Developmental Science 10:3 (2007), pp 288–297 DOI: 10.1111/j.1467-7687.2007.00590.x

BAYESIAN SPECIAL SECTION

Sensitivity to sampling in Bayesian word learning

Fei Xu[1] and Joshua B. Tenenbaum[2]

1. Department of Psychology, University of British Columbia, Canada
2. Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, USA

Abstract

We report a new study testing our proposal that word learning may be best explained as an approximate form of Bayesian inference (Xu & Tenenbaum, in press). Children are capable of learning word meanings across a wide range of communicative contexts. In different contexts, learners may encounter different sampling processes generating the examples of word–object pairings they observe. An ideal Bayesian word learner could take into account these differences in the sampling process and adjust his/her inferences about word meaning accordingly. We tested how children and adults learned words for novel object kinds in two sampling contexts, in which the objects to be labeled were sampled either by a knowledgeable teacher or by the learners themselves. Both adults and children generalized more conservatively in the former context; that is, they restricted the label to just those objects most similar to the labeled examples when the exemplars were chosen by a knowledgeable teacher, but not when chosen by the learners themselves. We discuss how this result follows naturally from a Bayesian analysis, but not from other statistical approaches such as associative word-learning models.

Introduction

Models for how children learn the meanings of words traditionally fall into two classes. One class of models treats the process as inferential in nature, akin to reasoning. Although the child presumably is not consciously working out each step of the reasoning process and the computations may be done implicitly, the child learner is assumed to draw on a set of hypotheses about candidate word meanings and to evaluate these hypotheses based on observed input using one or more principles of rational inference (e.g. Bloom, 2000; Carey, 1978; Markman, 1989; Siskind, 1996). In contrast, associative models assume that the learner represents a matrix of graded word–object mappings, and the strengths of these mappings are incrementally increased or decreased over time given repeated exposures (e.g. Colunga & Smith, 2005; Gasser & Smith, 1998; Regier, 2003, 2005).

We will argue for an alternative view that combines aspects of both approaches: the basic architecture is a form of rational hypothesis-driven inference, but the inferential logic is Bayesian and hence shows something of the graded statistical character of associative models (Tenenbaum & Xu, 2000; Xu & Tenenbaum 2005, in press). Confronted with a novel word, the learner constructs a hypothesis space of candidate word meanings (i.e. lexicalizable concepts) and a prior probability distribution over that hypothesis space. Given one or more examples of objects labeled by the new word, the learner updates the prior to a posterior distribution of beliefs based on the likelihood of observing these examples under each candidate hypothesis. The prior represents any knowledge (due to previous learning or innate endowment) about which meanings are more or less likely to be the target of the new word, independent of the observed examples. The likelihood is based on the sampling process presumed to have generated the observed object–label pairs.

Recent studies of word learning with adults and children provide some initial evidence for this account. These studies test generalization: participants are shown one or more examples of a novel word (e.g. 'blicket') and are asked to judge which objects from a test set the word also applies to. Xu and Tenenbaum (in press) demonstrated that in learning object kind labels at different levels of the hierarchy (i.e. subordinate, basic level, and superordinate), both the generalization patterns of adults and 4-year-old children were sensitive to the number and the span of the examples, in the ways predicted by a

Address for correspondence: F. Xu, Department of Psychology, 2136 West Mall, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada; e-mail: fei@psych.ubc.ca or J.B. Tenenbaum, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA; e-mail: jbt@mit.edu

simple Bayesian analysis. The most critical finding for distinguishing our Bayesian model from alternative approaches comes from what we call the 'one versus three' contrast, a contrast between patterns of generalization in two different kinds of trials. On a one-example trial, the experimenter pointed to an object, e.g. a toy basset hound, and taught the learner a new word with an utterance such as 'See this? This is a fep!' She continued talking and interacting with the example object, labeling it a total of three times. On a three-example trial, the experimenter pointed to each of three very similar-looking examples, e.g. three basset hounds varying subtly in color, and referred to each one as a 'fep'. In both cases, the learner was then asked to generalize the word to new objects, with test trials including subordinate matches (e.g. other basset hounds), basic-level matches (e.g. other dogs), and superordinate matches (e.g. other animals), as well as non-matching objects (e.g. vegetables and vehicles). In the one-example case, adults and children generalized to subordinate and basic-level matches, but not beyond. In the three-example case, although the input was still consistent with a basic-level interpretation, adults and children restricted their generalization to only subordinate matches.

Standard versions of associative learning models do not predict this contrast, because both types of trials presented the learner with three, practically identical word–object pairings. If anything, there was more variance among the objects in the three-example condition, so it is hard to see why a simple associative learner would generalize more conservatively there. Traditional rational inference accounts, based on eliminating hypotheses inconsistent with the observed examples, also do not predict this difference because exactly the same set of hypotheses is logically consistent with the examples in both cases. Adding a principle for selecting among multiple consistent hypotheses, such as a preference for mapping words to basic-level categories, would still predict the same generalizations in each case (i.e. basic-level generalization).

