Word learning as Bayesian inference

Joshua B. Tenenbaum
Department of Psychology
Stanford University
jbt@psych.stanford.edu

Fei Xu
Department of Psychology
Northeastern University
fxu@neu.edu

Abstract

We apply a computational theory of concept learning based on Bayesian inference (Tenenbaum, 1999) to the problem of learning words from examples. The theory provides a framework for understanding how people can generalize meaningfully from just one or a few positive examples of a novel word, without assuming that words are mutually exclusive or map only onto basic-level categories. We also describe experiments with adults and children designed to evaluate the model.

Introduction

Learning even the simplest names for object categories presents a difficult inference problem (Quine, 1960). Given a typical example of the word "dog", e.g. Rover, a black labrador, the possible inferences a learner might make about the extension of "dog" are endless: all (and only) dogs, all mammals, all animals, all labradors, all black labradors, all black things, all running things, this individual animal (Rover), all dogs plus the Lone Ranger's horse, and so on. Yet, even children under five can often infer the approximate extension of words like "dog" given only a few relevant examples of how they can be used, and no systematic evidence of how words are not to be used (Carey, 1978; Markman, 1989; Regier, 1996). How do they do it?

One influential proposal has been that people come to the task of word learning equipped with strong prior knowledge about the kinds of viable word meanings (Carey, 1978; Clark, 1987; Markman, 1989), allowing them to rule out *a priori* the many logically possible but unnatural extensions of a word. For learning nouns, one of the most basic constraints is the *taxonomic assumption*, that new words refer to taxonomic classes, typically in a tree-structured hierarchy of natural kind categories (Markman, 1989). Given the one example of "dog" above, the taxonomic assumption would rule out the subsets of all black things, all running things, and all dogs plus the Lone Ranger's horse, but would still leave a great deal of ambiguity as to the appropriate level of generalization in the taxonomic tree that includes labradors, dogs, mammals, animals, and so on. Other, stronger constraints try to reduce this ambiguity, at the cost of dramatically oversimplifying the possible meanings of words. Under the *mutual exclusivity* constraint, the learner assumes that there is only one word that applies to each object (Markman, 1989). This helps to circumvent the problem of learning without negative evidence, by allowing the inference that each positive example of one word is a negative example of every other word. Having heard Sox called "cat" as well as Rover called "dog", we can rule out any subset including both Rover and Sox (e.g. mammals, animals) as the extension of "dog". But some uncertainty in how far to generalize always remains: does "dog" refer to all dogs, all labradors, all black labradors, or just Rover himself?

Inspired by the work of Rosch et al. (1976), Markman (1989) suggested the even stronger assumption that a new word maps not to just any level in a taxonomy, but to an intermediate or *basic* level. Basic-level categories are intermediate nodes in a taxonomic tree that maximize many different indices of category utility and are widely recognized throughout a culture (Rosch et al., 1976). Whether children really have a bias to map words onto basic-level kinds is controversial (Callanan et al., 1994), but it is certainly a plausible proposal. Moreover, the basic-level constraint, together with the taxonomic constraint and mutual exclusivity, actually solves the induction problem, because each object belongs to one and only one basic-level category. However, this solution only works for basic-level words like "dog", and in fact is counterproductive for all the words that do *not* map to basic level categories. How do we learn all the other words we know at superordinate or subordinate levels? Some experimenters have found that seeing more than one labeled example of a word may help childern learn superordinates (Callanan, 1989), but there have been no systematic theoretical explanations for these findings. Regier (1996) describes a neural network learning algorithm capable of learning overlapping words from positive evidence only, using a weakened form of mutual exclusivity that is gradually strengthed over thousands of learning trials. However, this model does not address the phenomenon of "fast mapping" (Carey, 1978) – the meaningful generalizations that people make from just one or a few examples of a novel word – that is arguably the most remarkable feat of human word learning.

