# Word Learning as Bayesian Inference

Fei Xu
*Department of Psychology*
*University of British Columbia*

Joshua B. Tenenbaum
*Department of Brain and Cognitive Sciences*
*Massachusetts Institute of Technology*

Abstract

This paper presents a Bayesian framework for understanding how adults and children learn the meanings of words. The theory explains how learners can generalize meaningfully from just one or a few positive examples of a novel word's referents, by making rational inductive inferences that integrate prior knowledge about plausible word meanings with the statistical structure of the observed examples. The theory addresses shortcomings of the two best-known approaches to modeling word learning that are based on deductive hypothesis elimination or associative learning. Three experiments with adults and children test the Bayesian account's predictions in the context of learning words for object categories at multiple levels of a taxonomic hierarchy. Results provide strong support for the Bayesian account over competing accounts, both in terms of quantitative model fits and the ability to explain important qualitative phenomena. Several extensions of the basic theory are discussed, illustrating the broader potential for Bayesian models of word learning.

## Word Learning as Bayesian Inference

Learning even the simplest names for object categories presents a difficult induction problem (Quine, 1960). Consider a typical dilemma faced by a child learning English. Upon observing a competent adult speaker use the word "dog" in reference to Max, a particular Dalmatian running by, what can the child infer about the meaning of the word "dog"? The potential hypotheses appear endless. 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, all running 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 underdetermination, even 2- or 3-year-olds seem to be remarkably successful at learning the meanings of words from examples. In particular, children or adults can often infer the approximate extensions of words such as "dog" given only a few relevant examples of how the word can be used, and no systematic evidence of how words are not to be used (Bloom, 2000; Carey, 1978; Markman, 1989; Regier, 1996). How do they do it?

Two broad classes of proposals for how word learning works have been dominant in the literature: *hypothesis elimination* and *associative learning*. Under the hypothesis elimination approach, the learner effectively considers a hypothesis space of possible concepts onto which words will map, and (leaving aside for now the problem of homonyms and polysemy) 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. Some logically possible hypotheses may be ruled out a priori because they do not correspond to any natural concepts that the learner possesses, e.g., the hypothesis that "dog" refers to things which are dogs if first observed before next Monday but cats if first 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 cats, or 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, and we sometimes refer to these approaches as *deductive* approaches. Hypothesis elimination has its roots in early accounts of human and machine concept learning (Bruner, Goodnow & Austin, 1956; Mitchell, 1982), and it corresponds to one of the standard paradigms considered in formal analyses of natural language syntax acquisition (Gold, 1967; Pinker, 1979). It is also related to classic inferential frameworks that have been considered in the philosophy of science, including Popper's (1959) falsificationism, and the eliminative induction of Mill (1843) and Bacon (1620).

Many variants of hypothesis elimination are present in the word learning literature. Pinker (1984, 1989), Berwick (1986), and Siskind (1996) propose particularly clear and explicit formal models. For instance, Siskind (1996) presents an efficient algorithm for keeping track of just the necessary and possible components of word meaning hypotheses consistent with a set of examples. Most research on word learning does not work with such precise formal models, so it is not always so easy to identify the inference framework guiding the research. Whenever researchers speak of some process of "eliminating" or "ruling out" hypotheses about word meaning, as Pinker (1989) does, or of tracking some minimal set of necessary and sufficient meaning components, as Siskind (1996) does, we take them to be appealing to some kind of eliminative or deductive model, at least implicitly. This way of thinking about word learning serves as the foundation for many substantive proposals about children bring prior knowledge to bear on the inference problem (e.g., Carey, 1978; Clark, 1987; Markman, 1989).

The main alternatives to hypothesis elimination are based on some form of associative learning, such as connectionist networks (Colunga & Smith, 2005; Gasser & Smith, 1998; Regier, 1996, 2005; Smith, 2000) or similarity-matching to examples (Landau, Smith & Jones, 1988; Roy and Pentland, 2004).[1] 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 the simplest form of direct percept-word associations. The boundary between hypothesis-elimination approaches and associative-learning approaches is not always starkly clear. For

instance, Siskind (1996) keeps track of the frequencies with which specific words and world contexts are associated, to support rejection of noise and construction of homonymic lexical entries.

While both hypothesis elimination and associative learning models offer certain important insights, we will argue that neither approach provides an adequate framework for explaining how people learn the meanings of words. We will consider the following five core phenomena that have been highlighted in the literature of the last twenty years (e.g., Bloom, 2000; Carey, 1978; Colunga & Smith, 2005; Markman, 1989; Regier, 1996; Siskind, 1996; Tomasello, 2001), and which any model of word learning should account for:

1. Word meanings can be learned from very few examples. Often a reasonable guess can be made from just a single example, while two or three more examples may be sufficient in the right contexts to home in on the meaning with high accuracy.

2. Word meanings can be inferred from only positive examples – examples of what the word refers to. Negative examples – examples of what the word does *not* refer to – may be helpful but are often not necessary to make a reasonable guess at a word's meaning.

3. Word meanings carve up the world in complex ways, such that an entity, action, property or relation can typically be labeled by multiple words. The target of word learning is not simply a single partition of the world into mutually exclusive categories, with one word per category, but rather a system of overlapping concepts each with a distinct linguistic label.

4. Inferences about word meanings from examples may often be graded, with varying degrees of confidence reflecting the level of ambiguity or noise in the learner's experience.

5. Inferences about word meanings can be strongly affected by pragmatic or intentional reasoning about how the observed examples were generated given the relevant communicative context.

We do not mean to suggest that all of these phenomena apply in every case of word learning, only that they are pervasive and of central importance. They illustrate some of the severe challenges that word learning poses as a computational problem to be solved, as well as some of the powerful inferential capacities that children must be able to bring to bear in its solution. A satisfying framework for modeling word learning should thus present natural explanations for these phenomena. As we explain below,

traditional approaches based on hypothesis elimination or associative learning do not do so in general; at best each approach captures only a subset of these phenomena.

The main goal of this paper is to propose a new approach to understanding word learning based on principles of rational statistical inference. Our framework combines some of the principal advantages of both deductive and associative frameworks, while going beyond some of their major limitations. Our key innovation is the use of a Bayesian inference framework. Hypotheses about word meanings are evaluated by the machinery of Bayesian probability theory rather than deductive logic: hypotheses are not simply ruled in or out, but scored according to their probability of being correct. The interaction of Bayesian inference principles with appropriately structured hypothesis spaces can explain the core phenomena listed above. Learners can rationally infer the meanings of words that label multiple overlapping concepts, from just a few positive examples. Inferences from more ambiguous patterns of data lead to more graded and uncertain patterns of generalization. Pragmatic inferences based on communicative context affect generalizations about word meanings by changing the learner's probabilistic models.

The plan of the paper is as follows. We begin by pointing out some of the specific difficulties faced by the standard approaches to word learning. The core of the paper develops our Bayesian framework in the context of a particular case study: learning common nouns for object categories, such as "animal", "dog", or "terrier". We present the main ingredients of a computational model based on the principles of our Bayesian framework, providing an explanation for the key phenomena in learning overlapping extensions. The predictions of this model are then tested in three experiments, with both adult and child learners, demonstrating the importance of these inferences in an ostensive word learning context. These inferences may provide one means by which people can acquire words for concepts at multiple levels of an object-kind hierarchy (subordinate, basic, and superordinate) – traditionally considered a critical challenge of early word learning (e.g., Markman, 1989; Waxman, 1990). We then present a more quantitative fit of the model given the data. Finally, we show how the Bayesian framework can potentially address other difficulties faced by standard approaches, and we consider the challenges facing it as a general framework for word learning.

## Evaluating traditional approaches: the case of object-kind labels

Before describing our Bayesian approach and its experimental tests, it will be helpful to give some concrete illustrations of the core phenomena of word learning listed above, and to explain traditional deductive and associative accounts of these phenomena, as well as some of the difficulties facing them. Let us return to our opening question of how a child could infer the meaning of a common noun that labels an object-kind, such as the word "dog". Numerous studies have shown that children can make reasonable guesses about such word meanings from a single labeling event. For instance, Markman and Hutchinson (1984) taught 3-year-olds a new word (e.g., "fep") for a familiar object (e.g., a German shepard) and showed that children preferred to generalize new labels to taxonomically similar objects (e.g., a Poodle) rather than a thematically matching object (e.g., a bone). Markman and Wachtel (1988) found that 3-year-olds interpreted a novel word as referring to a whole object as opposed to a salient part of the object. Landau, Smith, and Jones (1988) showed that two-year-olds preferred to generalize category labels to objects matching in shape rather than texture, size or color. The ability to learn words from one or a few exposures may even be present in children as young as 13 to 18 months (Woodward, Markman, & Fitzsimmons, 1994). These rapid inferences are not restricted to object-kind labels. In the first fast mapping study of Carey and Bartlett (1978), an adult pointed children to two trays, one colored a prototypical blue and the other colored an unusual olive green. The adult then asked, "Bring me the chromium tray, not the blue one, the chromium one." Many of the children made the correct inference that "chromium" referred to the olive green color from only one or two experiences of this sort, and about half of them remembered the word-referent pairing about 5 weeks later. Furthermore, Heibeck and Markman (1987) showed that the ability to use linguistic contrast to infer word meanings applied to other semantic domains such as shape and texture. In sum, the ability to infer important aspects of a word's meaning from just a single positive example – what we have referred to as phenomena 1 and 2 above – seems to be present in children as young as two years of age.

How do children make these inferences about word meanings from such sparse data? One influential proposal within the hypothesis-elimination paradigm 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. Two classic constraints on the meanings of common nouns are 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. Given one example of "dog", the taxonomic assumption would rule out the subsets of all spotted things, all running things, all dogs plus the Lone Ranger's horse, or all dogs except Labradors.

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 (Figure 1). 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 hierarchical taxonomy is interesting in its own right, but more importantly as a special case of a fundamental challenge: the problem of learning with overlapping hypotheses – phenomenon 3 from the list above. 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. Thus a single example of a new word will typically fall under multiple nameable categories and thus be insufficient to fix the reference class of the word.

---

INSERT FIGURE 1 ABOUT HERE

---

Another example of learning with overlapping hypotheses arises in contexts where multiple dimensions of an object, such as its shape and material composition, might be relevant simultaneously. Consider the words that might apply to objects found in a furniture store: kinds of objects such as "table",

"chair", "shelf", and "vase", together with kinds of solid substances such as "wood", "metal", "plastic", and "stone" (Figure 2). Kinds of objects tend to have a fairly reliable perceptual correlate of shared shape, and kinds of substances refer to the material an object is made of. Both adults and children are capable of learning word meanings with this orthogonal pattern overlap from a small number of examples (Mintz & Gleitman, 2002; Akhtar, Jipson, & Callanan, 2001). Later in the paper we will discuss this case in more detail.

---

INSERT FIGURE 2 ABOUT HERE

---

Markman (1989) suggested one solution for dealing with overlapping hypotheses in the case of object categories: people may assume that new common nouns map not to just any level in a taxonomy, but preferentially to a basic level of categorization (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Basic-level categories, such as the category of dogs, are clusters of intermediate size that maximize many different indices of category utility (relative to smaller subordinate categories, e.g., Dalmatians, or larger superordinate categories that contain them, e.g., animals). Whether children really have a preference to map words onto basic-level kinds is controversial (Callanan, Repp, McCarthy & Latzke, 1994), but if this preference does exist, it is clear how it would enable names for basic-level categories to be learned after seeing just a single typical labeled example. In the context of learning words for kinds of objects and kinds of substances, Landau, Smith, and Jones (1988) made an analogous suggestion that children preferentially map words onto shape-based object categories, at least for simple or regular shapes (Imai & Gentner, 1997; Soja, Carey, & Spelke, 1991).

Assuming biases to map words onto basic-level object categories, together with other constraints mentioned above, appears to explain how a hypothesis-elimination learner could learn word meanings from just a single positive example, because each object now belongs to just one nameable category. But this solution only works for basic-level object labels like "dog", and in fact is counterproductive for all other kinds of words. How do we learn all the other words we know, for categories at superordinate or

subordinate levels, for substance concepts, and everything else? Admitting some kind of soft 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. In one sense, that is our goal in this paper.

Are the prospects any better for associative learning accounts, in trying to explain word learning from just one or a few positive examples? On the surface, associative learning would not seem well-suited to explaining any kind of rapid inference, because it is typically conceived of as a gradual process of accumulating associative strength over many experiences. Indeed, some classic associative models of language acquisition do not show enduring fast mapping (Plunkett, Sinha, Moller, & Strandsby, 1992), because of the potential for catastrophic interference. More recent models have tried to account for rapid inferences about word meaning (e.g., Colunga & Smith, 2005; Regier, 2005), through a combination of exemplar representations and attentional learning.

It is not clear, however, how the associative models can solve the problem of overlapping extensions. One standard mechanism in associative models is the presence of implicit negative evidence: the models implicitly assume that a positive example of one word is a negative example of every other word. This is precisely the issue concerning overlapping extensions. One attempt to address this problem was Regier (1996). He 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 strengthened over thousands of learning trials. However, this model does not address the phenomenon of learning from very few examples. Another class of models (Li & MacWhinney, 1996; MacWhinney, 1998; Merriman, 1999) uses competition among outputs to implement the idea of implicit negative evidence. However, the simple mechanism of competition embodied in these models is not designed to explain how children learn that multiple words can each apply to a single object.

Recent work in associative models of word learning has focused on the idea of tuning attentional biases. For example, a shape bias for object labels could be the result of shifting attention to shape (as opposed to material) over a set of training exemplars (e.g., Colunga & Smith, 2005**).** The model of Regier

(2005) acquires attentional biases in both meaning space and form space, to enable learning a system of form-meaning mappings. But the problem of learning words with overlapping extensions persists: how do learners acquire words at the subordinate or superordinate levels, or words for categories not based on shape, from just a few examples? These models have not tried to address this question, and it is not clear how they could.

We will argue that this essential problem of learning overlapping word meanings from sparse positive examples can be solved by a Bayesian approach to word learning. Relative to more traditional approaches, our approach posits a more powerful statistical-inference framework for combining prior knowledge with the observed examples of a word's referents. We will focus on a set of phenomena in the context of learning words for taxonomic categories, which strongly suggest that some inferential mechanism of this sort is at work. To illustrate with the ostensive learning problem introduced earlier, after observing Max the Dalmatian labeled a "fep", a learner guided by a taxonomic hypothesis space and perhaps some preference for labeling basic-level categories might reasonably guess that "fep" refers to all dogs. Now suppose that the learner observes 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. This intuition can be captured by a Bayesian inference mechanism that scores alternative hypotheses about a word's meaning according to how well they predict the observed data, as well as how they fit with the learner's prior expectations about natural meanings. An intuitive sensitivity to these sorts of suspicious coincidences is a core capacity enabling rapid word learning, and we will argue it is best explained within a Bayesian framework.

Several previous studies have shown that multiple examples help children learn subordinate or superordinate kind labels (Callanan, 1985, 1989; Liu, Golinkoff & Sak, 2001; Waxman, 1990), or adjectives (Mintz & Gleitman, 2002; Akhtar, Jipson, & Callanan, 2001). For instance, showing a dog, a horse, and a cow as examples of "animals" provides better evidence than just showing a "cow"; showing several differently shaped objects with a characteristic texture, as examples of a word for that texture, provides better evidence than just showing a single object. Intuitively, the additional examples help in these cases by ruling out compelling alternative hypotheses, and a formal account of this "cross-situational learning" was a key part of Siskind's (1996) hypothesis-elimination model of word learning. However, Siskind's hypothesis-elimintation approach cannot explain the phenomenon of learning from a suspicious coincidence, because no hypotheses are eliminated by the additional examples. This phenomenon also poses a challenge to associative learning approaches to word learning. Given one Dalmatian labeled three times as "a fep" and three Dalmatians labeled "a fep" once for each, the correlation between the appearance of *dog* features and the word "fep" is exactly the same as the correlation between the appearance of *Dalmatian* features and the word "fep": 100% in both cases. Associative models that attempt to infer word meaning from correlations between perceptual feature clusters and labeling events thus also get no inductive leverage from this coincidence.