Our Bayesian model accounts for the one-versus-three contrast with the assumption that the learner makes a rational statistical inference, treating the observed examples as a random sample from the extension of the word to be learned. In both conditions, the most plausible consistent hypotheses for the word's extension comprise a nested set of categories, e.g. all and only basset hounds, all and only dogs, all and only animals. Given just a single example, the probabilities assigned to these hypotheses are not greatly different. The ideal Bayesian learner thus averages the predictions of all these hypotheses, yielding a gradient of generalization that falls off around the basic level. However, given three highly similar examples, each assumed to be randomly drawn from the word's extension, this sample would be a highly suspicious coincidence unless the word had a very narrow extension. The posterior probability thus becomes concentrated on the smallest hypothesis containing the observed examples, leading the learner to generalize the word only to other subordinate matches.

More generally, the Bayesian model's sampling assumption is critical for explaining children's ability to infer a word's meaning from only a few passively observed positive examples. In contrast, standard learning models in both the associative and rational inference traditions take as their data just a set of word–object pairs without regard for how these examples are generated. An important feature of real-world word learning is that learning takes place in a communicative context, and the learner's data may be generated by very different processes in different contexts; the assumption of examples drawn at random from a word's extension will not always be true. A stronger test of the Bayesian approach over sampling-blind alternatives would be to place learners in different communicative contexts, varying the sampling process but keeping the word–object pairings constant, and to test whether learners generalize according to what will now be different ideal Bayesian predictions in these different conditions. The current study was designed to test just such a manipulation.

We compared two experimental conditions. The 'teacher-driven' condition was essentially a replication of the three-example subordinate trials in our previous studies (Xu & Tenenbaum, in press), as described above. The only significant difference was the use of novel artificial objects, rather than familiar objects for which participants might have already known English category labels. In the 'learner-driven' condition, again three examples in the same subordinate class were presented, but only the first example of a 'fep' was provided by the teacher. Then learners were asked to pick out two more feps and were promised a token prize (a sticker) if they correctly picked out these two examples. Ordinarily, if learners can actively choose examples to be labeled, we expect they would choose objects about which they are most uncertain, in order to gain the most information (Nelson, Tenenbaum & Movellan, 2001; Steyvers, Tenenbaum, Wagenmakers & Blum, 2003). However, in this context, we expected that learners would conservatively choose two more objects from the same subordinate category, in order to maximize their chance of being correct and getting the prize. All but one participant in fact made these choices and received the feedback that they had successfully picked out two more feps. Thus participants in the 'learner-driven' condition saw essentially the same three word–object pairings that participants in the 'teacher-driven' condition saw, but now in a context where these objects were chosen independently of the word's meaning (rather than as a random sample from the word's extension) – because the learner (unlike the teacher) does not know the word's meaning. Tenenbaum and Griffiths (2001) referred to this sampling mode as 'weak sampling', in contrast to 'strong sampling' where the examples are drawn from the concept's extension (as in the teacher-driven condition). Under weak sampling, the fact that all three examples fall under the same subordinate category is not a suspicious coincidence, and a Bayesian learner who uses this knowledge should generalize in this case essentially as they would from a single randomly drawn example – to all basic-level matches rather than just the subordinate matches.

A Bayesian model

Before describing our experimental design and results, we give a formal Bayesian analysis of the word-learning scenario studied here, highlighting the role of the sampling process. Our treatment in this section aims for generality, which requires a certain degree of abstraction. The Appendix illustrates these computations more concretely for the specific conditions of the experimental task.

The learner observes n examples $X = \{x_1, x_2, \ldots, x_n\}$ of how the word w is used. Each example x_i consists of an object o_i paired with a label l_i. For simplicity, we assume that the label variable is just binary: $l_i = 1$ denotes that the word w does apply to the ith example (a positive example); $l_i = 0$ denotes that the word does not apply (a negative example). The learner considers a hypothesis space M of candidate word meanings $m \in M$, and evaluates these hypotheses using Bayes' rule:

$$p(m \mid X) = \frac{p(X \mid m)p(m)}{p(X)} \tag{1}$$

$$= \frac{p(X \mid m)p(m)}{\sum\limits_{m' \in M} p(X \mid m')p(m')}. \tag{2}$$

The prior probability $p(m)$, along with the structure of the hypothesis space M, embodies the learner's beliefs about possible or likely word meanings. The likelihood $p(X \mid m)$ expresses the learner's beliefs about which examples would be expected to be observed, if m is in fact the true meaning of the word w to be learned. The posterior probability $p(m \mid X)$ measures the learner's degree of belief in each candidate word meaning m, given the combination of priors and evidence provided by the examples.