To sum up the problem: taking the taxonomic, mutual exclusivity, and basic-level assumptions literally as hard-and-fast constraints would solve the problem of induction for one important class of words, but at the cost of making the rest of language unlearnable. Admitting some kind of softer combination of these constraints seems like a reasonable alternative, but no one has offered a precise account of how these biases should interact with each other and with the observed examples of a novel word, in order to support meaningful generalizations from just one or a few examples. This paper takes some first steps in that direction, by describing one possible learning theory that is up to the task of fast mapping and applying it to model a simple experimental situation. Our experiments use real, everyday objects with an intuitively clear taxonomic organization, but they require subjects to learn multiple words at different levels of generality which violate the strict versions of mutual exclusivity and the basic-level constraint. Our theory is formulated in terms of Bayesian inference, which allows learners to combine probabilistic versions of *a priori* constraints with the statistical structure of the examples they observe, in order to acquire the sort of rich, multileveled vocabulary typical of natural languages.

The paper is organized as follows. Section 2 describes our basic word learning experiment and presents data from adult participants. Section 3 describes the Bayesian learning theory and its application to modeling the data in Section 2. Section 4 concludes and discusses some preliminary data from a parallel experiment with children.

Experiments with adult learners

Our initial experiments were conducted with adult learners, although the studies have been designed to carry over to preschoolers with minimal modification. The experiment consists of two phases. In the *word learning* phase, participants are given one or more examples of words in a novel language and asked to pick out the other instances that each word applied to, from a large set of test objects. In the *similarity judgment* phase, participants judge the similarity of pairs of the same objects used in the first phase. The average similarity judgments are then submitted to a hierarchical clustering algorithm, in order to reconstruct a representation of the taxonomic hypothesis space that people were drawing on in the word learning phase.

Participants. Participants were 25 students from MIT and Stanford University, participating for pay or partial course credit. All participants carried out the word learning task and the first nine also participated in the similarity judgment phase that followed.

Materials. The stimulus set consisted of digital color photographs of 45 real objects. This set was structured hierarchically to mirror, in limited form, the structure of natural object taxonomies in the world. Objects were distributed across three different superordinate categories (animals, vegetables, vehicles) and within those, many different basic-level and subordinate categories. The 45 stimuli were divided into a test set of 24 stimuli and a training set of 21 stimuli.

The training stimuli were grouped into 12 *nondisjoint* sets of examples. The first three sets contained one example each: a dalmatian, a green pepper, or a yellow truck, representing the three main branches of the microworld's taxonomy. The remaining nine sets contained three examples each: one of the three objects from the single-example sets (the dalmatian, green pepper, or yellow truck), along with two new objects that matched the first at either the subordinate, basic, or superordinate level of the taxonomy. For example, the dalmatian was paired with two other dalmatians, with two other dogs (a mutt and a terrier), and with two other animals (a pig and a toucan) to form three of these nine multiple-example sets.

The test set consisted of objects matching the labeled examples at all levels: subordinate (e.g., other dalmatians), basic (non-dalmatian dogs), and superordinate (non-dog animals), as well as many non-matching objects (vegetables and vehicles). In particular, the test set always contained exactly 2 subordinate matches (e.g. 2 other dalmatians), 2 basic-level matches (labrador, hushpuppy), 4 superordinate matches (cat, bear, seal, bee), and 16 nonmatching objects.

Procedure. Stimuli were presented on a computer monitor at normal viewing distance. Participants were told that they were helping a puppet who speaks a different language to pick out the objects he needs. Following a brief familiarization in which participants saw all 24 of the test objects one at a time, the experiment began with the word learning phase. This phase consisted of 32 trials in which learners were shown pictures of one or more labeled examples of a novel monosyllabic word (e.g. "blick") and were asked to pick out the other "blicks" from the test set of 24 objects by clicking on-screen with the mouse. On the first three trials, participants saw only one example of each new word, while on the next nine trials they saw three examples of each word.[1] Subject to these constraints, the 12 example sets appeared in a pseudo-random order that counterbalanced the order of example content (animal, vegetable, vehicle) and example specificity (subordinate, basic, superordinate) across participants. The frequencies with which each test objects was selected by participants when asked to "pick out the other blicks" were the primary data.