To see further how the effects of multiple examples reveal the inductive logic behind word learning, consider how a learner's beliefs about the meaning of "fep" might have changed had the first four examples been a Labrador, a Golden Retriever, a Poodle, and a Basset Hound – rather than three Dalmatians. Presumably the learner would become more confident that "fep" in fact refers to all and only dogs, relative to our initial belief given just the single example of one Dalmatian called a "fep". That is, the inference to a basic-level meaning is qualitatively similar given either one example or four examples from different subordinate classes, but becomes more confident in the latter case. This shift in confidence suggests that the initial inference after one example was not simply due to the application of a defeasible constraint ruling out all but the basic-level hypothesis of dogs; if so, then the additional examples would tell us nothing. A more plausible interpretation might be to say that given the first example, there was still some probability that the

word mapped only onto the subordinate category of Dalmatians. Subsequent evidence weighs against that overly specific hypothesis and shifts the corresponding weight of belief onto the basic-level hypothesis of all and only dogs. In the experiments described below, we show that both children and adults behave in agreement with this picture: they increase their tendency to generalize at the basic level given multiple examples spanning a basic-level hypothesis, relative to their base preference given just a single example. The explanation for this behavior, as with the restriction of generalization given three examples spanning a subordinate category discussed above, is that inferences to word meaning are not based purely on hypothesis elimination subject to hard and defeasible constraints. Rather they reflect some kind of statistical inference that may become sharper or more focused as additional consistent data are observed.

We will also show how these patterns of inference based on suspicious coincidence can be captured in a statistical framework based on Bayesian inference. In contrast with hypothesis elimination approaches, hypotheses are not just ruled in or out. Instead, the probability of each alternative hypothesis is evaluated. In contrast with associative learning approaches, the statistical information does not just come from correlations between words and referents. The inference mechanism is sensitive to how examples are generated and may disregard "outliers" or uninformative examples. The Bayesian framework can thus explain the more general graded character of generalization in word learning – Phenonemon 4 above – which causes difficulties for hypothesis elimination approaches in particular. A learner who has seen just a single example will typically be less confident in generalizing the word to new instances than a learner who has seen many consistent examples. Anomalous examples may be discounted as "outliers" rather than given full weight in revising learners' hypotheses (Jaswal & Markman, 2001).

One last phenomenon of word learning has been particularly central in recent studies: the role of intentional and pragmatic inferences – Phenomenon 5 above. These phenomena pose a challenge for all approaches to word learning, but particularly so for the associative tradition. As even advocates of this tradition have suggested (Regier, 2003), it is likely that some mechanisms beyond simple associative learning will be necessary to account for the social factors at work in word learning. For example, Baldwin and colleagues (Baldwin, 1991, 1993; Baldwin, Markman, Bill, Desjardins & Irwin, 1996) showed that by

18 months of age, children use speaker's eye gaze and joint attention to infer which object the speaker is referring to. If the speaker looks into a bucket and says "Look, a fep!" while the child is looking at another object, she would track the speaker's gaze and interpret the word as referring to the object inside the bucket. Or consider a study by Tomasello and Barton (1994). An adult said to children, "Let's glip the doll," then executed one action, followed by the exclamation "Oops!" and a second action, followed by the exclamation, "There!" Children correctly inferred that "glipping" referred to the second action using the emotional expression of the experimenter. Furthermore, some word learning constraints such as mutual exclusivity (Markman, 1989) have been reinterpreted as pragmatic constraints (Diesendruck & Markson, 2001). In other words, an example of a new word is not just a raw data point to be entered blindly into a matrix of word-object co-occurrences, but a potentially rich communicative experience to be explained by an inference to the word's most likely meaning. Although it is somewhat controversial whether it is in fact intentional reasoning, or more basic attentional cuing (e.g., Smith, 2000), that guides children solving these word learning problems, most agree that deciphering speaker's communicative intent is an important component of word learning. It is clear how these children's inferences could be cast naturally in the framework of hypothesis elimination, based on chains of pragmatic deductions about what the adult intended to refer to, but not so clear under an associative learning framework. Our Bayesian framework can address at least some of these inferences in terms of sensitivity to the sampling process, and we will give an example of this capacity later on in the paper.

In sum, traditional approaches based on hypothesis elimination or associative learning can account for some but not all of the five critical aspects of word learning we identified. In contrast, the Bayesian framework we propose can potentially handle all five phenomena. The rest of the paper will lay out the model in more detail, provide empirical evidence from both adults and 4-year-old children in learning words for different levels of a taxonomic hierarchy, and discuss some extensions and implications for our framework.

## The Bayesian framework

Our model is formulated within the Bayesian framework for concept learning and generalization introduced by Tenenbaum and his colleagues (1999; Tenenbaum & Griffiths, 2001; Tenenbaum, Griffiths & Kemp,2006). This framework aims to explain inductive learning at the level of "computational theory" (Marr, 1982) or "rational analysis" (Anderson, 1990; Oaksford & Chater, 1998) – to understand in functional terms the implicit knowledge and inferential machinery guides people in generalizing from examples – rather than to describe precisely the psychological processes involved.

We focus on the restricted problem of learning a single novel word *C* from a few examples, but as we discuss later on, the framework in principle extends to the more general problem of learning a whole lexicon from a large corpus of experience. Let $X = x^{(1)}, \ldots, x^{(n)}$ denote a set of $n$ observed examples of the novel word *C*. The examples are drawn from some known domain of entities *U*. We assume that the learner has access to a hypothesis space *H* of possible concepts and a probabilistic model relating hypotheses $h \in H$ to data *X*. Each hypothesis *h* can be thought of as a pointer to some subset of entities in the domain that is a candidate extension for *C*. We assume that the learner can identify the extension of each hypothesis (which entities fall under it). More generally, hypotheses could represent candidate intensions, but here we make the simplifying assumption that each intension yields a unique extension (a version of Clark's (1987) contrast principle), and we focus on how learners infer a word's extension.

Given the examples *X*, the Bayesian learner evaluates all hypotheses for candidate word meanings according to Bayes' Rule, by computing their *posterior probabilities* $p(h|X)$, proportional to the product of *prior probabilities* $p(h)$ and *likelihoods* $p(X|h)$:

$$p(h \mid X) = \frac{p(X \mid h)\,p(h)}{p(X)} \qquad (1)$$

$$= \frac{p(X \mid h)\,p(h)}{\sum_{h' \in H} p(X \mid h')\,p(h')} \quad . \qquad (2)$$

The prior $p(h)$, including the hypothesis space itself, embodies the learner's expectations about plausible meanings for the word *C*, independent of the examples *X* that have been observed. Priors may reflect

conceptual or lexical constraints, expectations about how different kinds of words are used in different contexts, or beliefs conditional on the meanings of other previously learned words. They may be innate or acquired. A taxonomic constraint or basic-level bias can be incorporated naturally through this term.

The likelihood $p(X|h)$ captures the statistical information inherent in the examples $X$. It reflects expectations about which entities are likely to be observed as examples of $C$ given a particular hypothesis $h$ about C's meaning, such as a default assumption that the examples observed will be a representative sample of the concept to be learned. The likelihood may also be sensitive to other data, such as the syntactic context of the examples, or examples of other words (which might contrast with $C$). We consider some of these possibilities later in the paper.

The posterior $p(h|X)$ reflects the learner's degree of belief that $h$ is in fact the true meaning of $C$, given a combination of the observations $X$ with prior knowledge about plausible word meanings. It is proportional to the product of the likelihood and prior for that hypothesis, relative to the corresponding products for all other hypotheses. This form embodies a principle of "conservation of rational belief": if the learner believes strongly in a particular hypothesis $h$ about the meaning of a word to be learned, i.e., she assigns a value near 1 to $p(h|X)$, then she must necessarily believe strongly that other hypotheses do not pick out the true meaning, i.e., she must assign values near 0 to $p(h'|X)$ for all other $h' \neq h$.

All of these probabilities – priors, likelihoods, and posteriors – are implicitly conditioned on a knowledge base, which could include the meanings of previously learned words, or abstract principles about possible word meanings, how words tend to be used, or how examples are typically generated. Later in the paper we consider more general analyses in which the likelihoods or prior probabilities change to incorporate different aspects of a learner's background knowledge.

The main work of the model is done in specifying the likelihoods and priors that enter into Bayes' rule. Before considering these components further, we note one further piece of machinery that is needed to relate the learner's beliefs about word meaning encoded in $p(h|X)$ to generalization behavior. The learner needs some way to decide whether or not any given new object $y$ belongs to the extension of $C$, given the observations $X$. If the learner is completely sure of the word's meaning – that is, if $p(h|X) = 1$ for exactly

one $h = h^*$ and 0 for all other $h$ – then generalization is trivial: $C$ applies to all and only those new objects $y \in h^*$. More generally, the learner must compute a probability of generalization, $p(y \in C|X)$, by averaging the predictions of all hypotheses weighted by their posterior probabilities $p(h|X)$:

$$p(y \in C \mid X) = \sum_{h \in H} p(y \in C \mid h)\, p(h \mid X) \qquad (3)$$

To evaluate Equation 3, note that $p(y \in C|h)$ is simply 1 if $y \in h$, and 0 otherwise, and $p(h|X) = 0$ unless the examples $X$ are all contained within $h$. Thus the generalization probability can also be written as

$$p(y \in C \mid X) = \sum_{h \supset y, X} p(h \mid X) \qquad (4)$$

or the sum of the posterior probabilities of all hypotheses that contain both the new object $y$ and the old examples $X$. Following Tenenbaum & Griffiths (2001), if we interpret hypotheses $h$ as features or feature bundles (that might define the intensions of the hypotheses), and the posterior probabilities $p(h|X)$ as feature weights, then Equation 4 captures the intuition that generalization from $X$ to $y$ will increase in proportion to the number or weight of features in common between $X$ and $y$ – as in classic models of similarity judgment by Tversky (1977) or Shepard and Arabie (1979). Yet because each hypothesis sharply picks out a subset of entities, Equation 4 can also produce essentially all-or-none, rule-like generalization if the posterior probability concentrates its mass on a single hypothesis.

*The hypothesis space*

Most generally, the hypothesis space $H$ is simply a set of hypotheses about the meaning of the novel word $C$. Each hypothesis $h$ points to a subset of entities in the domain $U$ that is a candidate for the extension of $C$. For the purposes of Bayesian inference, these hypotheses need not be structured or related to each other in any particular way. They may be simply a set of mutually exclusive and exhaustive candidate word extensions, carrying the assumption that the word to be learned maps onto one and only one of these subsets of the world. However, there are strong theoretical reasons – as well as practical motives – why we should

typically assume a more structured hypothesis space. Figures 1 and 2 show two examples of hypothesis spaces with different large-scale structures: a tree-structured taxonomy of object kinds, in which the hypotheses are nested (Figure 1), and an orthogonal "two-dimensional" matrix of object and substance categories, in which any two hypotheses from different dimensions overlap (Figure 2). As explained below, a structured hypothesis space can be thought of as an important component of the learner's prior, perhaps the most important component that supports successful learning from few examples. It is also the place where many candidate word-learning principles enter into the analysis. Practically speaking, assuming an appropriately structured hypothesis space can allow that space to be constructed in a fairly automatic fashion, based on independent behavioral data we collect from participants. Assuming no structure to the hypothesis space can force modelers to specify every hypothesis and its associated prior probability by hand (e.g., Heit, 1998), leading to a proliferation of free parameters in the model.

In our model of learning object-kind labels, we assume that the hypothesis space corresponds to a taxonomy of nested categories, which can be constructed automatically by hierarchical clustering ("average linkage"; Duda & Hart, 1973) on human participants' similarity ratings (see Figure 7). Each hypothesis corresponds to one cluster in this tree. We should emphasize that this intuitive taxonomy is intended only as a simple but tractable first approximation to the hypothesis space people could adopt for learning common object labels; it is not intended to be the only source of object-label hypotheses, nor to represent the structure of hypothesis spaces for learning other kinds of words.

*Probabilistic components of the model*

Two kinds of probabilities, prior probabilities and likelihoods, are defined over our hypothesis space of candidate word meanings. Here we describe the general character of these probabilities, saving the details of how they are computed in applying our model for the section on model evaluation (following the experimental sections).

*Likelihoods*. The likelihood function comes from assuming that the observed positive examples are sampled at random (and independently) from the true concept to be learned. Consider a hypothesis about the word's extension that picks out a finite set of $K$ objects. Then the likelihood of picking any one object at

random from this set of size $K$ would be $1/K$, and for $n$ objects (sampled with replacement), $1/K^n$. This reasoning leads to the following likelihood function:

$$p(X \mid h) = \left[\frac{1}{\text{size}(h)}\right]^n \quad (5)$$

if $x_i \in h$ for all $i$, and 0 otherwise. We refer to Equation 5 as the *size principle* for scoring hypotheses: hypotheses with smaller extensions 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 "fep", either all Dalmatians or all dogs seem to be fairly plausible hypotheses for the word's extension, but given three Dalmatians as the first three examples of "fep", the word seems much more likely to refer only to Dalmatians than to all dogs – because the likelihood ratio of these two hypotheses is now inversely proportional to the ratio of their sizes, *raised to the fourth power*. The size principle thus explains why a learner who observes four examples of "fep" that all happen to be Dalmatians will tend to infer that the word refers only to Dalmatians, rather than all dogs, even though both hypotheses are logically consistent with the examples encountered. Intuitively, this inference is based on noticing a suspicious coincidence; formally, it is justified on the grounds of maximizing the likelihood of the data given the hypothesis.

This proposal addresses a crucial shortcoming of traditional deductive or hypothesis-elimination approaches to word learning, which cannot explain how inferences may change without encountering falsifying examples. It also addresses an analogous shortcoming of associative approaches, which cannot explain why one feature may be preferred over another as the basis for a word's meaning, even though both features are equally correlated with the observed usage of the word. The rationality of the size principle depends on how widely applicable is the assumption of randomly sampled examples, and how defeasible it is when the learner is confronted with examples sampled in importantly different ways. The size principle can be viewed as a softer statistical version of the subset principle (Wexler & Cullicover, 1980; Berwick,

1986), a classic deductive approach to learning from positive examples in formal models of language acquisition. We discuss the connection between Bayesian learning and the subset principle in more detail later, when we compare alternative models with our experimental data.

*Priors.* Most generally, the prior should reflect all of people's implicit knowledge about how words map onto meanings and how meanings tend to be used in different contexts. Perhaps the most important component of the prior is simply the qualitative structure of the hypothesis space: here, the assumption that hypotheses correspond to nodes in a tree-structured taxonomy. This assumption is equivalent to assigning zero prior probability to the vast majority of logically possible hypotheses – all other subsets of objects in the world – that do not conform to this particular taxonomy.

Although a tree-structured hypothesis space is not necessary for our Bayesian approach, a rational statistical learner can only make interesting generalizations by adopting some bias that assigns zero or near-zero prior probability to most logically possible hypotheses. To see why, consider a Bayesian learner who assigned equal priors to all logically possible hypotheses – all subsets of entities in the domain. Then, under the size principle in the likelihood function, the best hypothesis for any set of examples would always be the one containing just those objects and no others – a hypothesis that calls for no generalization at all! Generalization in word learning, or any kind of inductive learning, is only possible with a prior that concentrates most of its mass on a relatively small number of hypotheses.

More fine-grained quantitative differences in prior probability will be necessary to explain the particular patterns of generalization that people make, as well as the different patterns shown by different groups of learners, such as adults versus children, or experts versus novices. One important kind of graded prior knowledge in word learning may be a preference for labeling distinctive clusters: more distinctive clusters are *a priori* more likely to have distinguishing names. In learning common nouns, a paramount goal is to acquire linguistic handles for natural kind categories. The perceptual distinctiveness of a cluster is a ready (if not infallible) indicator of how likely that cluster is to correspond to a natural kind. But distinctiveness (perceptual or conceptual) may also be important in its own right, as the utility and stability

of any word will depend in part on how easily speakers can pick out entities in its extension. Thus we expect that some kind of preference for distinctiveness will be a general aspect of prior probabilities in word learning.

*Summary of the basic modeling framework.* While both priors and likelihoods can be understood on their own terms, it is only in combination that they explain how people can successfully learn the reference of new words from just a few positive examples. Successful word learning requires both a constrained space of candidate hypotheses – provided by the prior – and the ability to re-weight hypotheses based on how well they explain a set of observed examples – provided by the likelihood. Without the constraints imposed by the prior, no meaningful generalizations would be possible. Without the likelihood, nothing could be learned from multiple examples beyond simply eliminating inconsistent hypotheses. In particular, priors and likelihoods each contribute directly to the main pattern of generalization that we described in the introduction and that we will look for in our experiments: given just a single example of a novel kind label, generalization to other objects should be graded, but given several examples, learners should apply the word more discriminatingly, generalizing to all and only members of the most specific natural concept that spans the observed examples. The prior determines which concepts count as "natural", while the likelihood generates the specificity preference and determines how the strength of that preference – and thus the sharpness of generalization – increases as a function of the number of examples.

The need for strong prior knowledge to constrain word learning has been a major theme of previous research in the rationalist tradition (Markman, 1989; Pinker, 1989; Bloom, 2000). The importance of statistical learning across multiple examples of word-object pairings has been stressed in associative learning approaches (e.g., Colunga & Smith, 2005; Regier, 2005). Our thesis here is that successful word learning depends on both prior knowledge and statistical inference – and critically, on their interaction. We have presented a theoretical framework for understanding how this interaction functions to support rational generalization from a few positive examples. We now turn to a series of empirical studies mapping out how adults and children generalize words from one or a few examples, followed by quantitative comparisons between these judgments and the generalization patterns of our Bayesian model.