Here we will assume simplified forms for the hypothesis space M and prior $p(m)$, because our focus is on how the learner thinks about the process generating the observed examples, which is captured in the likelihood. We will assume that the hypothesis space comprises a tree-structured taxonomy of kind categories, similar to the semantic hierarchies proposed by Collins and Qullian (1969) or Rosch, Mervis, Gray, Johnson and Boyes-Braem (1976). One hypothesis corresponds to each category in the taxonomy. The categories can be arranged into levels, such as the subordinate, basic, and superordinate levels of Rosch *et al*. (1976). For all categories at the same taxonomic level, the prior assigns equal probability. These choices reflect the minimal structure necessary to capture the experiments described below, as well as one of the essential inductive challenges faced by children learning words from real-world experience. Because any object can be construed as a member of multiple categories, at the point where a learner has seen only one or a small number of examples of a word's possible referents – the situation of 'fast mapping' – there is likely to be more than one *a priori* plausible hypothesis consistent with the observed data. However, there may be an expectation that words are more or less likely to map onto a particular level of a category hierarchy, such as an expectation that words are most commonly used to refer to objects at a basic level of categorization (e.g. Markman, 1989). For the specific experiments described below, we represent the objects in terms of a single superordinate category, with two basic-level categories each dividing further into three subordinate categories (Figure 1a). A parameter β represents the ratio of the prior probability assigned to one of the basic-level categories versus one of the subordinate categories. Given the constraint that the prior must sum to one over all hypotheses, fixing β uniquely determines the full prior distribution.

The likelihood is where the distinction between learning from strong sampling and weak sampling appears, and thus where we see effects of inferences about the process generating the observed examples. The full likelihood $p(X \mid m)$ can be decomposed as a product of terms for each example (always three in the experiments we describe below),

$$p(X \mid m) = \prod_{i=1}^{n} p(x_i \mid m). \tag{3}$$

Each term $p(x_i \mid m) = p(o_i, l_i \mid m)$ can be analyzed in terms of both the object o_i and the label l_i that are observed. The dependencies between these two components may be different for different sampling conditions. In particular, strong sampling and weak sampling correspond to qualitatively different structures of dependency between the object chosen for each example and the label that the object receives. Under strong sampling, when the learner

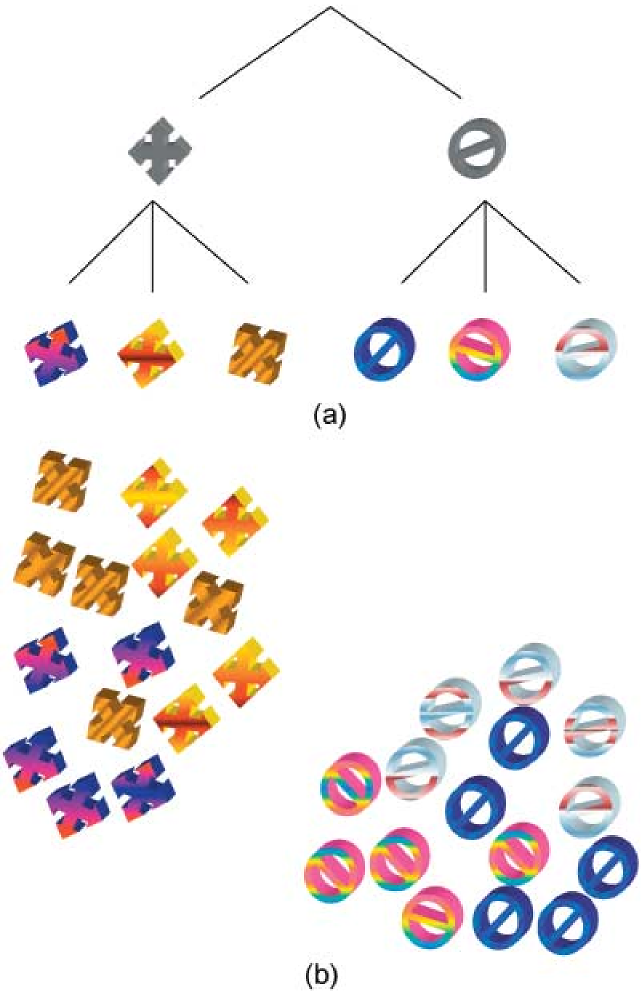

Figure 1 *(a) A schematic illustration of the hypothesis space used to model generalization in the experiment, for the stimuli shown in (b). (b) One set of stimuli used in the experiment, as they were shown to participants.*

is shown a positive example, the label is generated first (depending on what the speaker wants to communicate in a particular context), and then an object is chosen from the set of stimuli with that label. This causal order is reversed in weak sampling: the learner first picks an object to be labeled out of the total set of stimuli, and then the teacher responds with either a positive or negative label depending on whether the object chosen falls inside the word's positive extension.

This difference in causal order affects the inferences that the learner is licensed to draw under strong or weak sampling. Under strong sampling, the fact that a particular example is positive is not informative about the meaning of the word – only that the example is being generated by a knowledgeable speaker who either intends to provide positive examples in an ostensive teaching context, or is merely using words correctly (to refer to their positive instances) in the course of natural conversation with other competent speakers. However, the particular object chosen for labeling is informative about the word's meaning, because it is presumed to be a random sample from the word's positive extension. These dependencies are reversed under weak sampling. The choice of which object to label does not depend directly on the word's meaning, because the learner is choosing the examples and does not know the word's meaning. However, the label observed does depend on the meaning, because the labels are provided by a competent user who knows what the word means and will presumably label the learner's chosen object positively or negatively according to whether or not it falls in the word's extension. The differences in these patterns of causal dependencies, and the probabilistic dependencies they imply, lead a Bayesian learner to make qualitatively different kinds of inferences in these learning situations.