In the similarity judgment phase that followed these trials, participants were shown pairs of objects from the main study and asked to rate their similarity on a scale of 1 to 9. They were instructed to base their ratings on the same aspects of the objects that were important to them in making their choices during the main experiment. Similarity judgments were collected for all but six of the 45 objects used in the word learning experiment; these six were practically identical to six of the included objects and were omitted to save time. Each participant in this phase rated the similarity of all pairs of objects within the same superordinate class and one-third of all possible cross-superordinate pairs chosen pseudorandomly, for a total of 403 judgments per participant (executed in random order). Similarity ratings for all nine participants were averaged together for analysis.

Results and discussion. The results of the word learning phase are depicted in Figure 1. Figure 1a presents data collapsed across all category types (animals, vehicles, and vegetables), while Figures 1b-d show the data for each category individually. The four plots in each row correspond to the four different kinds of example sets (one, three subordinate, three basic, three superordinate), and the four bars in each plot correspond to test objects matching the example(s) at each of four different levels of specificity (subordinate, basic, superordinate, nonmatching). Bar height (between 0 and 1)

[1] The last 20 trials used different stimulus combinations to explore a different question and will not be analyzed here.

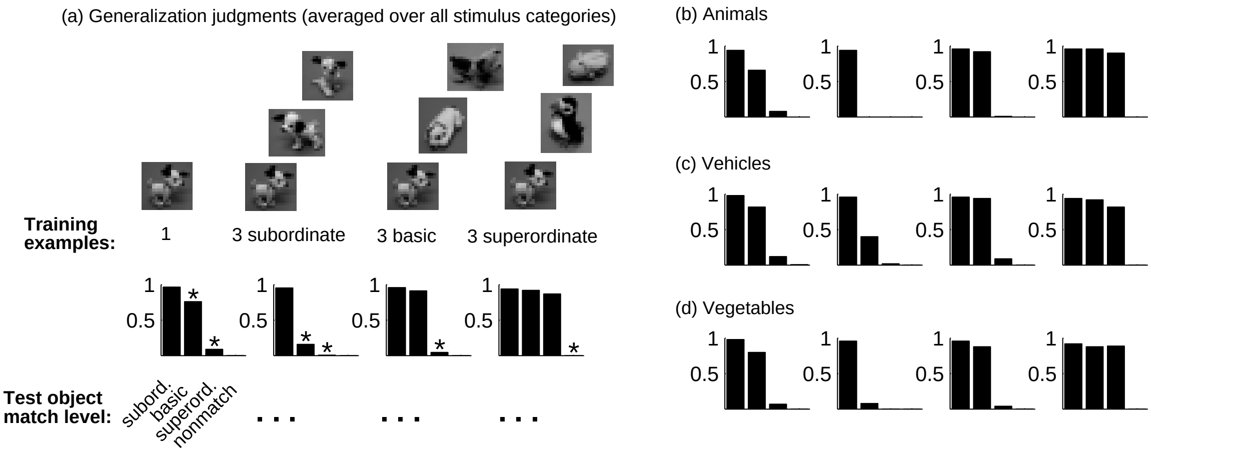


Figure 1: Generalization judgments averaged across categories (a) and broken down into individual categories (b-d).

represents the average probabilities with which participants chose to generalize to the corresponding kind of test object. In Figure 1a, asterisks denote probabilities that are significantly lower than the probabilities to the immediate left ($p < .05$, one-tailed paired t-tests (df = 24) with Bonferonni correction for 12 comparisons), indicating significant gradients of generalization.