## Experiment 1

Experiment 1 tested adults in a word learning situation. The experiment consisted of two phases, a word learning phase and a similarity judgment phase. In the word learning phase, adults were taught novel words (e.g., "This is a blicket.") and were asked to generalize the word to other objects. Two variables were manipulated: the number of examples (1 vs. 3) and the range spanned by the examples (e.g., three green peppers, or three different-colored peppers, or three different kinds of vegetables). In the similarity judgment phase, participants were asked to give similarity ratings for pairs of the same objects used in the word learning phase. Similarity judgments will be used to yield a hypothesis space for subsequent computational modeling.

The predictions are that adults would show graded generalization with one example, and more all-or-none generalizations with three examples. Furthermore, depending on the span of the three examples, adults would generalize to the most specific category that is consistent with the examples.

### *Method*

*Participants*. Participants were 22 students from MIT and Stanford University, participating for pay or course credit. All participants carried out the word learning task and also participated in the similarity judgment phase that followed. All participants were native speakers of English and had normal or corrected-to-normal vision.

*Materials*. The stimuli were digital color photographs of 45 real objects. They were distributed across three different superordinate categories (animals, vegetables, vehicles) and within those, many different basic-level and subordinate-level categories. These stimuli were divided into a training set of 21 stimuli and a test set of 24 stimuli.

---

INSERT FIGURE 3 ABOUT HERE

---

Twelve sets of labeled examples were used as training stimuli during the experiment (Figure 3). The first three sets contain one example each: a Dalmatian, a green pepper, and a yellow truck. The next nine sets contain three examples each: one of the three objects from the single-example sets (e.g., a Dalmatian), along with two new objects that match the first at either the subordinate-level (e.g., two other Dalmatians in different postures), basic-level (e.g., a terrier and a mutt), or superordinate-level (e.g., a pelican and a pig). Thus the nine sets arise from the combination of the three objects in the one-example set crossed with three levels of matching specificity.

---

INSERT FIGURE 4 ABOUT HERE

---

The 24 objects in the test set are shown in Figure 4. The objects were distributed across all three superordinate level categories (8 animals, 8 vegetables, and 8 vehicles). The set was constructed to provide matches at all levels: subordinate (2 other Dalmatians), basic-level (2 other dogs, a Labrador and a hush-puppy), and superordinate (4 other non-dog animals, a cat, a bee, a seal, and a bear), as well as many non-matching objects (vegetables and vehicles). Note that the test set was exactly the same for all trials, and for any set of exemplars always contains a total of 2 subordinate-level matches (e.g., the other Dalmatians), 4 basic-level matches (e.g., the Dalmatians and the other dogs), 8 superordinate-level matches (e.g., the dogs and the other animals), and 16 non-matching distractors (e.g., all the other objects). We chose to include more basic-level and superordinate-level matches because these categories have more members in the real world, although the actual ratio (1:2:4) is only a rough estimate of the size of the categories.

*Design and Procedure*. The first phase of the experiment was the *word learning* task. Stimuli were presented within a 15" x 15" square window on a color computer monitor, at normal viewing distance. Participants were told that they were helping a puppet (Mr. Frog) who speaks a different language to pick out the objects he wants. On each trial, the participants were shown pictures of either one or three labeled examples of a novel, monosyllabic word (e.g., "fep") and were asked to pick out the other "feps" from the

test set of 24 objects, by clicking on-screen with the computer mouse. The test items were laid out in a four-by-six array, with the order randomly permuted from trial to trial.

The experiment began with participants being shown all 24 test objects, one at a time for several seconds each, to familiarize them with the stimuli. This familiarization was followed by the instructions and 12 experimental trials. (Some subjects were then given an additional set of trials which are not reported here.) On the first three trials, participants saw only one example of each new word, e.g., "Here is a fep." On the next nine trials, they saw three examples of each new word, e.g., "Here are three feps." Within each set of trials, the example sets appeared in a pseudo-random order, counterbalancing content domain (animal, vegetable, and vehicle) and specificity (subordinate, basic, superordinate) across participants. On each trial, the participants were asked to choose the other objects that the word applied to (e.g., the other feps) and their responses were recorded. This phase last approximately 15 minutes in total.

The second phase of the experiment was a *similarity judgment* task. Participants were shown pictures of pairs of objects from the word learning study and asked to rate the similarity of the two objects on a scale of 1 (not similar at all) to 9 (extremely similar). 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 word learning phase. This instruction, along with the placement of the similarity judgment task after the word learning task, was adopted in the hope of maximizing the information that similarity judgments would provide about the hypothesis space that participants used in word learning. Similarity judgments took approximately 45 minutes to collect. Judgments were collected for all pairs of 39 out of 45 objects – 13 from each domain of animals, vegetables, and vehicles – including all test objects and all but 6 of the training objects (which were omitted to save time). The six omitted objects (two green peppers, two yellow trucks, and two Dalmatians) were each practically identical to three of the 39 included objects, and each was treated as identical to one of those 39 in constructing the model of learning reported below. Each participant rated the similarity of all pairs of animals, vegetables, and vehicles (78 x 3 judgments), along with one-third of all possible cross-superordinate pairs (e.g., animal-vegetable, vegetable-vehicle, etc.) chosen pseudo-randomly (169 judgments), for a total of 403 judgments per participant. The order of trials and the order of stimuli were

randomized across participants. These trials were preceded by 30 practice trials (chosen randomly from the same stimuli), during which participants were familiarized with the range of similarities they would encounter and were encouraged to develop a consistent way of using the 1-9 rating scale. They were also encouraged to use the entire 1-9 scale and to spread their judgments out evenly across the scale. The ratings were recorded and the average rating for each pair of objects was computed.

*Results*

The main results of Experiment 1 are shown in Figure 5. Adults clearly differentiated the one-example and the three-example trials, and they were sensitive to the span of the three examples. With one example, adults showed graded generalization from subordinate to basic-level to superordinate matches. These generalization gradients dropped off more steeply at the basic level, with a soft threshold: most test items from the same basic-level category were chosen but relatively few superordinate matches were chosen. With three examples, adults' generalizations sharpened up into a much more all-or-none pattern. Generalizations from three examples were almost always restricted to the most specific level that was consistent with the examples: for instance, given three Dalmatians as examples of "feps", adults generalized only to other Dalmatians; given three different dogs (or three different animals) adults generalized to all and only the other dogs (or other animals).

---

INSERT FIGURE 5 ABOUT HERE

---

With the above overview in mind, we turn to statistical analyses that quantify these effects. Later we present a formal computational model of this word learning task and compare it with the data from this experiment in more quantitative detail. All analyses in this section were based on one-tailed *t*-tests with planned comparisons based on the model's predictions. Data were collapsed over the three different superordinate categories, and over the different test items within a given level of generalization (subordinate, basic, and superordinate). For each of the four kinds of example sets (1, 3 subordinate, 3

basic-level, 3 superordinate) and each of the three levels of generalization, each participant received a set of percent scores measuring how often they chose test items at that level of generalization given that kind of example set. The means of these scores across subjects are shown in Figure 5. Because participants almost never (less than 0.1% of the time) chose any distractors (test items outside of the examples' superordinate category), subsequent analyses did not include these scores.

Two questions were addressed with planned *t*-tests. First, did participants generalize further 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 adults show a significant threshold in generalization 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 participants restrict 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, and superordinate levels. On the one-example trials, participants chose more subordinate (96%) and basic-level matches (76%) than superordinate matches (9%); the difference between the first two levels, 20%, is much less than the difference between the latter two levels, 69% ($t$ (73) = -10.7869, $p < .0001$). In contrast, when presented with three very similar exemplars from the same subordinate category, participants chose more subordinate matches (95%) than either basic-level (16%) or superordinate matches (1%) ($p < .0001$, for both comparisons). Similar comparisons were made between 1 example and 3 basic-level or 3 superordinate level examples. When presented with 3 examples from the same basic-level category, participants generalized even more to the basic-level than the one-example trials (76% vs. 91%). When presented with 3 examples from the same superordinate category, participants generalized to almost all exemplars from the superordinate category (87%).

As our model predicts (see below), given three examples spanning a single subordinate-level category, the generalization gradient should relate to the one-example trials as follows: equal generalization at the subordinate level, and a large decrease in generalization at the basic level, and a small decrease at the

superordinate level. Given three examples spanning a basic-level category, the generalization gradient should be modified as follows: equal generalization at the subordinate level, an increase in basic-level generalization, and a small decrease in superordinate-level generalization. Given three examples spanning a superordinate-level category, the generalization function should be modified as follows: equal generalization at the subordinate-level, and increases in both basic-level and superordinate-level generalization. All of these predictions follow from the model's general tendency to be relatively uncertain about the correct level of generalization when given one example, and to be more certain when given three examples. A set of two-tailed *t*-tests were conducted to test these predictions by comparing the mean percentages of all relevant pairs of conditions in adults' generalizations. The results of all tests were consistent with the model predictions, except for a nonsignificant difference in superordinate generalization between the 1 example and 3 basic-level examples conditions. This difference was in the predicted direction, and it was also predicted to be small, so a non-significant result is not surprising.

To investigate the second question, we tested a series of specific predictions from our model (see below), about how generalization given three examples at a certain level of specificity should differ from each other. 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 (95% vs. 16%, $p < .0001$). Given 3 examples from the same basic-level category, the model predicts a sharp drop between basic-level generalization and superordinate-level generalization (91% vs. 4%, $p < .0001$). Given 3 examples from the same superordinate category, the model predicts that generalization should include all exemplars from that superordinate category (94%, 91%, and 87%, n.s.).

The similarity data are analyzed later in the paper, when we describe the fits of our Bayesian learning model. The similarities will be used to construct the model's hypothesis space.

*Discussion*

Adults clearly generalized differently on the one-example and the three-example trials. With one example, they showed graded generalization from subordinate to basic-level to superordinate matches. In

addition, adults showed a basic-level bias: they generalized to all the other exemplars from the same basic-level category but much less to the superordinate category. With three examples, adults' generalizations were more all-or-none. They restricted their generalizations to the most specific level that is consistent with the examples.

## Experiment 2

Experiment 2 investigated how 3- and 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 Experiment 1, 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 3- and 4-year-old children (mean age 4 years 1 month, ranged from 3 years 6 months to 5 years 0 months; approximately half girls/boys). All participants were recruited from the Greater Boston area by mail and subsequent phone calls. Most children came from a middle-class non-Hispanic white background with about 10% of Asian, African-American, and Hispanic infants. The children received a token gift (i.e., a sticker) after the study. Five children were excluded because of unwillingness to play the game with the experimenter. English was the primary language spoken at home for all children.

*Materials.* The stimuli were the same 45 objects as in Experiment 1, but the children were presented with the real toy objects as opposed to photographs.

*Design and Procedure*. Each child was randomly assigned to one of two conditions: the One-Example condition or the Three-Example condition. Children in the One-Example condition always saw one example of each word, while children in the Three-Example condition always saw three examples of each word. Each child participated in a total of three trials, one from each of the three superordinate categories. On each trial in the Three-Example condition, the examples spanned a different level of generality (subordinate, basic-level, or superordinate).

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 experiment began with a dialog as follows. The experimenter held out 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."] "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 "Okay."] 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 picked out one toy, the experimenter reminded him/her, "Remember Mr. Frog wants all the blicks. Are there more blicks?" If a child picked 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 participated in three trials, each with an example drawn from one of the three superordinate categories: a Dalmatian (animal), a green pepper (vegetable), or 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 the following important difference. The experimenter first 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,” or “Look, this is a fep.” Three factors – the superordinate category (animal, vegetable, and vehicle), the range spanned by the examples (subordinate, basic, superordinate), and the novel word used (“blick”, “fep”, and “dax”) – were crossed pseudorandomly and counterbalanced across participants. Each level of each factor appeared equally often in the first, second, or third trial of the experiment.

*Results*

The patterns of generalization found were qualitatively similar to those found with adults in Experiment 1, and the quantitative analyses followed essentially the same logic. Analyses were based on one-tailed *t*-tests with planned comparisons. We collapsed across superordinate categories, novel words, and trial orders. For each type of example set children were shown, they received a set of percent scores measuring how often they chose test items at each of three levels of generalization (subordinate, basic, superordinate). The means of these scores across subjects are shown in Figure 6a. Children in the One-Example condition each received just a single set of scores, because their three trials all featured the same kind of example set. Children in the Three-Example condition each received three sets of scores, one for each trial, because each trial featured a different kind of example set (three examples clustering at the subordinate, basic or superordinate level). Because no child chose any distractors, subsequent analyses did not include these scores.

---

INSERT FIGURE 6 ABOUT HERE

---

The same two questions as in Experiment 1 were addressed here with planned *t*-tests. First, did children generalize differently in the 1-example trials compared with the 3-example trials in each case? Importantly, did they generalize differently given 1 vs. 3 virtually identical exemplars? More specifically, did children show a significant threshold in generalization 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 restrict 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%) and basic-level matches (31%) than superordinate matches (3%) ($p < .0001$, for both comparisons). In contrast, when presented with three very similar exemplars from the same subordinate category, participants chose more subordinate matches (83%) than either basic-level (13%) or superordinate matches (3%) ($p < .0001$, for both comparisons). Similar comparisons were made between 1 example and 3 basic-level or 3 superordinate-level examples. When presented with 3 examples from the same basic-level category, participants did not generalize more to the basic-level than the one-example trials (31% vs. 47%, n.s.). When presented with 3 examples from the same superordinate category, participants generalized more to both the basic-level and the superordinate level (31% vs. 53%, $p < .001$; 3% vs. 43%, $p < .0001$).

To investigate the second question, we tested a series of predictions based on our model as in Experiment 1. 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.

*Discussion*

Three- and 4-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. 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. Also, not all subordinate classes seemed to be equally salient or interesting to the children. The Dalmatians, as a class, seemed to be unusually interesting to some children, perhaps because of their distinctive coloration or the fact that they came from a line of toys based on a currently popular children's animated feature film. The green peppers, as a class, seemed not very salient to some children, perhaps because they differed from other members of the same basic-level class only in their color (and their coloration was not nearly as striking as the Dalmatians).

In the next experiment, we present children with each of 10 objects and ask for a yes/no response. This modification will ensure that all children provide us with judgement on each of the test objects. We also changed two of the subordinate classes slightly, in order to equalize salience of the subordinates across the animals, vegetables, and vehicles.

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 independent examples labeled or simply the number of labeling events (i.e., word-object pairings). This is particularly important because some associative models (e.g., Regier, 2005; Colunga & Smith, 2005) have suggested that children's word learning is built on keeping track of co-occurrences between words and object percepts. To distinguish our Bayesian approach from conventional associative approaches, it is important to tease apart these possibilities. In the next study, we equate the number of labeling events between the One-Example and Three-Example conditions by labeling the single example object in the One-Example condition three times, while each example object in the Three-Example condition is labeled just once.

## Experiment 3

This experiment sought to replicate and extend the results of Experiment 2 with slight modifications to the stimuli and two important methodological changes. First, we equated the number of labeling events in the One-Example and the Three-Example conditions. Second, instead of letting children choose among the 24 target objects, the experimenter chose 10 of these objects and asked for the child's judgment in each case.

*Method*

*Participants*. Participants were thirty-six 3- and 4-year-old children (mean age 4 years 0 months, ranged from 3 years 6 months to 5 years 0 months), approximately evenly divided by gender. Participants were recruited as in Experiment 2.

*Materials*. The stimuli were the same 45 objects as in Experiment 2, except that the Dalmatians were replaced by terriers, and the green peppers were replaced by chili peppers. Members of each subordinate class were now distinguished from other objects in the same basic-level class by both shape and color features of moderate salience.

*Design and Procedure*. The procedure was identical to that of Experiment 2, except for the following. In the One-Example Condition, each object was labeled three times. For example, the experimenter would pick out a green pepper, show it to the child, and say, "See? A fep." She would then put the pepper down on the floor and pick it up again, saying, "Look, a fep." She would it put down again and pick it up a third time, saying, "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 test objects were laid out in front of the child, the experimenter chose 10 of these objects to ask about. 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 distractors.

*Results*

The main results of Experiment 3 are shown in Figure 6b. A significance level of 0.05 was used in all statistical analyses. Preliminary analyses found no effects of sex, the order of domain, and the order of target type. Subsequent analyses collapsed over these variables. Only two children chose any of the distractors in this experiment on one trial, all analyses excluded the distractor scores.