More formally, in the strong sampling condition, we can write

$$\begin{aligned} p(x_i \mid m) &= p(o_i, l_i \mid m) \\ &= p(o_i \mid l_i, m)p(l_i \mid m) \\ &\propto p(o_i \mid l_i, m), \end{aligned}$$

where we have factorized $p(o_i, l_i \mid m)$ according to the causal order of the sampling process, and then dropped the dependence on $p(l_i \mid m)$ which we assume is independent of the word's meaning m, and thus will contribute an arbitrary constant multiple that cancels when we compute posterior probabilities in Equation (2). Assuming for simplicity that only positive examples are observed, and that the object o_i is sampled randomly from all objects in the word's extension, the remaining likelihood term $p(o_i \mid l_i, m)$ is just inversely proportional to $|m|$, the number of stimuli available that the word applies to under the hypothesis that its meaning is m, unless o_i is not contained in the subset of objects that m picks out, in which case the likelihood is 0. If the learner observes n examples generated by strong sampling, the total likelihood (for any hypothesis m consistent with those examples) is

$$p(x_1, \ldots, x_n \mid m) = \left(\frac{1}{|m|}\right)^n. \tag{4}$$

This likelihood function reflects what we have called the *size principle* (Tenenbaum, 1999; Tenenbaum & Griffiths, 2001): more specific meanings, with smaller extensions, are more likely than more general meanings, with larger extensions, when both are consistent with a given set of examples; and the preference for more specific meanings increases exponentially with the number of examples observed. We have previously shown (Xu & Tenenbaum, in press) that this principle is instrumental in explaining the dynamics of word learning for both adults and children, and Tenenbaum (1999, 2000) has shown its applicability to concept learning more broadly. However, the size principle does not hold under all sampling conditions.

Under weak sampling, the likelihood factorizes differently to reflect the different causal order of sampling:

$$\begin{aligned} p(x_i \mid m) &= p(o_i, l_i \mid m) \\ &= p(l_i \mid o_i, m)p(o_i \mid m) \\ &\propto p(l_i \mid o_i, m). \end{aligned} \tag{5}$$

The factorization in the second line reflects the learner first choosing which object is to be labeled, followed by a teacher or knowledgeable speaker providing a label for that object, positive or negative, according to whether it falls under the extension of w. We then drop the dependence on $p(o_i \mid m)$, since the learner does not know the word's meaning, and thus the learner's choice of which object to label does not depend directly on m. (The learner's choice may depend indirectly on m via the examples previously seen, but conditioned on those observations, o_i and m are independent.) The remaining likelihood term $p(l_i \mid o_i, m)$ is simply 1 or 0 depending on whether the label l_i correctly assesses whether object o_i falls under the candidate extension m. That is, $p(l_i \mid o_i, m) = 1$ if $l_i = 1$ and object o_i falls under the candidate extension m, or if $l_i = 0$ and object o_i falls outside the candidate extension m, and 0 otherwise.

The key difference between the strong-sampling and weak-sampling likelihoods is thus that the size principle only holds for the strong-sampling case. The clearest behavioral difference will emerge when the learner observes multiple examples of a new word in which the objects all fall within a single subordinate. The data are logically ambiguous between a subordinate interpretation of the word, and a basic-level interpretation. Under strong sampling, the size principle will heavily favor the subordinate interpretation. Intuitively, it would be a suspicious coincidence to observe a random sample of objects in a basic-level category that all happen to cluster within a single subordinate. However, under weak sampling, if for some reason learners choose or are required to focus their sampling within a single subordinate, then the likelihoods would be uninformative with respect to the basic versus subordinate interpretations of the word. Our experiments explore this contrast, by setting up two corresponding experimental conditions. In the 'teacher-driven' condition, where the learner observes three random examples provided by the teacher, the strong-sampling likelihood should apply for all three examples. Because the examples all fall within a single subordinate, participants are expected to interpret the word as a subordinate label. In the 'learner-driven' condition, a first example is provided by the teacher under strong sampling, but subsequent examples are chosen by the learner under weak sampling, and the pragmatics of the task (a sticker offered as reward) encourage the learner to choose essentially the same objects that are given under a strong-sampling process in the 'teacher-driven' condition. Although the object-label pairs observed are thus essentially the same in both conditions, our Bayesian analysis predicts that participants in the 'learner-driven' condition should favor a basic-level interpretation, because the weak-sampling likelihoods do not provide distinctive evidence in favor of the subordinate hypothesis that would overwhelm an initial tendency towards basic-level generalization.

The experiment presents results from adults and children, along with a concrete instantiation of the above analyses in a computational model of the same task.

Experiment

Methods

Participants

Twenty-four children and 14 adults participated in the study. The children were recruited from our subject database (mean age 4;0 [years;months], ranging from 3;8 to 4;10; half boys/girls). Two additional children were excluded due to refusing to answer any questions or choosing two additional examples from other subordinate categories in the learner-driven condition (see below). The children were randomly assigned to one of two conditions (N = 12 for each condition), the teacher-driven condition (mean age 4;0) and the learner-driven condition (mean age 3;11). The adult participants (mean age 35, ranging from 25 to 40; five men and nine women) were parents living in the greater Vancouver area who brought their children into the lab for other experiments. They were randomly assigned to one of two conditions (N = 7 for each condition).