The first plots in each row represent trials in which only a single labeled example was provided. Across all three major categories, participants generalized almost always (97% of trials) to test objects matching the example at the subordinate level (e.g., other dalmatians), often but not always (76% of trials) to basic-level matches (e.g., non-dalmatian dogs), rarely (9% of trials) to superordinate matches (e.g., non-dog animals), and practically never (< 1% of trials) to nonmatching test objects (e.g., vegetables or vehicles). Thus, generalization from one example appears to fall off according to a gradient of exemplar similarity, with a threshold located around the basic level.

A different pattern emerges in the last three plots of each row, representing trials on which three labeled examples were provided. Instead of a gradient of generalization decreasing with similarity to the example, there appears in most cases to be a sharp transition from near-perfect generalization to near-zero generalization. The cut-off occurs at the level of the most specific category containing all three labeled examples. That is, given three dalmatians as examples of “blicks”, participants generalized to all and only the other dalmatians; given three dogs, to all and only the dogs, and so on.

Two aspects of these results are consistent with the existing literature on word learning in children. First, we found what appears to be a basic-level bias in generalizing from one example. This interpretation is complicated by the fact that our participants already knew a very familiar word in English for each of the basic-level categories used in our study, “pepper”, “truck”, and “dog”. The tacit knowledge that objects are almost always named spontaneously at the basic level (Rosch et al., 1976) may have increased participants' propensity to map words in a new language onto these basic-level categories, and this bias could exist over and above any preference children or adults might have to map words for unfamiliar objects onto basic-level categories. Second, we found that giving participants more than one example had a dramatic effect on how they generalized to new objects, causing them to select all objects at the most specific taxonomic level spanned by the examples and no objects beyond that level. This finding is consistent with developmental studies in which children given two examples from different basic-level categories were significantly more likely to generalize to other objects of the same superordinate category, relative to children given only a single example (Callanan, 1989).

Our results also differ from the developmental literature in important ways. First, we found a qualitative difference in generalization from one labeled example versus several labeled examples. While generalization from a single example decreased according to a gradient of similarity to the test objects, generalization from three examples followed more of an all-or-none, threshold pattern. Second, we found that people could use multiple examples to infer how far to generalize a new word at any level of specificity in a multi-level taxonomy of object kinds, not just at the basic or superordinate levels.

Figure 2 shows the results of a hierarchical clustering (“average linkage”) analysis applied to participants' similarity judgments from the second phase of the experiment. Each leaf of the tree corresponds to one object used in the word learning phase. (For clarity, only objects in the training set are shown.) Each internal node corresponds to a cluster of stimuli that are on average more similar to each other than to other, nearby stimuli. The height of each node represents the average pairwise dissimilarity of the objects in the corresponding cluster, with lower height indicating greater average similarity. The length of the branch above each node measures how much *more* similar on average are that cluster's members to each other than to objects in the next nearest cluster, i.e., how distinctive that cluster is.

This cluster tree captures in an objective fashion much of people's intuitive knowledge about this domain of objects. Each of the main classes underlying the choice of

