues or constraints that have to be overridden later on.

In sum, our inductive models may be seen as probabilistic generalizations of the classic deductive approach to word learning based on hypothesis elimination. Our experiments in the domain of words for object categories, with both adults and children, showed that people's patterns of generalization are qualitatively and quantitatively consistent with the Bayesian model's behavior, but not with standard models based on hypothesis elimination, or associative or correlational learning. Bayesian inference may thus offer the most promising framework in which to explain the speed and success of fast mapping in word learning.

The field of cognitive and language development has often been polarized over debates about of whether nature or nurture is key in development. The Bayesian framework we advocate here can perhaps take us beyond this classic "either-or" dichotomy, by showing how both prior knowledge (probabilistic versions of constraints) and observed data (the statistical structure of the examples) can be combined in a rational inference process.

Acknowledgments. JBT was supported by the Paul E. Newton Career Development Chair and FX was supported by a Canada Research Chair. We thank the children for their participation.

References

Bloom, P. (2000) *How children learn the meanings of words.* Cambridge, MA: MIT Press.

Carey, S. & Bartlett, E. (1978) Acquiring a single new word. *Papers and reports on child language development, 15,* 17-29.

Gasser, M. & Smith, L.B. (1998) Learning nouns and adjectives: A connectionist approach. *Language and Cognitive Processes, 13,* 269-306.

Keil, F.C. (1979) *Semantic and Conceptual Development: An ontological perspective.* Cambridge, MA: Harvard University Press.

Markman, E.S. (1989) Categorization and naming in children. Cambridge, MA: MIT Press.

Markson, L. & Bloom, P. (1997) Evidence against a dedicated word learning system in children. *Nature, 385*, 813-815.

Quine, W.V.O. (1960) *Word and object.* Cambridge, MA: MIT Press.

Regier, T. (1996) *The human semantic potential: Spatial language and constrained connectionism.* Cambridge, MA: MIT Press.

Tenenbaum, J.B. & Griffiths, T. (2001) Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629-641.

Tenenbaum, J.B. & Xu, F. (2000) Word learning as Bayesian inference. In L. Gleitman and A. Joshi (eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 517-522). Hillsdale, NJ: Erlbaum.

Tomasello, M. (2001) Perceiving intentions and learning words in the second year of life. In Bowerman & Levinson (eds.), Language acquisition and conceptual development. Cambridge University Press.

Waxman, S.R. & Booth, A.E. (2001) Principles that are invoked in the acquisition of words, but not facts. *Cognition, 77,* B33-B43.