Design and procedure

Two color pictures of sets of novel objects were used in the study. One set of novel objects is shown in Figure 1b. Each picture depicted two basic-level categories of objects, spatially segregated, and within each of those categories, three subordinate categories of objects differing subtly in color, texture, and orientation. There were five objects in each subordinate category, and 15 objects in each basic-level category. Four nonsense words were used (‘blicket’, ‘tupa’, ‘wug’, and ‘fep’). Which word was used to refer to which picture and which picture was shown first were counterbalanced across participants.

Teacher-driven condition

With child participants, the experimenter began by asking the child to play a game with her. She pointed to an object in the picture and said to the child, ‘See this? It’s a blicket.’ She then pointed to two other objects from the same subordinate category, one at a time, and said to the child, ‘See this one? It’s a blicket.’ She then asked the child to choose a sticker for paying attention and doing a good job in the game. The experimenter then proceeded to the test phase, in which the child was asked to decide which other objects were also blickets. The experimenter pointed to a total of five other objects, and for each one she asked the child, ‘Is this a blicket?’ The five test objects were, in order of questioning, a subordinate match, a non-match from the other basic-level category, a basic-level match from a different subordinate, another subordinate match, and another basic-level match. Participants received no feedback on their answers to test questions. The experimenter then presented the child with the second set of objects and went through the same procedure, using a different novel word. At the end of the study, each child was allowed to choose another sticker for doing a good job.

The procedure for adults was identical to that for children. The adults were told that the study was initially designed for preschoolers, and that stickers were given out to keep the child on task.

Learner-driven condition

The procedure was identical to that of the teacher-driven condition with the following critical difference. After presenting the first example with the phrase, ‘See this? It’s a blicket.’ The experimenter then asked the child, ‘Can you point to two other blickets? If you get both of them right, you get a sticker!’ Once the child picked two more objects, the experimenter confirmed that the child had correctly found two blickets (regardless of the child’s selections) and the child was allowed to choose a sticker as a reward. The experimenter then proceeded to the test phase, pointing to five test objects and asking for each one, ‘Is this a blicket?’ just as in the teacher-driven condition. In both conditions, these five test questions were asked only after participants had seen three labeled examples; the only difference between conditions was the process by which the second and third labeled examples were sampled. The procedure for adults was again identical to that for children. With the exception of one child (whose data were excluded from the analyses), all participants pointed to two other objects from the same subordinate category when asked to point to two other blickets.

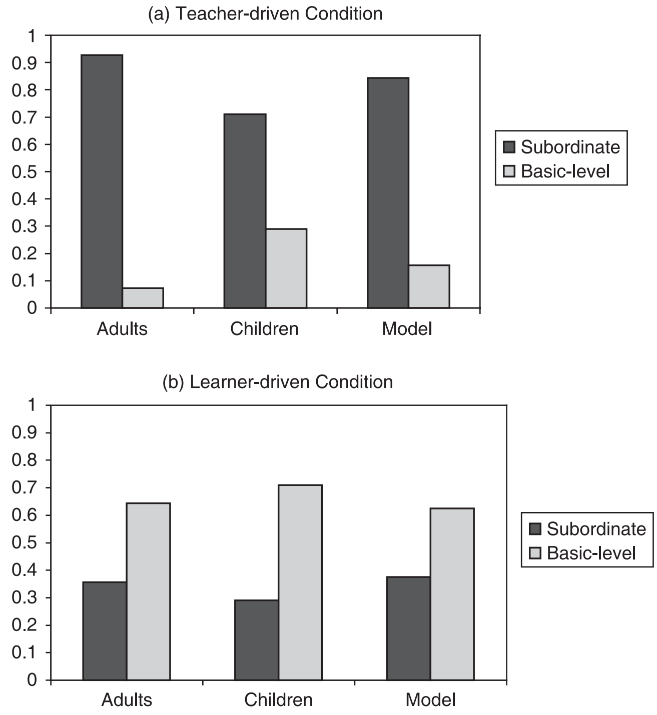

Figure 2 *Percentages of generalization responses at the subordinate and basic levels, for adults and children in both teacher-driven (a) and learner-driven (b) conditions. Corresponding posterior probabilities for subordinate and basic-level hypotheses are shown for the Bayesian model.*

Results

Figure 2 summarizes responses to the test questions in both conditions, in terms of the frequencies with which participants generalized to different levels of the category hierarchy. Over the five test trials, no participant generalized a new word to any object from the other basic-level category, and every participant responded in a way that was consistent with a preferred level of generalization for a given word. That is, for each novel word, each participant either generalized to just the two subordinate-level matches, or to those objects and the two basic-level matches. We calculated the percentages of ‘yes’ responses for both the subordinate and the basic-level matches. In the teacher-driven condition, children generalized the novel word at the subordinate level 71% of the time, and at the basic level 29% of the time. In the learner-driven condition, this preference was exactly reversed: children generalized the novel word at the subordinate level only 29% of the time and at the basic level 71% of the time (shown in Figure 2). To assess whether the difference between these two conditions was statistically significant, each child was assigned a score of 0, 1, or 2 depending on whether they generalized at the basic level for 0, 1, or 2 of the two novel words learned. A Mann-Whitney test between the scores for children in the teacher-driven and learner-driven conditions showed a statistically reliable difference, $z = -1.95$, $p < .05$.