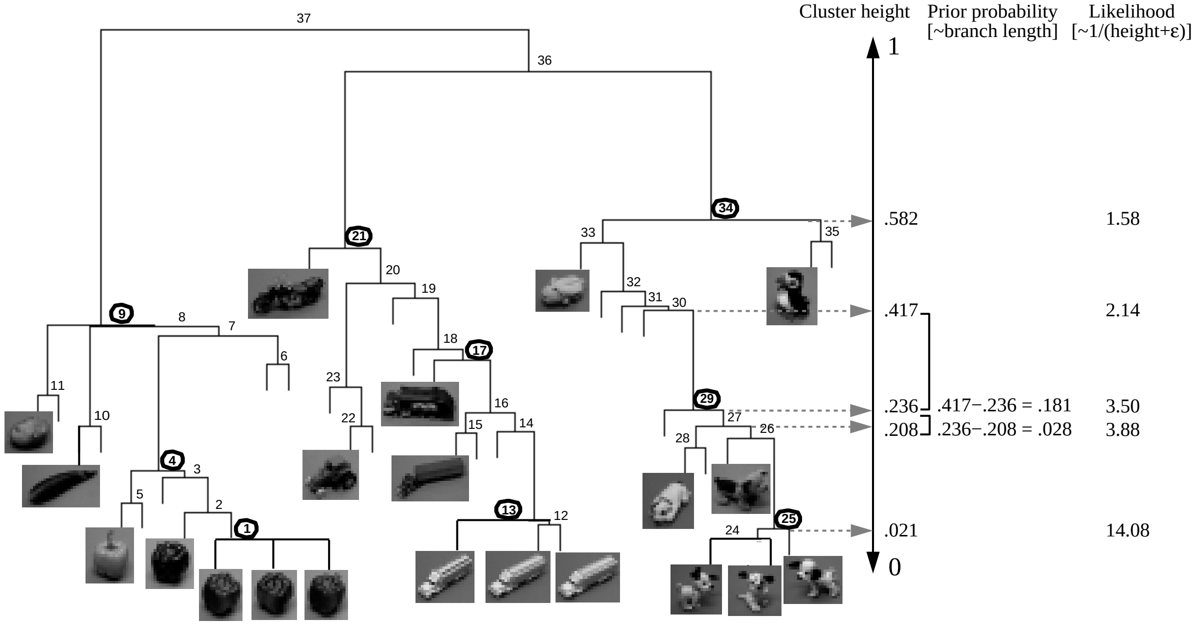


Figure 2: Hierarchical clustering of similarity judgments yields a taxonomic hypothesis space for word learning.

stimuli (vegetable, vehicle, animal, pepper, truck, dog, green pepper, yellow truck, and dalmatian) corresponds to a node in the tree (marked by a circled number). Moreover, most of these clusters are highly distinctive, i.e., well-separated from other clusters by long branches, as one would expect for the targets of kind terms. Other naturally "nameable" nodes include cluster #23, containing the tractor, the bulldozer, and the crane, but no other vehicles, or cluster #33, containing all and only the mammals. Still other clusters reflect more subtle similarities. For example, cluster #18 includes all of the trucks and also the yellow schoolbus. While the schoolbus does not fall into the class of trucks, it intuitively comes much closer than any other non-truck object in the set. This intuitive taxonomy of objects will form the basis for the formal Bayesian model of fast mapping described next.

A Bayesian model

We first describe the general approach, saving the details for below. We assume that the learner has access to a hypothesis space $\mathcal{H}$ of possible concepts and a probabilistic model relating hypotheses $h \in \mathcal{H}$ to data X. Let $X = \{x^{(1)}, \ldots, x^{(n)}\}$ denote a set of n observed examples of a novel word C. Each hypothesis h can be thought of as a pointer to some subset of objects in the world that is a candidate extension for C. The Bayesian learner evaluates these hypotheses by computing their *posterior* probabilities $p(h|X)$, proportional to a product of *prior* probabilities $p(h)$ and *likelihoods* $p(X|h)$:

$$p(h|X) \propto p(X|h)p(h) \qquad (1)$$

The prior, along with the structure of the hypothesis space, embodies the learner's pre-existing (though not necessarily innate) biases, such as the taxonomic or basic-level assumptions. The likelihood captures the statistical information inherent in the examples. The posterior reflects the learner's degree of belief that h is in fact the true extension of C, given a rational combination of her observations X with her relevant prior knowledge about possible word meanings.

The hypothesis space. Tenenbaum (1999) introduced this Bayesian framework for learning simple concepts with hypotheses that could be represented as rectangular regions in a multidimensional continuous feature space. Here we adapt that framework to the task of word learning, assuming that the hypotheses can be represented as clusters in a tree-structured taxonomy (e.g., Figure 2). Such a hypothesis space is clearly not appropriate for learning all kinds of words, but it may be a good first approximation for learning common nouns under the taxonomic assumption. Assuming a tree-structured hypothesis space makes the model more tractable but is by no means a requirement of the Bayesian framework. In principle, any subset of objects could be a hypothesis under consideration.