The same two questions as in Experiments 1 and 2 are addressed with planned *t*-tests. First, did children behave differently in the 1-example trials compared with the 3-example trials? Importantly, did they generalize differently 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 the superordinate level. On the one-example trials, participants chose more subordinate (96%) and basic-level matches (40%) 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 either basic-level (6%) and superordinate matches (0%) ($p < .0001$, for both comparisons). Similar comparisons were made between 1 example and 3 basic-level or 3 superordinate level examples. When presented with 3 examples from the same basic-level category, participants generalized more to the basic-level than the one-example trials (75% vs. 40%, $p < .005$). When presented with 3 examples from the same superordinate category, participants generalized more to both the basic-level and the superordinate level (88% vs. 40%, $p < .0001$; 62% vs. 17%, $p < .0001$).

To investigate the second question, we tested a series of predictions based on our model as in Experiments 1 and 2. As can be seen in Figure 6b, 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 generalization 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 generalization 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 a simplified testing procedure, preschool children generalized new words in ways that looked more like the adults in Experiment 1. However, they still showed a much lower tendency for basic-level generalization given a single example, which suggests that adults' strong tendency for one-shot basic-level generalization may reflect a convention acquired through extensive experience with learning and using words. The differences in generalization between the One-Example and Three-Example conditions of Experiment 2 persisted (or became stronger) here, even though the number of labeling events was equated across conditions. This finding suggests that preschool children make statistical inferences about word meanings which are computed over the number of examples labeled, not just the number of word-object pairings.

## Discussion of Experiments

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). We also tested both adults and children. 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 often associated with statistical learning. Rather, there was a strong shift in generalization behavior from one to three examples, reflecting the rational statistical principle that observing the span of three independent, randomly-sampled examples warrants a sharp increase in confidence about which hypothesis for generalization is correct. Both

adult and child learners appear to be sensitive to "suspicious coincidences" in how the examples given for a novel word appear to cluster in a taxonomy of candidate categories to be named.

By varying the span of examples, we found that labels for subordinate and superordinate categories may not be as difficult for children 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 within the superordinate categories of animal, vehicle, and vegetable. Special linguistic 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.

Lastly, we found an interesting difference between adults and preschool children in how likely they were to extend novel words from one example to other instances in the same basic-level category: adults showed much greater basic-level generalization than did children. This is consistent with Callanan et al.'s (1994) finding that children do not show robust basic-level generalization when taught unfamiliar words. Our results are broadly consistent with Callanan et al., in that they suggest a basic-level bias may not part of the foundations for word learning. Rather, such a bias may develop as children learn more about general patterns of word meanings and how words tend to be used. Further research using a broader range of categories in the same experimental paradigm developed here will be necessary to establish a good case for this developmental proposal. If further research does support the notion that a basic-level bias develops through experience, we expect that this development could also be modeled as an instance of Bayesian learning, in which people come to realize that basic-level object labels are used much more frequently than subordinate or superordinate labels.

It is important to note a few caveats. We have argued that in our experiments preschool children learned words for categories at multiple levels of the taxonomic hierarchy – superordinate, basic-level, and subordinate – but it is an open question whether children understand these categories as part of a hierarchically organized system of kinds. Like most previous studies, we did not include an explicit test for the child's understanding of class inclusion relations, which is often taken to be the ultimate test for understanding hierarchical structures. Smith (1979) asked 4- to 6-year-old children inference questions based on class inclusion and found that 4-year-olds showed a fragile but statistically reliable understanding. It is possible that children simply use the span of perceptual similarity as a first approximation for larger versus smaller categories that are akin to a set of nested categories in the mature conceptual system. This alternative possibility assumes that children may have a somewhat different hypothesis space from adults – instead of having a nested set of categories, they may have mapped the words onto regions of perceptual space (e.g., Shepard, 1987; Tenenbaum & Griffiths, 2001), some broad and some narrow.

One indication that children may have had a somewhat different hypothesis space than adults is their pattern of generalization with superordinates. Given three examples that spanned a superordinate-level category, children chose superordinate matches most of the time, and far more often than with other example sets, but still less often than adults (62% of the time in Experiment 3, versus 87% in Experiment 1). There are several possible explanations for this finding, which could be explored in future work. Children may simply have had a different tree-structured hypothesis space than adults – a stable hypothesis space with stable superordinate-level hypotheses that just happen not to include exactly the same objects as adults'. Children could also have less stable hypothesis spaces. There could be more variance across children in the hypothesis spaces they use, or each individual child might not have a single tree-structured hypothesis space so clearly articulated as adult learners might have. Children might also need to acquire deeper knowledge about superordinate categories – for example, integrating their representation of the category *animals* with an intuitive theory of biology, or understanding the functional and social significance of *vehicles* – before these categories can become stable hypotheses for generalizing word meanings.

Another potential concern is that in our experiments we used only relatively familiar categories. It is possible that children had already acquired the superordinate or subordinate level terms and they simply translated those words into our nonsense labels during the experiments. This is unlikely because Waxman (1990) found that only about half of her 4-year-old children knew the superordinate term "animal," and both "vegetable" and "vehicle" are less commonly known to preschoolers. In our sample, the 3-year-olds (who presumably were less likely to know these words) and the 4-year-olds did not behave differently on our task. Some of the subordinate-level concepts we used had an existing label, e.g., *Bassett Hound* or *Dalmatian*, whereas others did not, e.g., *yellow truck* or *green pepper*. Thus it is unlikely that the children simply translated the new words into words they already know. Still, future work using a broader range of categories and novel categories could help to clarify the generality of our findings. Xu and Tenenbaum (in press) and Schmidt and Tenenbaum (unpublished data) have studied word learning with different sets of novel objects, each of which can be classified into a tree-structured hierarchy of object kinds, and found behavior consistent with the Bayesian framework we present here.

Lastly, we stress that when we say these words are not too hard to learn from examples, we are not saying that all aspects of these words are easy to learn. Both in our experiments and our model, we have only addressed word meaning in terms of extension, i.e., which entities the word refers to. Other aspects of word meaning having more to do with the word's intension, such as the essence of the concept labeled by the word, how that concept relates to a domain theory, and how it relates to other concepts, may not be so easily grasped from just a few examples. (See Bloom (2000) for a discussion of the differences between extensions and intensions in word meaning.) In developing models based on statistical inference, it is most natural to begin by focusing on the extension of words, because that is the component of meaning with most directly measurable statistical consequences. However, our framework is not limited to extensions. Other aspects of word meaning also have statistical consequences for how and when a word is likely to be used, and thus in principle could be learned from observations given an appropriate hypothesis space.

If category labels at different levels of the conceptual hierarchy are not very difficult to learn, as we have suggested here, why is it that in young children's early vocabulary we tend to see more basic-level

category labels? This is, after all, a critical observation that motivated the standard picture of how children acquire kind terms at multiple levels of the taxonomy. Several factors may be important in explaining the time lag between acquiring basic-level labels and subordinate and superordinate labels. First, subordinate and superordinate labels may require multiple examples. If each example is labeled on different occasions and spread out in time, children may forget the examples over time. Second, subordinate- and superordinate-level category labels are used much less frequently in adult speech, so the relevant examples are harder to come by. Middle-class American parents tend to point to objects and label them with basic-level terms. Lastly, superordinates are often used to refer to collections (Markman, 1989), so children may be misled by the input in interpreting these words. In our studies, we have presented children with a simplified learning situation in order to uncover the underlying inferential competence that guides them in – but is not exclusively responsible for – real world performance.

## Evaluating a Bayesian model of learning object-kind labels

In this section we assess the quantitative fit between a Bayesian model of word learning and participants' generalization judgments in the kind-label learning experiments just presented. We also consider the predicted generalization patterns of several alternative models, including weaker versions of the full Bayesian approach as well as a number of non-Bayesian models intended to capture the essences of the major hypothesis-elimination and associative-learning approaches.

### *Constructing the hypothesis space*

Based on participants' similarity judgments in Experiment 1, we generated a hierarchical cluster tree to approximate the taxonomy of nested categories (Figure 7). Each internal node of the tree corresponds to a cluster of objects that are on average more similar to each other than to other, nearby objects. The height of each node represents the average pairwise dissimilarity of the objects in the corresponding cluster. 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.

---

INSERT FIGURE 7 ABOUT HERE

---

Each of the main classes underlying the choice of stimuli corresponds to a node in the tree: vegetable (EE), vehicle (HH), animal (JJ), pepper (J), truck (T), dog (R), green pepper (F), yellow truck (G), and Dalmatian (D). 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.[2] Other easily describable nodes include cluster U containing all and only the construction vehicles (tractor, bulldozer, and crane) or cluster II, containing all and only the mammals. The only clusters that do not appear to correspond to conceivably lexicalizable concepts are three defined only by subtle perceptual variation below the subordinate level: A, including two of the three Dalmatians; B, including two of the three green peppers, and E, including two of the three yellow trucks. We take each cluster to correspond to one hypothesis in *H*, with the exception of these three clusters below the subordinate level. In so doing, we are assuming that each learner maintains only a single hypothesis space, and that its structure does not change as new words are learned. We are also assuming that a single tree structure is sufficient to model the hypothesis spaces of all word learners. While these assumptions greatly simplify modeling, none is a fundamental commitment of our theoretical framework, and we expect that they will need to be relaxed in future work.

*Computing numerical values for likelihoods and priors*

For learning common nouns under the taxonomic constraint, the geometry of the cluster tree suggests general-purpose procedures for computing both likelihoods and priors. These methods are convenient for modeling purposes, but we view them as, at best, just a first approximation to the knowledge people actually bring to bear on this problem. 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*.

The likelihood of each hypothesis is a function of the size of its extension. 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 or cluster height in the tree. Thus equating node height with approximate cluster size, we have for the likelihood:

$$p(X \mid h) \quad \propto \quad \left[ \frac{1}{\text{height}(h) + \varepsilon} \right]^n \qquad (6)$$

if $x_i \in h$ for all $i$, and 0 otherwise. We add a small constant $\varepsilon > 0$ 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 $\varepsilon$ is not critical. We generally find best results with $\varepsilon$ around 0.05 or 0.1; the simulations in this paper use $\varepsilon = 0.05$. (In Figure 7, all nodes are shown at a height of 0.05 above their true height, reflecting this value of $\varepsilon$.) Larger values of $\varepsilon$ may also be appropriate in situations where the sizes of the concepts are not apprehended so distinctly by the learner. Likelihoods will be monotonically related to heights in the cluster tree for any finite $\varepsilon > 0$, but they become increasingly uniform (and hence uninformative) as $\varepsilon$ increases.

A preference for cluster distinctiveness in the prior can be captured by taking $p(h)$ to be proportional to the branch length separating node $h$ from its parent:

$$p(h) \quad \propto \quad \text{height}(\text{parent}[h]) - \text{height}(h)$$

This measure is maximized for clusters of entities that have high average within-cluster similarity relative to their similarity to the most similar entities outside the cluster. For example, in Figure 7, the class containing all and only the dogs (R) is highly distinctive, but the classes immediately under it (P) or above it (Z) are not nearly as distinctive; accordingly, R receives a much higher prior than P (proportional to 0.131 vs. 0.023). This example shows why a distinctiveness bias in the prior is necessary. In terms of the likelihood, hypothesis P (effectively, *dogs with significant body area colored white*) will typically be slightly preferred to hypothesis R (effectively, *all dogs*), because P is slightly smaller. Yet the strong distinctiveness prior favoring R will ensure that this much more conceptually natural hypothesis receives the

higher posterior probability when a learner observes random examples of dogs (which will tend to fall under both hypotheses).

In general, distinctiveness will be high for basic-level categories, but a prior probability based on distinctiveness is not the same thing as a basic-level bias. Distinctiveness may also be high for other conceptually natural clusters, such as superordinate or subordinate categories. In Figure 7, the superordinate categories are significantly more distinctive than any other categories, which accords with the intuition that most fundamental differences between contrasting ontological categories occurs at the superordinate-level (e.g., animal versus vehicle), rather than at the basic or subordinate-levels (e.g., dog versus cat, or Dalmatian versus terrier). Independent of this general preference for distinctiveness, people may also have a preference to map new words onto basic-level categories (Markman, 1989; Golinkoff, Mervis & Hirsh-Pasek 1994). The existence of a basic-level bias in children's learning is a matter of controversy (Callanan et al., 1994; Waxman, 1990), but the original studies of Rosch et al. (1976) certainly provide strong reasons to think that such a bias would be useful, over and above the preference for distinctiveness we have already introduced. Rosch et al. (1976) found that in spontaneous labeling of objects, adults almost always use basic-level names. This preference was much more extreme than the other basic-level preferences Rosch et al. (1976) reported based on nonlinguistic (perceptual or motor-action) criteria, which suggests that in learning kind labels, it would be appropriate to adopt a basic-level bias over and above a general bias towards more natural (e.g., more distinctive) concepts. Note that this basic-level bias does not reflect learners' beliefs about which word meanings are more natural, but rather their beliefs about how words (specifically, kind labels) tend to be used. The latter belief, as Rosch showed, is strongly supported in naming statistics. The former belief would not be statistically valid: the majority of kind labels do not in fact pick out basic-level concepts, because there are many more subordinate kinds than basic-level kinds that receive labels.

In order to test the utility of this sort of basic-level bias in word learning, we will consider two versions of our model: one that contains no preference to map words onto the basic level other than as instantiated in the distinctiveness prior (Equation 7), and one that contains an extra bias in the prior probability for just those hypotheses corresponding to basic-level words in English: "dog", "truck", and

"pepper" in Figure 7. For those hypotheses, the basic-level bias is implemented by replacing $p(h)$ with $\beta$ times its value given in Equation 7, where $\beta$ is a single free numerical parameter that will be adjusted to provide the best fit to the data.[3]

*Model results*

We first consider the basic Bayesian model using the distinctiveness prior, Equation 7. Figure 8a compares $p(y \in C|X)$ computed from this model with the generalization judgments of our adult participants (Figure 5 and 8d), averaged across participants, superordinate classes (animal, vehicle, vegetable), and test items within a given level of generalization. On the averaged data shown in Figure 8d, the model achieves a reasonable quantitative fit ($r = 0.89$).[4] It also captures the main qualitative features of the data: graded generalization given one example, and more all-or-none, rule-like generalization at the level of the most specific consistent natural concept, given three examples. However, there are also several differences between the model's generalizations and people's judgments: the model produces 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.

---

INSERT FIGURE 8 ABOUT HERE

---

Figure 8b shows the fit of the Bayesian model after incorporating a bias in the prior that favors the three basic-level hypotheses. The strength of the basic-level bias is a free parameter, here set to $\beta = 10$. With this one free parameter, the model now provides an almost perfect fit to the average data ($r = 0.99$). All of the main qualitative trends are captured, including those not accounted for by the Bayesian model without a basic-level bias (in Figure 8a). These results suggest that, at least for adults, hypotheses for word learning are biased specifically towards basic-level object categories, over and above a general preference for more distinctive categories that was captured in the branch length prior (Equation 7 and Figure 8a).

A different picture emerges when we compare these two versions of the Bayesian model with preschool-age children's generalizations (Experiment 3; Figures 6b and 8c). In some ways, children's performance looks more like the Bayesian model's predictions *without* the basic-level bias, particularly in the shift from one example to three subordinate examples. Correlation coefficients for the two models are similar ($r = 0.91$ without the basic-level bias, $r = 0.89$ with the basic-level bias). Because the additional parameter $\beta$ does not contribute significantly to the variance accounted for, and leads to a fit that is qualitatively worse in some ways, these results suggest that child word learners may not have the strong basic-level bias that adults exhibit. Their tendency to extend new words to basic-level matches is much weaker than adults, and may simply be explained as the combination of Bayesian hypothesis averaging (Equation 3) with a general preference for hypotheses corresponding to distinctive categories (Equation 7). We return to this issue in the discussion below.

*Comparison with other models*

Figure 9 illustrates respectively the complementary roles played by the size principle (Equations 5 and 6) and hypothesis averaging (Equation 3) in the Bayesian framework. If instead of the size principle we weight all hypotheses strictly by their prior (including the basic-level bias), 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 (Figure 9a). Mathematically, this corresponds to replacing the size-based likelihood in Equations 5 and 6 with a simpler measure of consistency: $p(X|h) = 1$ if the examples $X$ are consistent with the hypothesis $h$ (i.e., $x_i \in$ h for all $i$), and $p(X|h) = 0$ otherwise. Tenenbaum and Griffiths (2001) called this approach *Weak Bayes*, because it uses only a weak binary measure of consistency in the likelihood rather than the strong assumption of randomly sampled examples implicit in using the size principle. Essentially this algorithm has been proposed by Mitchell (1997), Haussler et al. (1994), and Shepard (1987).