Adult participants gave similar results. In the teacher-driven condition, adults generalized the novel word at the subordinate level 92.8% of the time and at the basic-level 7.2% of the time. In the learner-driven condition, adults generalized the novel word at the subordinate level 35.7% of the time and at the basic-level category 64.3% of the time. (Figure 2 compares these percentages graphically to children's patterns of generalization.) Each adult's overall pattern of generalization was scored on the same 0–2 scale, and a Mann-Whitney test between the scores for adults in the teacher-driven and learner-driven conditions was again statistically significant, $z = -1.98$, $p < .05$.

Model results

Figure 2 also compares the generalization patterns of both children and adults in the teacher-driven and learner-driven conditions with the predictions of our Bayesian model in those same two conditions. The size of each hypothesis is set equal to the number of objects shown in the corresponding category. The basic-level bias in the prior, β, is set to 5. Qualitatively, the model predictions do not depend on the precise value of β, as long as it is somewhat but not vastly greater than one. The specific value chosen provides the best fit found in a coarse search over a range of parameter values. The model results shown in Figure 2 reflect the posterior probabilities of the two hypotheses that are consistent with the observed examples in each case – a basic-level extension and a subordinate-level extension.

To simulate each condition of the study, the model is given the same set of three examples observed by participants in that condition, and the likelihood of each hypothesis is updated based on those three examples assuming the appropriate sampling processes. In the teacher-driven condition (where all three examples were generated under strong sampling), the model assigns posterior probabilities of 84.4% to the subordinate hypothesis and 15.6% to the basic-level hypothesis. In the learner-driven condition (with one example generated under strong sampling and two under weak sampling), the model assigns posterior probabilities of 37.5% to the subordinate hypothesis and 62.5% to the basic-level hypothesis. The Appendix shows concretely how these predictions are derived.

Discussion

In deciding how to generalize a novel word beyond the examples given, children and adults were shown to be sensitive to the sampling process responsible for generating the examples they observed. Their generalizations are in accord with the predictions of a Bayesian analysis, assuming appropriately different likelihood functions – strong sampling and weak sampling – for examples generated by a knowledgeable speaker or by the learner, respectively. Only when learners were justified in assuming that the three very similar examples they observed represented a random sample from the extension of the new word did they restrict their generalization to a higher specific subordinate meaning. When essentially the same three examples were given but with only one as a genuinely random sample of the word's referents, learners recognized this and generalized more broadly, to the basic level, just as in the one-example conditions of our previous studies (Tenenbaum & Xu, 2000; Xu & Tenenbaum, 2005, in press).

Our Bayesian analysis makes a number of simplifying assumptions, which are not essential to the framework and which could be weakened and further explored in future work. We have neglected interactions between multiple words in the learner's developing lexicon, assuming that we can model learning of a single new word on its own. We have considered only the extensional aspects of a word's meaning – the set of objects that it applies to, as opposed to the conceptual content that allows the learner to establish this reference class – and we have assumed that any word either does or does not apply to an object in a binary fashion. We also have not considered the possibility that a word could have more than one meaning. Perhaps most importantly, we have assumed that children (and adults) are able to infer which sampling model would be appropriate to use given the pragmatics of the learning situation, but we have not modeled *how* they make that inference. This is surely a deep and hard problem, arguably drawing on an understanding of discourse, theory-of-mind, and more general social reasoning abilities that might support word learning (e.g. Bloom, 2000; Tomasello, 2001). Our goal here is to present and test a formal approach to modeling inductive inferences about word meaning, which can incorporate multiple relevant capacities and sources of input within a principled and unifying framework for statistical inference. This stands in contrast to typical associative models, which do not naturally incorporate capacities such as the pragmatic and intentional reasoning that appear to be necessary here, or traditional rational inference models, which have not attempted to explain the graded, statistical character of people's generalizations about word meaning in the face of ambiguous evidence.

Standard associative and rational inference models do not predict our results, but could they be reasonably modified to do so? Associative models might be modified to have greater learning rates for trials on which a knowledgeable speaker selects the objects to be labeled, relative to trials on which learners themselves select the examples. However, besides being rather arbitrary and post hoc, this modification seems inconsistent with other general findings about differences between active and passive learning; active data selection typically leads to *more* effective learning than passive observation (e.g. Lagnado & Sloman, 2004; Sobel & Kushnir, in press; Steyvers *et al.*, 2003). It is critical in our learner-driven condition that learners were encouraged by the promise of a reward to select examples conservatively. If they had been encouraged to explore more freely and select the most informative examples, we expect that they would have chosen quite different examples, and thereby converged quickly and confidently on the appropriate level of generalization. Our Bayesian model would be able to explain such a different learning trajectory without fundamental modification, based only on the different informational content of the examples encountered, while a variable learning-rate associative model might need substantial fine adjustments to account for all of these different kinds of learning curves.