Priors and likelihoods. Both priors and likelihoods can be defined in terms of the geometry of the cluster tree. The crucial geometrical feature is the height of node h in the tree, which is scaled to lie between 0 (for the lowest node) and 1 (for the highest node) and measures the average dissimilarity of objects within h.

We take the prior $p(h)$ to be proportional to the branch length separating node h from its parent:

$$p(h) \propto \text{height}(\textsc{parent}(h)) - \text{height}(h). \qquad (2)$$

This captures the intuition that more distinctive clusters are *a priori* more likely to have distinguishing names. For example, in Figure 2, the class containing all and only the dogs (#29) is highly distinctive, but the classes immediately under it (#27) or above it (#30) are not nearly as distinctive; accordingly, #29 receives a much higher prior than #27 (proportional to .181 vs. .028).

The likelihood function comes from assuming that the observed positive examples are sampled at random (and independently) from the true concept to be learned. Imagine that each hypothesis consisted of a finite set of K objects. Then the likelihood of picking any one object at random from a set of size K would be $1/K$, and for n objects (sampled with replacement), $1/K^n$. Hence set size is crucial for defining likelihood. While we do not have access to the "true" size of the set of all dogs in the world, or all vegetables, we do have access to a psychologically plausible proxy, in the average within-cluster dissimilarity (as measured by cluster height in Figure 2). Moving up in the tree, the average dissimilarity within clusters increases as they become larger. Thus equating node height with approximate cluster size, we have for the likelihood

$$p(X|h) \propto \left[\frac{1}{\text{height}(h) + \epsilon}\right]^n, \tag{3}$$

if $x_i \in h$ for all i, and 0 otherwise. (We add a small constant $\epsilon > 0$ to height(h) to keep the likelihood from going to infinity at the lowest nodes in the tree (with height 0). The exact value of ϵ is not critical; we found best results with $\epsilon = 0.05$.) Equation 3 embodies the *size principle* for scoring hypotheses: smaller hypotheses assign greater likelihood than do larger hypotheses to the same data, and they assign exponentially greater likelihood as the number of consistent examples increases. This captures the intuition that given a dalmatian as the first example of "blick", either all dalmatians or all dogs seem to be fairly plausible hypotheses for the word's extension (with a likelihood ratio of $14.08/3.50 \approx 4$ in favor of just the dalmatians). However, given three dalmatians as the first three examples of "blick", the word seems much more likely to refer only to dalmatians than to all dogs (with a likelihood ratio now proportional to $(14.08/3.50)^3 \approx 65$ in favor of just the dalmatians).

Generalization. Given these priors and likelihoods, the posterior $p(h|X)$ follows directly from Bayes' rule (Equation 1). Finally, the learner must use $p(h|X)$ to decide how to generalize the word C to new, unlabeled objects. $p(y \in C|X)$, the probability that some new object y belongs to the extension of C given the observations X, can be computed by averaging the predictions of all hypotheses weighted by their posterior probabilities $p(h|X)$:

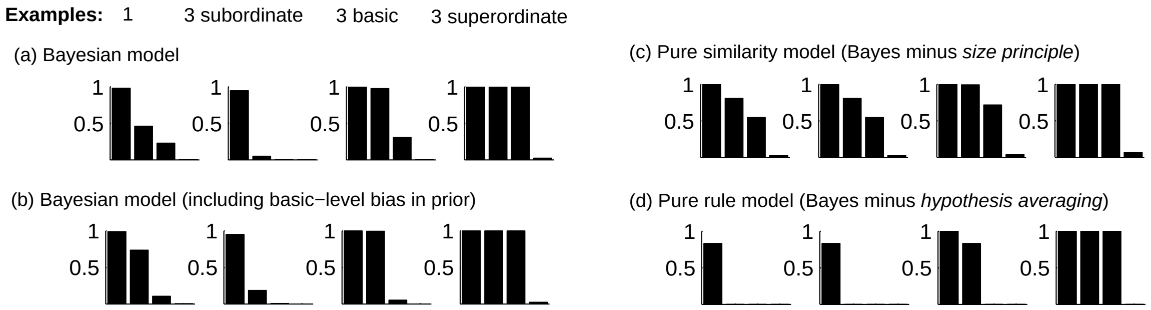
$$p(y \in C|X) = \sum_{h \in \mathcal{H}} p(y \in C|h) p(h|X). \tag{4}$$