---

INSERT FIGURE 9 ABOUT HERE

---

If instead of averaging the predictions of all consistent hypotheses we base generalization on just the single most probable hypothesis, Bayes reduces to an all-or-none rule-like computation. Priors (again including the basic-level bias) and likelihoods cooperate to rank hypotheses, but only the highest-ranking hypothesis – rather than a probability distribution over hypotheses – is used in generalization. Mathematically, this corresponds to replacing hypothesis averaging in Equation 3 with a simpler decision rule: $p(y \in C|X) = 1$ if $y \in h^*$, and 0 otherwise, where $h^*$ is the hypothesis with maximal posterior probability $p(h|X)$ (in Equation 2). This approach is called *Maximum A Posteriori Bayes*, or *MAP Bayes* for short. As Figure 9b shows, MAP Bayes captures the qualitative trends in how adults and children generalize from multiple examples, including the restriction of generalization after 3 subordinate examples have been observed.[5] However, it does not capture the graded nature of generalization from a single example. It also does not capture the increasing confidence in basic-level generalization that comes from seeing three basic-level examples; unlike both adults and children, MAP Bayes makes exactly the same generalizations from three basic-level examples as it does from just a single example.

---

INSERT FIGURE 10 ABOUT HERE

---

Figure 10 shows the predictions of four alternative learning models. None of these models have been specifically proposed for word learning, but they are generic approaches from the literature on computational models of learning and generalization, and they are representative of previous suggestions for how word learning might be viewed computationally. None are explicitly Bayesian, but to varying degrees they correspond to the two special cases of Bayesian learning shown above. Figure 10a presents the predictions of a simple exemplar-similarity model, in which $p(y \in C|X)$ is computed by averaging the similarity of $y$ to each exemplar in $X$. (We use the mean similarity judgments of the adult participants in

Experiment 1, normalized to a 0-1 scale.) For each set of examples the generalization function is scaled linearly to have a maximum at 1.

Figure 10b shows the predictions of an alternative approach to exemplar similarity, inspired by proposals of Goldstone (1994) and Osherson et al. (1990), in which $p(y \in C|X)$ is computed by taking the maximum similarity of $y$ to all exemplars in $X$. Like Weak Bayes, the pure hypothesis-averaging version of the Bayesian model shown in Figure 9a, both exemplar-similarity models give a soft gradient of generalization from one example but fail to sharpen generalization to the appropriate level given three examples.

More flexible similarity-based models of category learning that incorporate selective attention to different stimulus attributes (e.g., Kruschke, 1992) might be better able to accommodate our data, but not without major modification. These models typically rely on error-driven learning algorithms, which are not designed to learn how broadly they should generalize from just one or a few positive examples without any negative examples, and low-dimensional spatial representations of stimuli, which are not suited to representing a broad taxonomy of object kinds.

Several authors have suggested that associative or correlational learning algorithms, perhaps instantiated in neural networks, can explain how children learn the meanings of words (Regier, 1996, 2003; Colunga & Smith, 2000; Smith & Gasser, 1998). It is not possible here to evaluate all extant correlational learning algorithms, but we do consider the standard approach of Hebbian learning (Hertz, Krogh, and Palmer, 1991). Figure 10c shows the predictions of a Hebbian learning network that is matched as closely as possible in structure to our Bayesian models. The Hebbian model uses input features corresponding to the same hypotheses used in our Bayesian models, but instead of evaluating and averaging those hypotheses with the machinery of Bayesian inference, it uses the Hebb rule to compute associative weights between each input feature unit and an output unit representing the occurrence of the novel word to be learned (e.g., "fep"). This network produces generalization patterns very much like those produced by the exemplar-similarity models (Figures 10a,b) or weak Bayes (Figure 9a), capturing something of the graded character of

one-shot generalization but failing to account for how generalization sharpens up to the appropriate level after seeing three examples.

The similar predictions of these various models reflect two underlying computational commonalities. First, learning in the Hebbian network is strictly based on the frequency with which input features occur in the observed examples: each exemplar leaves a trace of its feature values in the weights connecting input features to the output unit, and the final pattern of generalization is proportional to the average of the generalization (or "similarity") gradients produced by each exemplar individually. Second, all of these models fail to converge to the appropriate level of specificity given multiple examples, because they all lack the size principle for re-weighting multiple consistent hypotheses to prefer the hypothesis most likely to have produced the observed examples. When the Hebbian learning network is presented with multiple examples in the same subordinate category (e.g., three Dalmatians), the correlation between the output unit and input features specific to Dalmatians is no greater than the correlation between output and input features that apply to all dogs, or all animals, because every Dalmatian exemplar activates the dog and animal units as well as the Dalmatian units. Because these correlations are independent of the number of examples observed, the Hebbian model cannot explain why generalization beyond the subordinate concept *decreases* as more examples lying strictly within the subordinate are observed (e.g., why seeing three Dalmatian exemplars leads to lower generalization to other dogs, relative to seeing just one Dalmatian exemplar). The same problem afflicts more powerful associative learning mechanisms, such as standard neural networks trained using backpropagation of errors (Rumelhart, Hinton & Williams, 1985), or the recent associative models of word learning (Colunga & Smith, 2005; Regier, 2003, 2005), which are also defined solely in terms of the statistics of input-output co-occurrence.

The Hebb rule, or other associative learning algorithms, could be modified to include some version of the size principle. For instance, we could allow learning rates to vary for different input features as a function of feature specificity. Such a move might allow Bayesian and associative models of word learning to interact productively. Our point here is not that connectionist models of word learning are untenable, but rather that generic associative learning mechanisms based purely on correlations between observable

features are not sufficient to explain how children or adults learn the meanings of new words. More powerful mechanisms of statistical inference, such as our Bayesian framework, are necessary.

Figure 10d shows the predictions of a standard learning algorithm in the hypothesis elimination paradigm, known as the subset principle (Berwick, 1986; Pinker, 1989; Siskind, 1996; see also Wexler and Cullicover (1980) for a discussion of subset-based learning in syntax acquisition, and Bruner et al., 1957, and Feldman, 1997, for analogous proposals in category learning). The subset principle ranks all hypotheses by inclusion specificity: hypothesis $h_i$ ranks higher than hypothesis $h_j$ if $h_j$ strictly includes $h_i$, that is, if $h_j$ includes every object in $h_i$ as well as at least one other object. A subset learner eliminates all hypotheses inconsistent with the observed examples and then generalizes in an all-or-none fashion according to the highest-ranking remaining hypothesis – the most specific hypothesis consistent with the observed examples. This approach is intuitively sensible and produces reasonable generalizations from multiple examples, but is far too conservative given just a single example.

The patterns of generalization from multiple examples under the subset principle are essentially identical to those of MAP Bayes. Both approaches use just a single all-or-none hypothesis to guide generalization, chosen based on inclusion specificity or posterior probability, respectively. These two criteria often converge because one component of posterior probability is likelihood, and under the size principle the likelihood is proportional to specificity. The posterior probabilities of MAP Bayes also depend on the priors, which exert their strongest role when only one example has been observed. Here, the basic-level bias in the prior accounts for why MAP Bayes generalizes differently than the subset principle on one example – to all basic-level matches rather than to just the subordinate matches. The more examples have been observed, the stronger the influence of the specificity preference in the likelihood over the prior, and the more likely it is that MAP Bayes and the subset principle will coincide. In the limit of infinite data, MAP Bayes (as well as our full Bayesian models) become equivalent to maximum likelihood, which under the size principle is also equivalent to subset learning. Thus the subset principle can be justified as a rational statistical inference when large numbers of examples have been observed, and it is precisely this case of "learnability in the limit" that has been the focus of most previous uses of the subset principle. Our

Bayesian models, based on the size principle, can be viewed as extensions of the subset principle to explain the dynamics of learning from just one or a few examples – arguably the most important regime of learning for many real-world words and concepts. More broadly, various phenomena of entrenchment and conservatism in language acquisition (Braine & Brooks, 1995; Goldberg, 2003) may be more consistent with our softer, statistical model than with the hard commitments of the subset principle.

*Summary*

In sum, our inductive models may be seen as probabilistic generalizations of the classic deductive approach to word learning based on hypothesis elimination. As in hypothesis elimination accounts, a constrained hypothesis space makes possible meaningful generalization from examples. But in contrast to these accounts, hypotheses are not just ruled in or out. Using Bayes' rule, they are assigned a probability of being correct based on how well they explain the pattern of examples observed. The assumption that the observed examples are randomly sampled from the word's extension provides a powerful statistical lever, yielding strong but reliable generalizations from just a few examples. Our experiments in the domain of words for object categories 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, exemplar-similarity, or associative or correlational learning. In particular, the Bayesian approach naturally explains the spectrum of generalization behavior observed given one or a few positive examples. Graded generalization with one example follows straightforwardly from the mechanism of hypothesis averaging, while the sharpening from one to three examples follows straightforwardly from the size principle. Bayesian inference may thus offer the most promising framework in which to explain the speed and success of fast mapping.

Could other models from the hypothesis-elimination or associative traditions be extended to accommodate our findings? Not easily we think, and not without positing additional machinery that either is inelegant or fundamentally departs from the original spirit of these approaches. Using the deductive framework of hypothesis elimination, in order to explain the sharpening of generalization 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

subset bias (choosing 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 kind-labels difficult to learn. Since children do eventually learn names at these levels, hypothesis elimination approaches would have to posit further provisions, such as overriding the basic-level bias with time, or incorporating some other linguistic cue to the appropriate level of generalization for a label.

As mentioned earlier, it might be possible to extend an associative model of word learning to account for the range of generalization behavior we observed, by building the size principle and hypothesis averaging into its learning and activation rules. However, even those extensions might not be sufficient. Computations in associative models are typically defined not over a core ontology of objects and object-kind concepts, but over relationships between perceptual features: e.g., the visual features of objects and the sound features of words (Regier, 2005). This focus on learning correlations at the level of perceptual features could stand in the way of appropriate generalization. Our Bayesian learner sees a critical difference between one object labeled three times and three distinct but perceptually highly similar objects each labeled once; so do four-year-old children, as we showed in Experiment 3. Although both cases provide three observations of word-object pairings, and the object features are almost the same in both cases, the latter case provides three independent samples of objects in the concept, while the former case provides only one independent sample. Thus only the latter case provides strong evidence about the extent of the concept, and only in the latter case do children restrict their generalization to just objects in the same subordinate category. For an associative learner to appreciate this difference, it would need to gain not only something like the size principle in its learning rule, but also some kind of ontology that understands about the differences between objects, percepts, and categories (e.g., Keil, 1989; Spelke, 1990). It would also need to build some kind of taxonomic hierarchy of object-kind categories on top of that ontology. Although these capacities are not currently part of conventional associative models, they might not be incompatible with a more general, predictive-learning view of associationism (Smith, 2000). Still, adding these capacities would

seem to abandon a core associationist claim that word learning can be explained without sophisticated inferential mechanisms or sophisticated representations of the world (e.g., Landau, Smith, & Jones, 1996).

## Extending the Bayesian framework

Our models here have focused on the inductive problem of word learning in its simplest form: learning the meaning of a single new word in a fairly restricted hypothesis space, given observations of how that word is used to label one or more entities in the world. We have tried to keep our models as simple as possible, to capture some fundamental insights about how the meanings of important classes of words may be learned from very limited data. But word learning in the real world is considerably more complex, in terms of the kinds of hypothesis space the learners entertain, the kinds of inferences required, and the kinds of data brought to bear on those inferences. Word learning is also a dynamic process, in which knowledge gained from previous word learning experience – both the specific meanings of particular words and abstractions about the general principles of word meaning and usage – leads to crucial constraints on future word learning (e.g., Baldwin, 1993; Bloom, 2000; Gleitman, 1990; Markman, 1989; Regier, 2003; Tomasello, 2001). This section briefly sketches some of the possible avenues for extending our models to handle these complexities.

### *A differently structured hypothesis space: objects and solid substances*

The fundamental problem of induction in word learning is how to choose among the multiple potential concepts – hypotheses for word meanings – that are consistent with the observed examples of a new word. So far we have addressed this problem in the context of learning count nouns for object kinds, where multiple consistent hypotheses come from hierarchically nested kind-concepts. In learning other sorts of words, different kinds of inductive ambiguities can arise when hypotheses overlap in other ways. Here we sketch a Bayesian analysis of one such case, cross-cutting object kinds and solid substance kinds, showing how the same general framework we developed for learning kind-labels applies even when the relevant concepts do not conform to a nested hierarchy.

Consider the "furniture store" context raised in the introduction. The entities in a furniture store may be referred to in terms of either their object category, e.g., "chair", "table", "shelf", "vase", or the

material they are made out of, e.g., “wood”, “plastic”, “metal”, “stone”. More generally, any solid entity may be construed in at least two modes, as an object of a particular kind or as the solid substance(s) that comprises it, and words are available to refer to either of these two modes. Learning words for solid entities thus poses a challenge of learning with overlapping hypotheses – but not nested hypotheses. Object-kind categories cross-cut solid-substance categories, with each kind of object realizable in many different substances, and each substance capable of taking on many different shapes (Figure 2).

Prasada, Ferenz and Haskell (2002) explored the conditions under which people would construe a solid entity in terms of an object-kind or substance-kind concept, when both cross-cutting hypotheses were available. They showed people solid entities composed of unfamiliar materials, with either regular shapes or complex irregular shapes, and asked whether they would prefer to call one of these entities “a blicket” or “some blicket”. Choosing “a blicket” suggests that “blicket” refers to an object category, while choosing “some blicket” suggests a substance kind. Prasada et al. found that given a single regularly-shaped entity, people tended to choose an object category, while given a single irregularly-shaped entity, people tended to choose the substance interpretation. Inferences given multiple examples were generally consistent with these single-example cases, with one interesting exception. When people were shown multiple essentially identical objects, each with the same complex irregular shape and novel material, their preference for labeling an object in this set switched from a substance interpretation to an object interpretation.

Our Bayesian framework can explain these inferences, if we assume that people are treating the examples given as random samples from one of two hypotheses for the meaning of the novel word “blicket”, an object-kind category or a substance category. The goal is to infer which hypothesis is more probable given the examples observed. Technical details are beyond our scope here, but there are several basic assumptions from which the results are derived. First, each object category is organized around a prototypical shape. Second, object categories with regular shapes should have higher priors than substance categories, which in turn should have higher priors than object categories with irregular shapes. This is consistent with English word frequencies, which are higher for regularly-shaped object category labels than for material or irregularly-shaped object category labels (e.g., Landau, Smith, & Jones, 1988; Bloom, 2000).

Third, there are more conceptually distinct shapes that support object categories than there are material properties that support substance categories. Thus the effective size of each object-kind hypothesis is smaller than each substance hypothesis, and it is more of a suspicious coincidence to observe three randomly sampled entities with the same novel shape than with the same novel material. Speakers of English do tend to find more salient and potentially nameable differences among object shapes than among material substances, but there are also cross-linguistic differences (Imai and Gentner, 1997).

Together, these ingredients allow us to explain Prasada et al's finding of a shift from a substance-based interpretation for the novel word given one irregularly-shaped example, to an object-based interpretation given three essentially identical examples with the same irregular shape and material. The prior initially favors the substance interpretation, but the shift in generalization with multiple examples comes from detecting a suspicious coincidence – a reflection of the size principle in the likelihood term. It would be a strong coincidence to observe three random samples all with the same irregular shape if the novel word is intended to label a substance kind, which suggests that an object-kind construal is more likely to be correct.

Prasada et al. (2002) interpret their findings in similar terms, arguing that people will be more likely to interpret an entity as an instance of an object kind if its form appears non-arbitrary. This interpretation also explains a further finding of theirs, that a single entity with an irregular shape can be construed as an instance of an object kind if that shape is shown to have functional significance – that is, in another sense, if its shape appears not to be a coincidence. By framing these interpretations explicitly in terms of statistical inference, as we do, we can see how they reflect more general rational inferential mechanisms at work in understanding and learning the meanings of words. The same mechanisms that underlie people's ability to infer the appropriate scope or range of generalization, in learning names for hierarchically nested categories, also support the ability to infer the appropriate directions or dimensions for generalization, where multiple plausible hypotheses cross-cut each other. Recent work has shown how to extend this Bayesian approach to learning other aspects of linguistic meaning, using differently structured hypothesis spaces appropriate for

learning verb frames (Niyogi, 2002), color terms (Dowman, 2002), or principles of anaphora resolution (Regier and Gahl, 2004).

*Transforming the likelihood function*

In the basic Bayesian framework presented above, any information about the meaning of a new word contributes through its influence on either the likelihood or the prior. By changing or expanding on one or both of these terms, we can address more complex kinds of inferences or incorporate additional sources of constraint on the learner's inferences.