It is easier to accommodate our findings from the standpoint of rational hypothesis-driven inference. A growing body of research suggests that word learning depends on the child's sensitivity to pragmatic and intentional cues (e.g. Baldwin, 1991, 1993; Bloom, 2000; Diesendruck & Markson, 2001; Tomasello, 2001; Tomasello & Barton, 1994; Tomasello, Strosberg & Akhtar, 1996; Xu, Cote & Baker, 2005), and our results could be seen as further evidence for the importance of these factors. However, previously there have not been systematic attempts to give a formal model of word learning as hypothesis-driven inference that integrates prior knowledge about conceptual structures underlying word meanings, principles of word–concept mapping, statistical information carried by patterns of observed object–word pairings, and the effects of pragmatic and intentional reasoning. This is essentially the goal of our Bayesian approach. Although we are far from giving a complete account of word learning, even the simple Bayesian model presented here combines aspects of all these major sources of input. Future work should further explore the potential for Bayesian models in these directions: to integrate the diverse sources of constraint that make it possible for children to learn so many words so quickly, and to illuminate the connections between word learning and other aspects of cognition where Bayesian models have recently gained traction – such as the acquisition of syntactic knowledge, intuitive domain theories, and causal knowledge (Perfors, Tenenbaum & Regier, 2006; Regier & Gahl, 2004; Tenenbaum & Griffiths, 2001; Tenenbaum, Griffiths & Niyogi, in press) – which place words and word learning at the center of human mental life.

Appendix

Here we show concretely how the Bayesian model is applied to our experimental task, in both the teacher-driven and learner-driven conditions. The tree-structured hypothesis space is illustrated in Figure 1a. Each hypothesis m for the novel word's meaning refers to one node of the tree. Let m_1 and m_2 represent hypotheses that the word maps onto one or the other of the two basic-level categories in the tree. Let m_{11}, m_{12}, and m_{13} represent the three subordinate hypotheses under m_1, and m_{21}, m_{22}, and m_{23} represent the three subordinate hypotheses under m_2. For simplicity, we do not include a hypothesis for the single superordinate category that includes all objects. We reasoned that participants would assign this hypothesis a very low prior probability on pragmatic grounds; including it, with the same prior probability as the subordinate hypotheses, does not qualitatively change the results.

Prior probabilities are determined by two constraints: basic-level hypotheses are preferred over subordinate hypotheses by a factor of β, and the sum of the prior probabilities must equal one. Hence, $p(m_1) = p(m_2) = \beta/(2\beta + 6)$, and $p(m_{11}) = p(m_{12}) = p(m_{13}) = p(m_{21}) = p(m_{22}) = p(m_{23}) = 1/(2\beta + 6)$. The choice of $\beta = 5$ yields reasonable fits to people's judgments, where $p(m_1) = p(m_2) = 0.3125$, and $p(m_{11}) = p(m_{12}) = p(m_{13}) = p(m_{21}) = p(m_{22}) = p(m_{23}) = 0.0625$.

In both conditions of the experiment, learners observe three labeled examples $X = \{x_1, x_2, x_3\}$ prior to being asked to generalize in the test phase. The examples always fall under a single subordinate category; for concreteness, assume this category corresponds to m_{11}. Thus the likelihoods will be 0 for all hypotheses except two: m_{11}, and m_1 which represents a superset of m_{11}; all other hypotheses are inconsistent with the observed examples. The relative likelihoods of m_{11} and m_1, and hence their posterior probabilities, will depend on the sampling conditions.

Teacher-driven condition

Here all three examples are assumed to be generated independently by strong sampling, so the likelihood of each hypothesis is inversely proportional to its size raised to the third power:

$$p(X|m) = p(x_1|m) \times p(x_2|m) \times p(x_3|m)$$
$$= \frac{1}{|m|} \times \frac{1}{|m|} \times \frac{1}{|m|}.$$

For simplicity we assume that the size of each hypothesis is equal to the number of objects shown in the corresponding category, so $|m_{11}| = 5$ and $|m_1| = 15$. Then we have

$$p(X|m_{11}) = \frac{1}{5} \times \frac{1}{5} \times \frac{1}{5} = 0.008$$
$$p(X|m_1) = \frac{1}{15} \times \frac{1}{15} \times \frac{1}{15} = 0.0002963.$$

Posterior probabilities for the subordinate and basic-level hypotheses can now be computed as follows:

$$p(m|X) = \frac{p(X|m)p(m)}{\sum_{m' \in M} p(X|m')p(m')}$$
$$p(m_{11}|X) = \frac{0.008 \times 0.0625}{0.008 \times 0.0625 + 0.0002963 \times 0.3125} = 0.844$$
$$p(m_1|X) = \frac{0.0002963 \times 0.3125}{0.008 \times 0.0625 + 0.0002963 \times 0.3125} = 0.156.$$