To evaluate Equation 4, note that $p(y \in C|h)$ is simply 1 if $y \in h$, and 0 otherwise.

Model results. Figure 3a compares $p(y \in C|X)$ computed from the Bayesian model with the average generalization data from Figure 1a. The model achieves a reasonable quantitative fit ($R^2 = .93$) and also captures the main qualitative features of the data: a similarity-like gradient of generalization given one example, and more all-or-none, rule-like generalization at the most specific consistent level, given three examples. The main errors seem to be too little generalization to basic-level matches given one example or three subordinate examples, and too much generalization to superordinate matches given three basic-level examples. All of these errors would be explained if participants in the word learning task had an additional basic-level bias that is not captured in their similarity judgments. Figure 3b shows the fit of the Bayesian model after adding a bias to the prior that favors the three basic-level hypotheses. With this one free parameter, the model now provides an almost perfect fit to the average data ($R^2 = .98$). Figures 3c and 3d illustrate respectively the complementary roles played by the size principle (Equation 3) and hypothesis averaging (Equation 4) in the Bayesian framework. If instead of the size principle we weight all hypotheses strictly by their prior, Bayes reduces to a similarity-like feature matching computation that is much more suited to the generalization gradients observed given one example than to the all-or-none patterns observed after three examples ($R^2 = .74$ overall). If instead of averaging hypotheses we choose only the most likely one, Bayes essentially reduces to finding the most specific hypothesis consistent with the examples. Here, that is a reasonable strategy after several examples but far too conservative given just one example ($R^2 = .78$ overall). Similarity-based models of category learning that incorporate selective attention to different stimulus attributes (Kruschke, 1992) could in principle accomodate these results, but not without major modification. These models typically rely on error-driven learning algorithms, which are not capable of learning from just one or a few positive examples and no negative examples, and low-dimensional spatial representations of stimuli, which are not well-suited to representing a broad taxonomy of object kinds.

Conclusions and future directions

Research on word learning has often pitted rule-based accounts (Clark, 1973) against similarity-based accounts (Jones & Smith, 1993), or rationalist accounts (Bloom, 1998) versus empiricist accounts (Quine, 1960). In contrast, our work suggests both a need and a means to move beyond some of these classic dichotomies, in order to explain how people learn a hierarchical vocabulary of words for object kinds given only a few random positive examples of each word's referents. Rather than finding signs of exclusively rule- or similarity-based learning, we found more of a transition, from graded generalization after only one example had been observed to all-or-none generalization after three examples had been observed. While special cases of the Bayesian framework corresponding to pure similarity or rule models could accomodate either extremes of this behavior, only the full Bayesian model is capable of modeling the transition from similarity-like to rule-like behavior observed on this task. The Bayesian framework also brings together theoretical constraints on possible word meanings, such as the taxonomic and basic-level biases, with statistical principles more typically associated with the empiricist tradition, such as the size principle and hypothesis averaging. No one of these factors works without the others. Constraints provide sufficient structure in the learner's hypothesis space and prior probabilities to enable reasonable statistical inferences of word meaning from just



(a) Bayesian model

(b) Bayesian model (including basic−level bias in prior)

(c) Pure similarity model (Bayes minus *size principle*)

(d) Pure rule model (Bayes minus *hypothesis averaging*)

Figure 3: Predictions of the basic Bayesian model and three variants for the data in Figure 1.

a few random positive examples.