*Other sources of input*. Earlier we mentioned two sources of information about word meaning that we did not explicitly incorporate into our formal analyses: negative examples – examples of entities that a word does not apply to – and special linguistic cues that relate the meaning of the new word to familiar words – as in saying, "This is a Dalmatian. It's a kind of dog." Although we have focused above on learning from positive examples without special linguistic cues, either negative examples or relational linguistic cues could sometimes be crucial in inferring the scope of a new word's extension. Our Bayesian framework can naturally accommodate these sources of information through straightforward modifications to the likelihood function. We can assign zero likelihood to any hypothesis that includes one or more negative examples, essentially treating negative examples as a deductive constraint on candidate word meanings. A cue such as "This is a Dalmatian. It's a kind of dog." can also be treated as a deductive constraint, by assigning zero likelihood to any hypothesis for "Dalmation" that is not contained within the extension of the word "dog." A Bayesian learner could then rationally infer a subordinate meaning for the new word "fep", given just one positive example of a "fep" (e.g., a Dalmatian), and either of these two additional sources of input – a negative example (e.g., a Labrador that is not a "fep"), or a relational cue in language (e.g., "Feps are a kind of dog").

*Theory-of-mind reasoning and sensitivity to sampling*. A vital source of information about word meaning comes from theory-of-mind reasoning (e.g., Baldwin, 1991, 1993; Bloom, 2000; Tomasello, 2001). The fact that a certain kind of object can be labeled with a certain word is not just a simple perceptual feature to be associated with the corresponding object concept. Words are tools used by intentional agents

to refer to aspects of the world, and most examples of words that a learner observes are the consequences of intentional acts of reference. Inferences based on theory-of-mind reasoning are often put in opposition to statistical inferences about word meaning (Bloom, 2000; Regier, 2003), when the latter are construed as bottom-up associative processes. But in the top-down, knowledge-based approach to statistical inference that we propose here, theory-of-mind considerations could play a critical role. Making statistical inferences about the meanings of words from examples may demand from the learner, in addition to other abilities, a sensitivity to the intentional and epistemic states of the speakers whose communicative interactions produce the examples observed. This sensitivity may enter into our Bayesian framework in specifying the sampling assumption that determines the appropriate likelihood function to use.

Although we have not yet explored our framework's predictions in settings with strong theory-of-mind demands, we have tested whether children and adults are sensitive to the sampling process generating the examples that they see labeled, and whether they adjust their likelihoods accordingly. Xu and Tenenbaum (in press) studied generalization for novel object-kind labels in two conditions, one in which the labeled examples appeared to be sampled randomly from the set of objects the word applies to, and another in which essentially the same examples were observed, but they were clearly not randomly sampled from the word's extension. The stimuli were simple novel objects, generated by a computer drawing program. As in the studies reported above, the objects could be classified at multiple levels of a clear, salient hierarchy of classes. The "teacher-driven" condition was similar to the 3-subordinate trials of the experiments above. The experimenter pointed to an object and said to the child, "This is a fep!" Then she pointed to two distinct (but very similar-looking) objects, from the same subordinate category, and labeled each one "a fep." Here it is reasonable (to a first approximation) to treat these examples as random samples of feps. In the "learner-driven" condition, however, after the first labeled example was chosen by the experimenter, she said to the child, "Can you point to two other feps? If you can get both of them right, you will get a sticker." In this case, children were motivated to choose two other objects from the same subordinate category in order to get a sticker, both of which were labeled as correct by the experimenter. The children in the learner-driven condition received essentially the same data as the children in the teacher-driven condition. However, in the

learner-driven condition, the three examples cannot be treated as a random sample drawn from the word's extension, since the child does not know the meaning of the novel word (and is in fact trying to choose objects that are as similar to the one labeled by the experimenter).

Our results showed that both adults and preschoolers were sensitive to the sampling conditions. In the teacher-driven condition, we replicated our results from the current experiments – the learners restricted their generalization of the novel word to other subordinate exemplars. In the learner-driven condition, however, both adults and children generalized more broadly, to the basic-level category that included these examples. This is just what a Bayesian analysis would predict in a situation where the examples to be labeled are sampled independently of the meaning of the word (Xu and Tenenbaum, in press). The likelihood, instead of reflecting the size principle, now becomes simply a measure of consistency: it is proportional to 1 for hypotheses consistent with the labeled examples, and 0 for all inconsistent hypotheses. Without an increasing preference for smaller hypotheses, a Bayesian learner will maintain the same basic-level threshold of generalization as additional examples are observed beyond the first, as long as they are all consistent with the same set of hypotheses.

This sensitivity to sampling conditions is a distinctive feature of our Bayesian approach. It is not predicted by either traditional associative or deductive accounts of word learning, because they do not view word learning as fundamentally a problem of making statistical inferences from samples to underlying explanatory hypotheses. Associative approaches typically embody some implicit statistical assumptions, but they do not make these assumptions explicit and grant learners the power to make inferences about the sampling process. They thus forgo not only an important aspect of rational statistical inference, but also an important contribution of intentional reasoning to the word learning process.

*Transforming prior probabilities*

*The effects of previously learned words.* There are several ways in which word meanings learned previously can constrain the meanings of new words to be learned. One way is through the development of abstract syntax-semantics mappings, such as a bias to map count nouns onto object kinds and mass nouns onto substance kinds (Colunga & Smith, 2005; Kemp et al., in press). Another way is through lexical

contrast, an assumption that the meanings of all words must somehow differ (Clark, 1987). Both of these influences can be captured in a Bayesian framework by modifying the learner's prior probabilities. Although the technical details are beyond the scope of this paper, here we sketch a Bayesian analysis for lexical contrast.

Mutual exclusivity is one simple form of lexical contrast: a constraint that each entity has only one label, and thus no two words can have overlapping extensions (Markman, 1989). The simplest way to capture mutual exclusivity in our framework is in the prior. If mutual exclusivity is assumed to be a hard constraint, we simply set the prior to zero for any hypothesis about the extension for a new word that overlaps the extension of a previously learned word. If mutual exclusivity is taken to be only a soft bias rather than a hard constraint, then the prior probability for hypotheses with extensions overlapping those of known words could be set to some fraction of its default value. Regier (2003) suggests an alternative way that mutual exclusivity could enter into a Bayesian analysis, via an alternate formulation of the likelihood.

Mutual exclusivity could be useful in early stages of word learning, but it excludes all cases of meaning overlap we have studied here and makes it impossible to learn word meanings like *animal*, *Dalmatian*, *pet*, and so on, just for the sake of learning the one basic-level kind term "dog". Clark's principle of contrast (Clark, 1987) is a weaker version of lexical contrast that is more suited to the mature lexicon: we assume that no two words have exactly the same meaning, although their extensions may overlap in any way other than complete identity. Formally, this principle could be implemented just like mutual exclusivity, by setting the prior probability of any hypothesis that corresponds to the extension of a known word to zero, or to some small fraction of its default value if a softer bias is called for.

Our analysis of lexical contrast effects has so far assumed a highly idealized scenario, in attributing to the learner a completely fixed lexicon of previously learned words. In practice, learners will be learning many words at a time, with varying degrees of experience and confidence in meaning. A more realistic Bayesian formulation of the word learning problem would construe the hypotheses and data as language-wide structures rather than learning individual word-concept mappings. The learner would evaluate hypotheses about possible sets of word-concept mappings for the entire language, based on the full body of

data for all words in the language seen to date. The size principle in the likelihood would still apply separately for each word. The prior over candidate lexicons might incorporate all the factors discussed so far, including a principle of contrast and a bias to map words onto a priori natural and distinctive concepts. Directly implementing this language-wide approach would be computationally intractable, but some kind of online approximation could usefully describe the trajectory of large-scale vocabulary acquisition.

## Open issues

Although our theoretical framework aims for generality, many important questions of word learning are beyond its current scope. Here we sketch several of these open questions.

First, we have emphasized the phenomenon of fast mapping in both adults and children, showing how our Bayesian models naturally give rise to very efficient learning from just a few examples. But many researchers have suggested that very early word learning is a fundamentally different kind of process. It is often characterized as a slow and laborious enterprise (e.g., Dromi, 1987; Golinkoff et al., 1994). Children between 12 and 18 months require many exposures to a single word in order to learn it (but see Woodward et al., 1994), and sometimes words appear to drop out of their lexicon. It is unclear why very early word learning appears so much less efficient than later stages, and how that reflects on the applicability of our Bayesian framework to the earliest stages of word learning.

There are at least four possible reasons why word learning in the youngest children might not look like the fast-mapping behavior of our Bayesian models. First, the necessary capacity for Bayesian inference may not be available to the youngest children, but may itself develop (through simple maturation, or in a way that depends on the development of other general-purpose cognitive capacities). Associative models of word learning (e.g., Regier, 2005; Colunga & Smith, 2005) often focus on the earliest stages of word learning, and it is certainly possible that word learning is best characterized as initially associative but Bayesian in the more mature state that we have studied here. Second, the capacity for Bayesian inference may be available, but very young children may have much weaker, less constrained hypothesis spaces, which do not support learning with high confidence from just a few examples. That is, they may be viewed as Bayesian word learners but without the appropriate hypothesis spaces. Third, very young children could

possess domain-general Bayesian inference capacities but not yet be able to apply these mechanisms to the task of word learning. For instance, they might not yet grasp the concepts of reference and intention necessary to treat observations of word-object labeling events as randomly sampled examples of a word's reference class, and thus not be able to set up the likelihood functions appropriate for word learning. Finally, the youngest children could possess all of these core conceptual capacities but still suffer from processing limitations that prevent them from remembering words stably over time or fixing the referent of a word quickly. More research is necessary to distinguish among these and other possible accounts of the earliest stages of word learning.

We have so far treated word learning as a mapping problem: learners possess concepts – hypotheses for candidate word meanings – independent of those words, and their task is to map word forms onto these concepts. This view does not imply that the concepts are innate, just that they are mostly in place by the time the words are being learned. But it is quite possible that word learning and concept formation proceed in parallel to some extent (e.g., Bowerman & Levinson, 1996; Gentner & Goldin-Meadow, 2003; Xu, 2002, 2005). In terms of our Bayesian framework, perhaps the observation of new words that cannot be mapped easily onto the current hypothesis space of candidate concepts somehow triggers the formation of new concepts, more suitable as hypotheses for the meanings of these words. Bayesian models of the relation between word learning and concept learning more generally are one focus of our ongoing work (Perfors, Kemp & Tenenbaum, 2005).

More generally, questions about the origins of the learner's hypothesis space are clearly important targets for future work. These questions can be asked on at least two levels. First, and most deeply, how does the learner acquire the abstract knowledge that a certain class of words should map onto a hierarchy of object kinds, and that certain kinds of perceptual features are typically diagnostic of kind membership? Second, given this abstract knowledge, how does the learner construct a concrete tree-structured hierarchy onto which words for object kinds will be mapped? In principle, both of these questions can be addressed within a hierarchical Bayesian framework (Tenenbaum, Griffiths, and Kemp,2006; Kemp, Perfors, and Tenenbaum, 2004, in press), an extension of the approach we have developed here to include hypothesis

spaces at multiple levels of abstraction, with probabilistic models linking each level in the hierarchy. The second question is easier to address, and in some sense is already addressed implicitly in the work presented above. Given the goal of searching for a tree-structured hierarchy of object kinds, and a sense of which perceptual features are characteristic of how kinds cohere, the learner just needs to perform some kind of hierarchical clustering on the objects it has observed in the world. As Kemp, Perfors, and Tenenbaum (2004) discuss, this hierarchical clustering can be viewed as a Bayesian inference, searching for the simplest tree that assigns a high likelihood to the observed object features, under a probabilistic model in which objects that are nearby in the tree are expected to look more similar than objects that are far apart in the tree. The first question can be addressed by the same logic. The learner considers different classes of structures that could generate a hypothesis space of word meanings, including tree-structured object-kind hierarchies as well as other kinds of structure. Each of these abstract organizing principles can also be scored according to how well it predicts the observed object features, although in practice computing this score could be quite difficult, as it involves summing or searching over all specific structures consistent with each class of structures (Kemp, Perfors, and Tenenbaum, 2004; Perfors, Kemp, and Tenenbaum, 2005).

Our analysis of word learning focuses on what Marr (1982) called the level of computational theory. We have tried to elucidate the logic behind word learners' inductive inferences, without specifying how that logic is implemented algorithmically in the mind or physiologically in neural hardware. We make no claim that Bayesian computations are implemented exactly in the mind or brain, with explicitly represented probabilities. On the contrary, it is more likely that the details of mental or neural processing correspond to some efficient approximation to the Bayesian computations we propose here. We also make no claim that any of these computations have consciously accessible intermediate steps. The fact that people are typically not aware of considering many hypotheses for a word's meaning does not mean that the mind does not implicitly behave in accord with our Bayesian principles.

Lastly, the learning mechanism we have proposed here is unlikely to be specific to word learning or language acquisition. Recent research has shown that other domains of inductive learning and reasoning may be explained in Bayesian terms, including causal learning (Gopnik, Glymore, Sobel, Schulz, Kushnir,

& Danks, 2004; Griffiths & Tenenbaum,2005; Sobel, Tenenbaum, & Gopnik, 2004; Steyvers, Tenenbaum, Wagenmakers & Blum, 2003), category-based induction (Heit, 1998; Kemp & Tenenbaum, 2003), conditional reasoning (Oaksford & Chater, 1994) and covariation assessment (McKenzie, 1994; McKenzie & Mikkelsen, in press). Whether word learning requires specialized mechanisms or assumptions is the subject of a lively debate in the field of cognitive and language development (e.g., Bloom, 2000; Behrend et al., 2001; Deisendruck & Bloom, 2003; Diesendruck & Markson, 2001; Waxman & Booth, 2002; Xu, Cote, & Baker, 2005). Although word learning may require certain language-specific principles or structures, it is plausible that the inference mechanisms, as we suggest here, are domain-general.

## Conclusion

In this work, we have taken a close-up look at only a few pieces of a big puzzle. We have argued that a Bayesian approach provides a powerful computational framework for explaining how people solve the inductive problem of learning word meanings, by showing how the approach gives distinctive insights into several core phenomena of word learning as well as strong quantitative fits to behavioral data from our experiments with adult and child learners. We should caution against concluding too much from the studies presented here. The specific experimental tasks and models we have worked with simplify the real challenges that children face in so many ways, and they leave many aspects of word learning completely unaddressed – even if they suggest a number of promising extensions. Yet we still think there are valuable lessons to be drawn, about the computational basis of word learning and cognitive development more generally.

Accounts of cognitive development typically view statistical learning and sophisticated representational machinery as competing – or even mutually exclusive – explanations for how we come to know so much about the world. Here we have presented a theoretical framework for explaining one aspect of development, word learning, based on the operation of powerful statistical inference mechanisms defined over structured mental representations. In contrast to the associative tradition, our approach has critical roles for conceptual hierarchies, individuated objects as distinct from word-percept correlations, and abstract linguistic or communicative principles. Unlike traditional rationalist approaches, ours is at heart about

statistical inference, in which knowledge about word meanings can be more or less graded depending on the probabilistic evidence provided by different degrees of data. A fully satisfying computational model of word learning remains as remote as a model of general-purpose cognition, but our work suggests at least one good bet about what such models will have to look like. Only a combination of sophisticated mental representations and sophisticated statistical inference machinery will be able to explain how adults and children can learn so many words, so fast and so accurately.

References

Akhtar, N., Jipson, J., & Callanan, M.A. (2001) Learning words through overhearing. Child Development, 72, 416-430.

Anderson, J.R. (1990) The adaptive character of thought. Hillsdale, NJ: Erlbaum.

Bacon, F. (1620/1960) The new organon, and related writings. Indianapolis, IN: Bobbs-Merrill.

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.

Baldwin, D., Markman, E.M., Bill, B., Desjardins, R. & Irwin, J. (1996). Infants' reliance on a social criterion for establishing word-object relations. Child Development, 67, 3135-3153.

Behrend, D.A., Scofield, J. & Kleinknecht, E.E. (2001). Beyond fast mapping: young children's extensions of novel words and novel facts. Developmental Psychology, 37, 698-705.

Berwick, R.C. (1986). Learning from positive-only examples: The subset principle and three case studies. In J. G. Carbonell, R. S. Michalski & T. M. Mitchell (eds.), Machine Learning: An Artificial Intelligence Approach (vol. 2). Los Altos, CA: Morgan Kauffman, pp. 625-645.

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

Bowerman, M. & Levinson, S.C. (1996). Language acquisition and conceptual development. Cambridge University Press.

Braine, M. & Brooks, P. (1995). Verb argument structure and the problem of avoiding an overgeneral grammar. In M. Tomasello & W. Merriman (Eds.), Beyond names for things: Young children's acquisition of verbs. Hillsdale, NJ: Erlbaum.

Brown, R. (1957). Linguistic determinism and the part of speech. Journal of Abnormal and Social Psychology, 55, 1-5.