Learner-driven condition

Here the first example is assumed to be generated by strong sampling, but the second and third examples are assumed to be generated by choosing the objects to be labeled independently of the word's true meaning (since the learner has chosen the examples and the learner does not know what the word means). As explained in the main text, the likelihood function associated with these latter examples is effectively just 1 for any hypothesis consistent with the labeled examples, and 0 for all inconsistent hypotheses. We thus have

$$p(X|m) = p(x_1|m) \times p(x_2|m) \times p(x_3|m)$$
$$= \frac{1}{|m|} \times 1 \times 1,$$

or for each of the two consistent hypotheses:

$$p(X|m_{11}) = \frac{1}{5} \times 1 \times 1 = 0.2$$
$$p(X|m_1) = \frac{1}{15} \times 1 \times 1 = 0.0667.$$

Posterior probabilities for the subordinate and basic-level hypotheses can now be computed as follows:

$$p(m|X) = \frac{p(X|m)p(m)}{\sum_{m' \in M} p(X|m')p(m')}$$
$$p(m_{11}|X) = \frac{0.2 \times 0.0625}{0.2 \times 0.0625 + 0.0667 \times 0.3125} = 0.375$$
$$p(m_1|X) = \frac{0.0667 \times 0.3125}{0.2 \times 0.0625 + 0.0667 \times 0.3125} = 0.625.$$

Acknowledgements

The authors contributed equally to this research. JBT was supported by the Paul E. Newton Career Development Chair. FX was supported by a Canada Research Chair and a grant from the Natural Science and Engineering Research Council of Canada. We thank members of the UBC Baby Cognition Lab for their help with data collection, and Paul Bloom, Geoff Hall and Terry Regier for helpful discussion. We owe a particular debt to Liz Bonawitz, for discussions and pilot work on an earlier version of this work.

References

Baldwin, D. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, **62**, 875–890.

Baldwin, D. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, **20**, 394–419.

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

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G.A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.

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

Collins, A.M., & Quillian, M.R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, **8**, 240–248.

Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: a pragmatic account. *Developmental Psychology*, **37**, 630–641.

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

Kemp, C.S., Perfors, A., & Tenenbaum, J.B. (2004). Learning domain structures. *Proceedings of the 26th Annual Conference of the Cognitive Science Society*.

Lagnado, D., & Sloman, S.A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **30**, 856–876.

Markman, E.M. (1989). *Naming and categorization in children*. Cambridge, MA: MIT Press.

Nelson, J.D., Tenenbaum, J.B., & Movellan, J.R. (2001). Active inference in concept learning. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 381–384). Edinburgh: Edinburgh University Press.

Perfors, A., Tenenbaum, J.B., & Regier, T. (2006). Poverty of the stimulus? A rational approach. Manuscript under review.

Regier, T. (2003). Emergent constraints on word-learning: a computational review. *Trends in Cognitive Sciences*, **7**, 263–268.

Regier, T. (2005). The emergence of words: attentional learning in form and meaning. *Cognitive Science*, **29**, 819–866.

Regier, T., & Gahl, S. (2004). Learning the unlearnable: the role of missing evidence. *Cognition*, **93**, 147–155.

Rosch, E., Mervis, C.B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8**, 382–439.

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

Sobel, D.M., & Kushnir, T. (in press). The importance of decision demands in causal learning from interventions. *Memory and Cognition.*

Steyvers, M., Tenenbaum, J.B., Wagenmakers, E.J., & Blum, B. (2003). Inferring causal networks from observations. *Cognitive Science*, **27**, 453–489.

Tenenbaum, J.B. (1999). Bayesian modeling of human concept learning. In M. Kearns, S. Solla, & D. Cohn (Eds.), *Advances in neural information processing systems 11* (pp. 59–68). Cambridge, MA: MIT Press.

Tenenbaum, J.B. (2000). Rules and similarity in concept learning. In S. Solla, T. Leen, & K. Muller (Eds.), *Advances in neural information processing systems 12* (pp. 59–65). Cambridge, MA: MIT Press.

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

Tenenbaum, J.B., Griffiths, T.L., & Niyogi, S. (in press). Intuitive theories as grammars for causal inference. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation*. Oxford: Oxford University Press.

Tenenbaum, J.B., & Xu, F. (2000). Word learning as Bayesian inference. In L. Gleitman & 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 M. Bowerman & S.C. Levinson (Eds.), *Language acquisition and conceptual development*. Cambridge: Cambridge University Press.

Tomasello, M., & Barton, M. (1994). Learning words in non-ostensive contexts. *Developmental Psychology*, **30**, 639–650.

Tomasello, M., Strosberg, R., & Akhtar, N. (1996). Eighteen-month-old children learn words in non-ostensive contexts. *Journal of Child Language*, **23**, 157–176.

Xu, F., Cote, M., & Baker, A. (2005). Labeling guides object individuation in 12-month-old infants. *Psychological Science*, **16**, 372–377.

Xu, F., & Tenenbaum, J.B. (2005). Word learning as Bayesian inference: evidence from preschoolers. In B.G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 2381–2386). Mahwah, NJ: Erlbaum.

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