Still, the hardest questions of learning remain unsolved. Where does the hypothesis space come from? Are constraints on the hypothesis space learned or innate? In ongoing work, we are exploring how unsupervised learning algorithms might be used to bootstrap a hypothesis space for supervised concept learning. For example, can clustering algorithms like the one we used to construct our taxonomic hypothesis space still be successfull when applied to more primitive perceptual representations of objects, instead of adult humans' similarity judgments? Generalizations of the Bayesian framework also hold some promise as bootstrapping mechanisms, in virtue of their ability to propagate probabilistic information from raw data up to increasingly higher levels of abstraction. Perhaps we begin life with a hypothesis space of hypothesis spaces – each embodying different possible constraints on word meanings – and grow into the most useful ones – those which consistently contain the best explanations of the word-to-world pairings we encounter – through the same mechansims of Bayesian inference used to learn any one novel word.

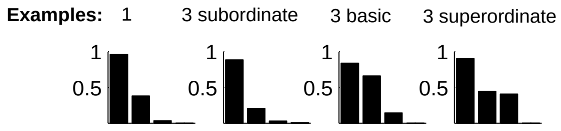


Figure 4: Data from child word learners.

We are also working to extend this line of research to studies of child learners, and to studies of both adults and children learning words for novel objects. Figure 4 shows some promising pilot data from a study with 4-year-old children, using familiar objects in a design approximately parallel to the above adult study. Like the adults, children given three examples of a novel word adapt their generalizations to the appropriate level of specificity, although their superordinate generalizations are less consistent. When given just one example, children show a gradient of generalization much like the adults, but with significantly fewer responses at the basic level and above. If anything, children's overall patterns of responses look more like the Bayesian model's predictions *without* the added basic-level bias (Figure 3a) than *with* that added bias (Figure 3b). Consistent with Callanan et al. (1994), this suggests that a strong basic-level bias may not be a fundamental building block of early word learning – at least, not as distinct from the more general bias in favor of labeling distinctive clusters that the Bayesian model assumes – but rather develops later as the child gains experience about how words are typically used. This issue is one aspect of a broader question: to what extent should differences between child and adult word learners be attributed to differences in their hypothesis spaces, probability models (e.g., priors), or learning algorithms? We hope to answer these questions as we conduct more extensive studies with child learners.

References

Bloom, P. (1998). Theories of word learning: Rationalist alternatives to associationism. In Bhatia, T. K. and Ritchie, W. C. (eds.), *Handbook of Language Acquisition.* Academic Press.

Carey, S. (1978). The child as word learner. In Halle, M., Bresnan, J., and Miller, G. A. (eds.), *Linguistic Theory and Psychological Reality.* MIT Press.

Callanan, M. A. (1989). Development of object categories and inclusion relations: Preschoolers' hypotheses about word meanings. *Developmental Psychology*, 25(2):207–216.

Callanan, M. A., Repp, A. M., McCarthy, M. G., and Latzke, M. A. (1994). Children's hypotheses about word meanings: Is there a basic level constraint? *Journal of Experimental Child Psychology*, 57:108–138.

Clark, E. V. (1973). What's in a word? On the child's acquisition of semantics in his first language. In Moore, T. E. (ed.), *Cognitive Development and the Acquisition of Language.* Academic Press.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In MacWhinney, B. (ed.), *The 20th Annual Carnegie Symposium on Cognition.* Erlbaum.

Jones, S. S. and Smith, L. B. (1993). The place of perception in children's concepts. *Cognitive Development*, 8:113–140.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44.

Markman, E. M. (1989). *Categorization and Naming in Children: Problems of Induction.* MIT Press.

Quine, W. V. (1960). *Word and Object.* MIT Press.

Regier, T. (1996). *The Human Semantic Potential: Spatial Language and Constrained Connectionism.* MIT Press.

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

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In Kearns, M. J., Solla, S. A., and Cohn, D. A. (eds.), *Advances in Neural Information Processing Systems 11.* MIT Press.