Bruner, J.A., Goodnow, J.S. & Austin, G.J. (1956). A study of thinking. New York: Wiley.

Callanan, M.A. (1985). How parents label objects for young children: the role of input in the acquisition of category hierarchies. Child Development, 56, 508-523.

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

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

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

Carey, S. & Bartlett, E. (1978). Acquiring a single new word. Papers and Reports on Child Language Development, 15, 17-29.

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

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

Diesendruck, G. & Bloom, P. (2003). How specific is the shape bias. Child Development, 74, 168-178.

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

Duda, R.O. & Hart, P.E. (1973). Pattern classification and scene analysis. New York: Wiley.

Feldman, J. (1997). The structure of perceptual categories. Journal of Mathematical Psychology, 41, 145-170.

Dowman, M. (2002). Modelling the Acquisition of Colour Words. In B. McKay and J. Slaney (eds.), Advances in Artificial Intelligence. Berlin: Springer, pages 259-271.

Dromi, E. (1987). Early lexical development. Cambridge University Press.

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

Gentner, D. & Goldin-Meadow, S. (2003). Language in mind. MIT Press.

Gleitman, L.R. (1990). The structural sources of verb meanings. Language Acquisition, 1, 3-55.

Gold, E.M. (1967). Language identification in the limit. Information and Control, 16, 447-474.

Goldberg, A. (2003). Constructions: A new theoretical approach to language. Trends in Cognitive Sciences 7(5), 219-224.

Goldstone, R.L. (1994). The role of similarity in categorization: providing a groundwork. Cognition, 52, 125-157.

Golinkoff, R.M., Mervis, C., & Hirsch-Pasek, K. (1994). Early object labels: the case for a developmental lexical principles framework. Journal of Child Language, 21, 125-155.

Gopnik, A., Glymore, C., Sobel, D.M., Schulz, L.E., Kushnir, T. & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. Psychological Review, 111, 3-32.

Griffiths, T.L. & Tenenbaum, J.B. (2005). Structure and Strength in causal induction. Cognitive Psychology 51(4), 285-386.

Griffiths, T.L., Steyvers, M. & Tenenbaum, J.B. (submitted). Topics in semantic association. Psychological Review.

Haussler, D., Kearns, M. & Schapire, R.E. (1994). Bounds on the sample complexity of Bayesian learning using information theory and the VC-dimension. Machine Learning, 14, 83-113.

Heibeck, T. & Markman, E.M. (1987). Word learning in children: an examination of fast mapping. Child Development 58, 1021-1034.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford and N. Chater (Eds.), Rational models of cognition (pp. 248-274). Oxford University Press.

Hertz, J., Krogh, A., & Palmer, R.G. (1991). Introduction to the theory of neural computation. Boston, MA: Addison-Wesley.

Imai, M. & Gentner, D. (1997). A cross-linguistic study of early word meaning: universal ontology and linguistic influence. Cognition, 62, 169-200.

Jaswal, V.K. & Markman, E.M. (2001). Learning proper and common nouns in inferential vs ostensive contexts. Child Development, 72, 768-786.

Keil, F.C. (1979). Semantic and conceptual development: An ontological perspective. Cambridge, MA: Harvard University Press.

Kemp, C., Perfors, A. F., and Tenenbaum, J. B. (2004). Learning domain structure. Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society.

Kemp, C., Perfors, A. F., and Tenenbaum, J. B. (in press). Learning overhypotheses with hierarchical Bayesian models. Developmental Science.

Kemp, C. & Tenenbaum, J.B. (2003). Theory-based induction. Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society.

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

Landau, B., Smith, L.B., & Jones, S. (1988). The importance of shape in early lexical learning. Cognitive Development, 5, 287-312.

Landau, B., Smith, L.B., & Jones, S. (1997). Object shape, object function and object name. Journal of Memory and Language, 36, 1-27.

Landauer, T. & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 104, 211-240.

Li, P. & MacWhinney, B. (1996). Cryptotype, overgeneralization, and competition: A connectionist model of the learning of English reversive prefixes. Connection Science 8, 3-30.

Liu, J., Golinkoff, R., & Sak, K. (2001). One cow does not animal make: young children can extend words at the superordinate level. Child Development, 72, 1674-1694.

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

Markman, E.M. & Hutchinson, J.E. (1984). Children's sensitivity to constraints on word meaning: taxonomic versus thematic relations. Cognitive Psychology 8, 561-577.

Markman, E.M. & Wachtel, G.F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. Cognitive Psychology 20, 121-157.

Marr, D.C. (1982) . Vision. San Francisco: W.H. Freeman.

Merriman, W. (1999). Competition, Attention, and Young Children's Lexical Processing. In Brian MacWhinney (Ed.) The Emergence of Language (pp. 331-358). Lawrence Erlbaum Associates, Mahwah, NJ.

MacWhinney, B. (1998). Models of the emergence of language. Annual Review of Psychology 49, 199-227.

McKenzie, C.R.M. (2004). The accuracy of intuitive judgment strategies: covariation assessment and Bayesian inference. Cognitive Psychology, 26, 209-239.

McKenzie, C.R.M. & Mikkelsen, L.A. (in press). A Bayesian view of covariation assessment. Cognitive Psychology.

Mill, J.S. (1843). A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence, and methods of scientific investigation. London: J.W. Parker.

Mintz, T.H. & Gleitman, L.R. (2002). Adjectives really do modify nouns: the incremental and restrictive nature of early adjective acquisition. Cognition,84, 267-293.

Mitchell, T.M. (1982). Generalization as search. Journal of Artificial Intelligence, 18, 203-226.

Mitchell, T.M. (1997). Machine learning. New York: McGraw Hill.

Niyogi, S. (2002). Bayesian learning at the syntax-semantics interface. In W. Gray and C. Schunn (eds.), Proceedings of the 24th Annual conference of the Cognitive Science Society. Lawrence Erlbaum Associates, pages 697-702.

Oaksford, M. & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. Psychological Review 101(4), 608-631.

Oaksford, M. & Chater, N. (1998). Rational models of cognition. Oxford University Press.

Osherson, D.N., Smith, E.E., Wilkie, O., Lopez, A. & Shafir, E. (1990). Category-based induction. Psychological Review, 97, 185-200.

Perfors, A., Kemp, C., & Tenenbaum, J.B. (2005). Modeling the acquisition of domain structure and feature understanding. Proceedings of the 27th Annual Conference of the Cognitive Science Society.

Pinker, S. (1979). Formal models of language learning. Cognition, 1, 217-283.

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

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

Plunkett, K., Sinha, C., Moller, M., & Strandsby, O. (1992). Symbol Grounding or the Emergence of Symbols? Vocabulary Growth in Children and a Connectionist Net. Connection Science 4, 293-312.

Popper, K. (1959). The logic of scientific discovery. New York: Basic Books.

Prasada, S., Ferenz, K. & Haskell, T. (2002). Conceiving of entities as objects and as stuff. Cognition, 83, 141-165.

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

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. Cognitive Science 22(4), 425-469.

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

Regier, T. (2003). Emergent constraints on word-learning: A computational review. Trends in Cognitive Science, 7, 263-268.

Regier, T. & Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. Cognition, 93(2), 147-155.

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

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

Roy, D. & Pentland, A. (2004). Learning words from sights and sounds: A computational model. Cognitive Science 26(1), 113-146.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning representations by back-propagating errors. Nature, 323, 533-536.

Shepard, R.N. (1987). Towards a universal law of generalization for psychological science. Science, 237, 1317-1323.

Shepard, R.N. & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. Psychological Review, 86(2), 87-123.

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

Smith, L. B. (2000). Avoiding associations when it's behaviorism you really hate. In R. M. Golinkoff and K. Hirsh-Pasek (eds.), Becoming a word learner: A debate on lexical acquisition. Oxford University Press.

Sobel, D., Tenenbaum, J.B. & Gopnik, A. (2004). Children's causal inferences from indirect evidence: backwards blocking and Bayesian reasoning in preschoolers. Cognitive Science, 28, 303-333.

Soja, N.N., Carey, S., & Spelke, E.S. (1991). Ontological categories guide young children's induction of word meaning: object terms and substance terms. Cognition, 38, 179-211.

Spelke, E.S. (1990). Principles of object perception. Cognitive Science, 14, 29-56.

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

Tenenbaum, J.B. (1999). A Bayesian framework for concept learning. Unpublished doctoral dissertation. Massachusetts Institute of Technology.

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., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. Trends in Cognitive Sciences, 10, 309-318.

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.

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

Tversky, A. (1977). Features of similarity. Psychological Review 84(4), 327-352.

Waxman, S.R. (1990). Linguistic biases and the establishment of conceptual hierarchies: Evidence from preschool children. Cognitive Development, 5, 123-150.

Waxman, S.R. & Booth, A.E. (2002). Word learning is "smart": Evidence that conceptual information affects preschooler's extension of novel words. Cognition, 84(1), 11-22.

Wexler, K. & Cullicover, P. (1980). Formal principles of language acquisition. Cambridge, MA: MIT Press.

Woodward, A., Markman, E.M., & Fitzsimmons, C. (1994). Rapid word learning in 13-and 18-month-olds. Developmental Psychology 30, 553-566.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. Cognition, 85, 223-250.

Xu, F. (2005). Categories, kinds, and object individuation in infancy. In L. Gershkoff-Stowe and D. Rakison (Eds.), Building object categories in developmental time: Papers from the 32nd Carnegie Symposium on Cognition (pp. 63-89). New Jersey: Lawrence Erlbaum.

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. (in press). Sensitivity to sampling in Bayesian word learning. Developmental Science.

## Author Note

The authors contributed equally to this work and they are listed in order of birth. This research was made possible by a grant from Mitsubishi Electric Research Labs and the Paul E. Newton chair to JBT, and a Canada Research Chair and grants from National Science Foundation and Natural Science and Engineering Research Council of Canada to FX. We thank Eliza Calhoun and Sydney Goddard for assistance with behavioral experiments. We thank Dare Baldwin, Paul Bloom, Jeff Elman, Jesse Snedeker, and Elizabeth Spelke for helpful discussions about this work, and Nick Chater, Susan Carey, Terry Regier, and several anonymous reviewers for valuable comments on this manuscript. We also thank the parents and children for their participation. Portions of this work were presented at the Annual Conference of the Cognitive Science Society (2000, 2005), the Boston University Conference on Language Development (2000, 2002), and the Biennial Conference of the Society for Research in Child Development (2001).

Address correspondence to F. Xu at Department of Psychology, 2136 West Mall, University of British Columbia, Vancouver, B.C., V6T 1Z4, Canada. Phone: (604) 822-5972. Fax: (604) 822-6923. Email: fei@psych.ubc.ca.

Address correspondence to J. B. Tenenbaum at Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. Phone: (617) 452-2010. Fax: (617) 258-8654. Email: jbt@mit.edu.

Notes

1. There is another important approach to modeling the acquisition of word meaning, broadly within the associative-learning tradition, which we do not discuss here because it focuses on a different learning task – complementary to our focus on learning word meanings from examples. This is the problem of learning the associative contexts of words from observing how those words tend to be used together in conversation or writing. Statistical approaches that look for clusters of words occurring in similar contexts (Redington, Chater & Finch, 1998; Griffiths, Steyvers & Tenenbaum, submitted), or a latent space that best explains patterns of word co-occurrence (Landauer & Dumais, 1997), have recently produced intriguing results. It would be of interest to see how these approaches could profitably be combined with the approaches we discuss here, to yield models of how word-learning draws on both observed examples and patterns of linguistic usage, but that is beyond the scope of the present work.
2. A notable exception is the cluster corresponding to trucks (T), which is barely separated from the next highest cluster (V) that contains the trucks plus a long yellow school bus. Cluster V itself *is* fairly well-separated from the next highest cluster, suggesting that the perceptually basic category here is not quite trucks but something more like "truck-shaped motor vehicles".
3. We should note that our use of the term "basic-level bias" differs from many uses in the literature. Typically it is unclear whether a putative word-learning bias, such as a "basic-level bias", refers to a behavioral tendency or to an aspect of mental representation: a greater prior degree of belief in some concepts (e.g., basic-level kinds) as candidate word meanings. Our interest primarily concerns the latter, and we would like to reserve the term "bias" for that sense, but empirical studies have mostly focused on the former. It is an empirical phenomenon, demonstrated in previous studies (Callanan et al., 1994; Waxman, 1990) as well as in our studies here, that generalization of a taxonomic label from a single example appears to follow a gradient falling off around the basic level. That is, children or adults tend to extend a novel label almost always to new objects matching at the subordinate level, much of the time (between 40% and 80% in our studies) to objects matching only at the basic level, and rarely to objects

matching at only the superordinate level. Instead of referring to this behavioral tendency as a "basic-level bias", we will refer to it as "one-shot basic-level generalization," to distinguish it from possible cognitive structures that might be proposed to account for it.

4. All correlation ($r$) values in this section are computed using only judgments for test items within the same superordinate class as the observed examples. Participants almost never chose test items that crossed superordinate boundaries, and most models give these test items zero or near-zero probability of generalization.
5. Figure 9b shows the median pattern of generalization over the three superordinate categories, rather than the mean, because the MAP generalizations are always either 0 or 1 and thus the mean is sometimes not representative of the model's all-or-none predictions.

**Figure Captions**

*Figure 1.* The extensions of words that label object kind categories may overlap in a nested fashion, in accord with the tree-structured hierarchy of an object-kind taxonomy.

*Figure 2.* The extensions of words that label object shape and substance categories may overlap in a cross-cutting fashion, because these two dimensions of object appearance are approximately independent.

*Figure 3*. Twelve training sets of labels objects used in Experiment 1, drawn from all three domains (animals, vegetables, and vehicles) and all four test conditions (1-example, 3-example subordinate, 3-example basic-level, and 3-example superordinate).

*Figure 4*. The test set of 24 objects used to probe generalization of word meanings in Experiment 1. For each training set in Figure 3, this test set contains 2 subordinate matches, 2 basic-level matches, and 4 superordinate matches.

*Figure 5*. Adults' generalization of word meanings in Experiment 1, averaged over domain. Results are shown for each of four types of example set (1 example, 3 subordinate examples, 3 basic-level examples, and 3 superordinate examples). Bar height indicates the frequency with which participants generalized to new objects at various levels. Error bars indicate standard errors.

*Figure 6*. Children's generalization of word meanings in Experiments 2 and 3, averaged over domain. Results are shown for each of four types of example set (1 example, 3 subordinate examples, 3 basic-level examples, and 3 superordinate examples). Bar height indicates the frequency with which participants generalized to new objects at various levels. Error bars indicate standard errors.

*Figure 7*. Hierarchical clustering of similarity judgments yields a taxonomic hypothesis space for Bayesian word learning. Letter codes refer to specific clusters (hypotheses for word meaning), as discussed in the text.

*Figure 8.* Predictions of the Bayesian model, both with and without a basic-level bias, compared to the data from adults in Experiment 1 and those from children in Experiment 3.

*Figure 9*. Predictions of two variants of the Bayesian model. (a) Without the size principle, Bayesian generalization behaves like an exemplar-similarity computation. (b) Without hypothesis averaging, Bayesian generalization follows an all-or-none, rule-like pattern.

*Figure 10*. Predictions of four alternative, non-Bayesian models.

Figure 1

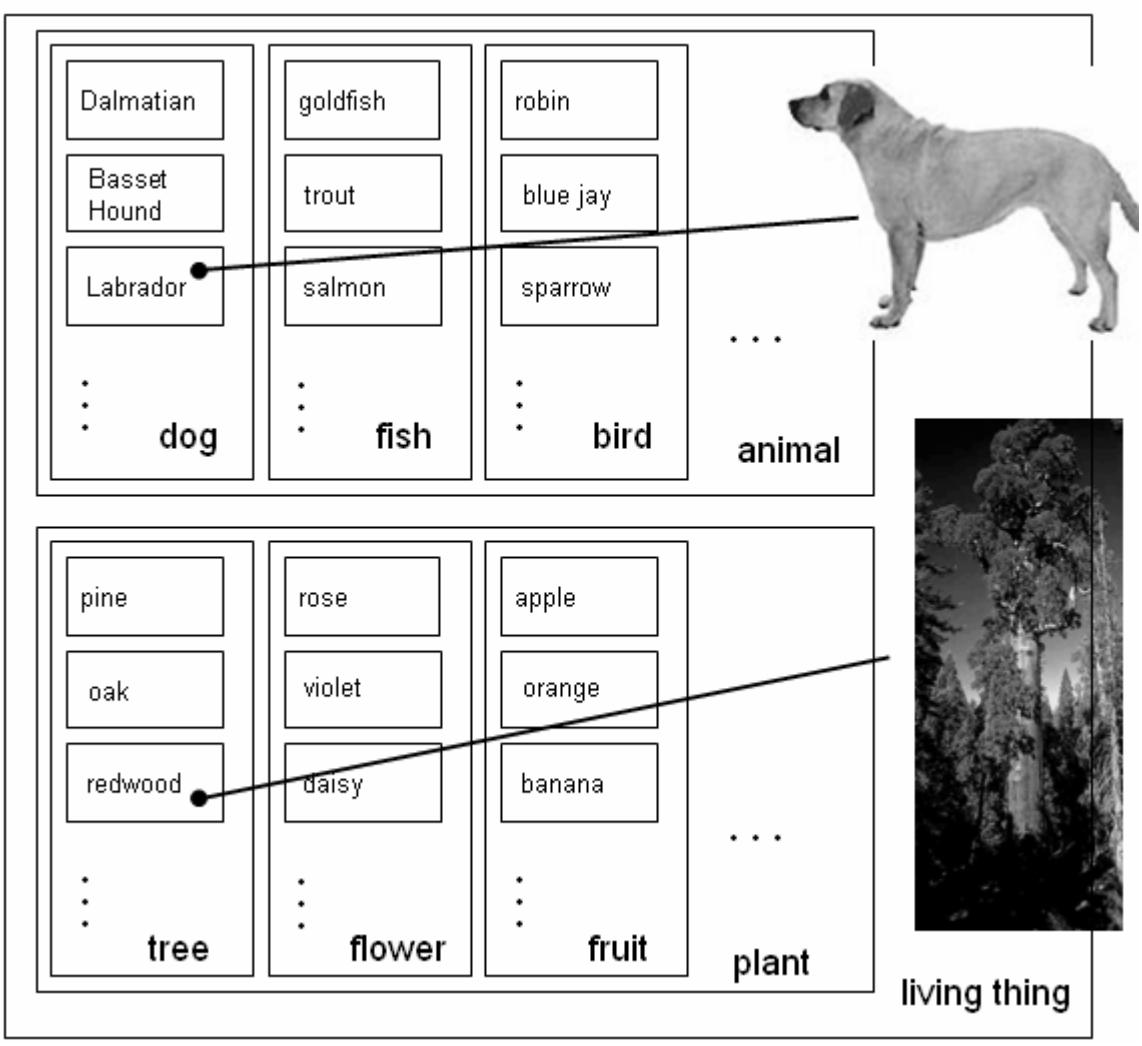

Dalmatian
Basset Hound
Labrador
goldfish
trout
salmon
robin
blue jay
sparrow
dog
fish
bird
animal
pine
oak
redwood
rose
violet
daisy
apple
orange
banana
tree
flower
fruit
plant
living thing

Figure 2

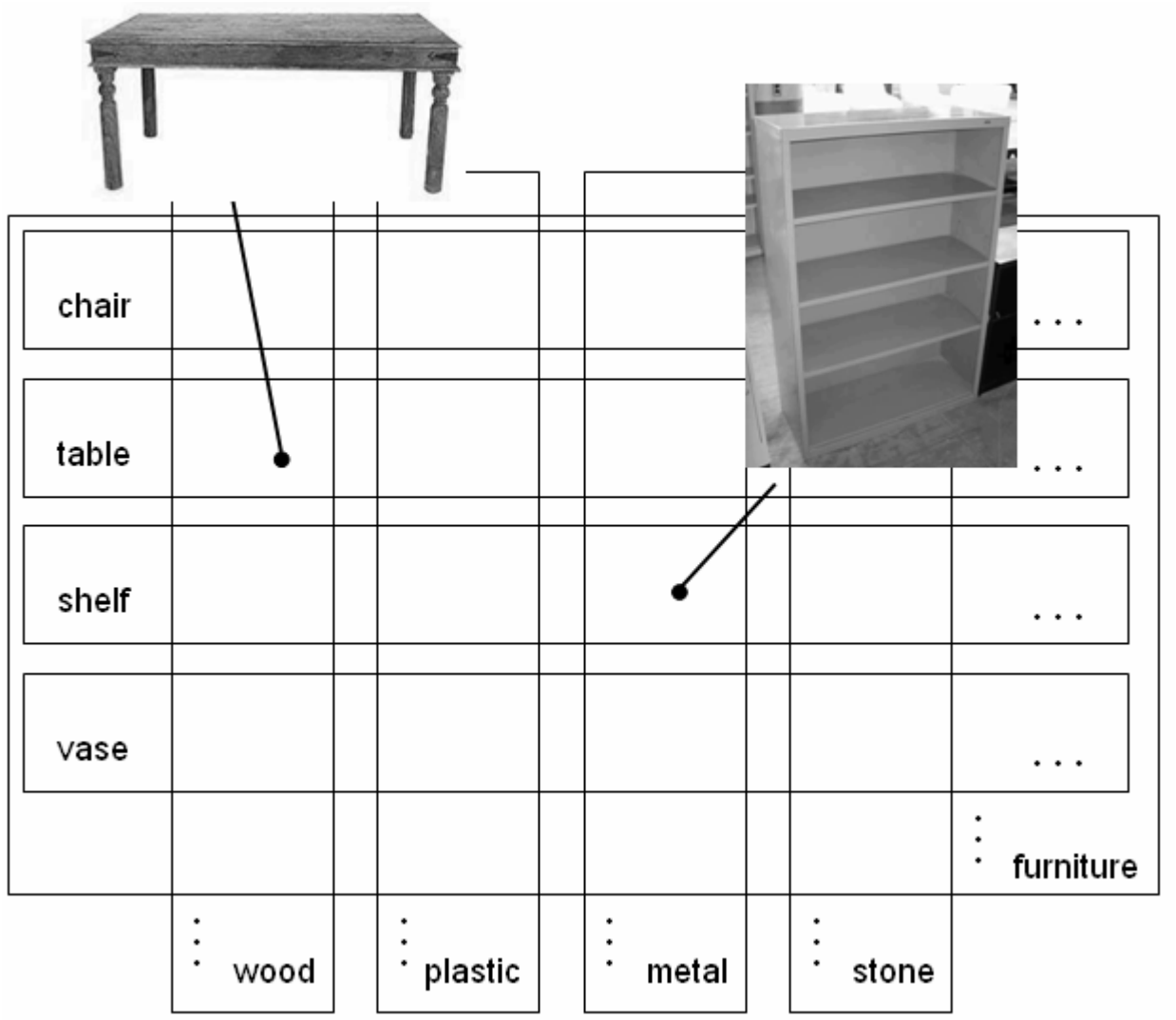


Figure 3

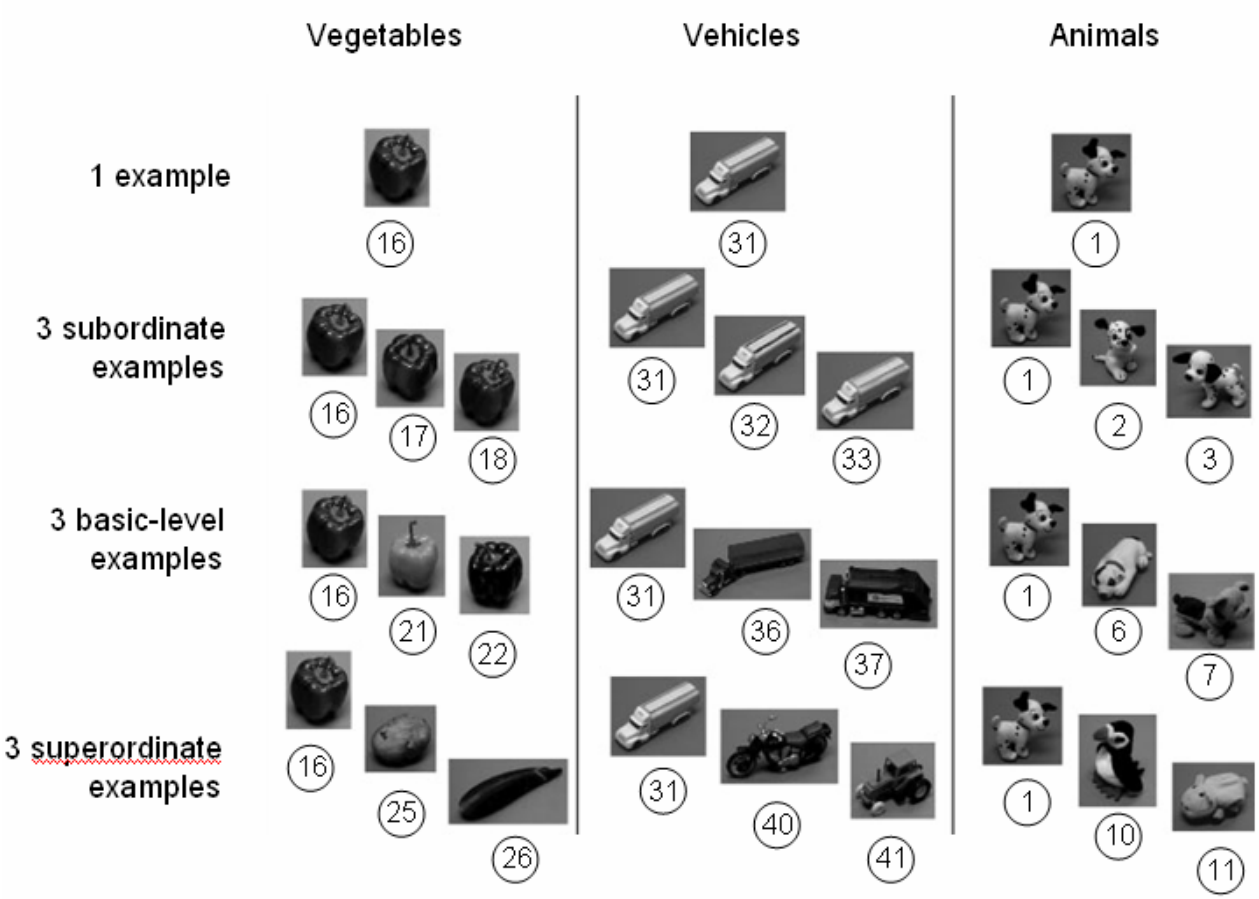

Vegetables
Vehicles
Animals
1 example
3 subordinate examples
3 basic-level examples
3 superordinate examples
16
31
1
16
17
18
31
32
33
1
2
3
16
21
22
31
36
37
1
6
7
16
25
26
31
40
41
1
10
11

Figure 4

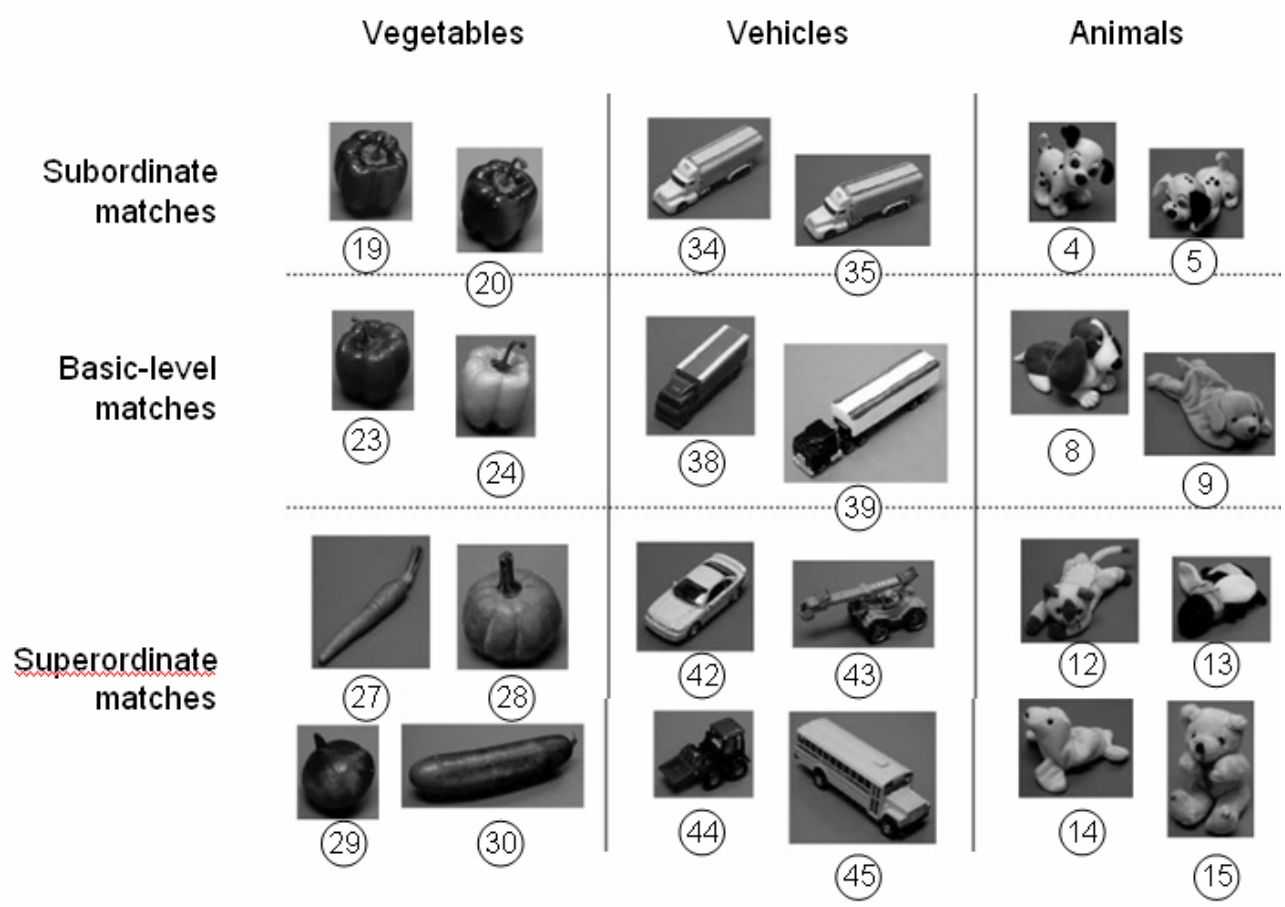

Vegetables
Vehicles
Animals
Subordinate matches
Basic-level matches
Superordinate matches
19
20
34
35
4
5
23
24
38
39
8
9
27
28
42
43
12
13
29
30
44
45
14
15

Figure 5

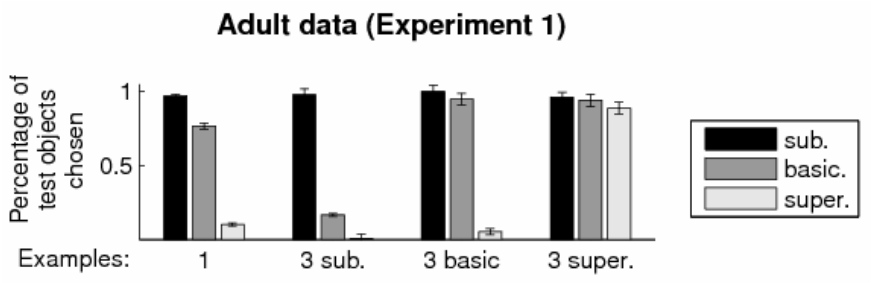

Adult data (Experiment 1)
Percentage of test objects chosen
1
0.5
Examples:
1
3 sub.
3 basic
3 super.
sub.
basic.
super.

Figure 6

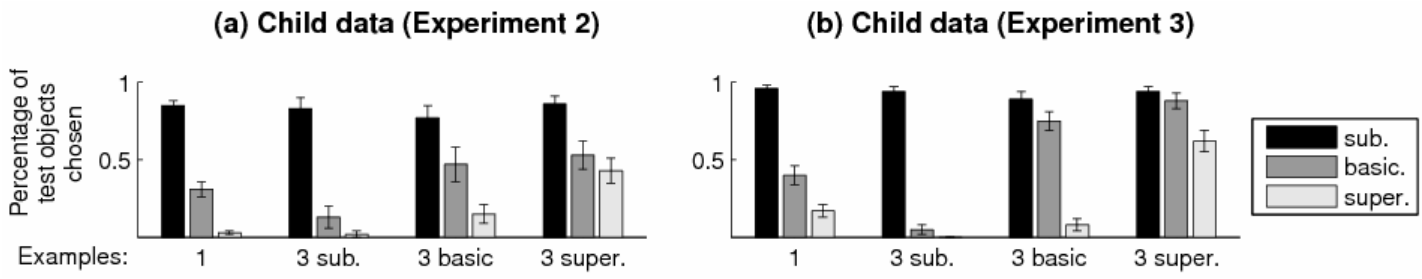

(a) Child data (Experiment 2)
(b) Child data (Experiment 3)
Percentage of test objects chosen
1
0.5
Examples:
1
3 sub.
3 basic
3 super.
sub.
basic.
super.

Figure 7

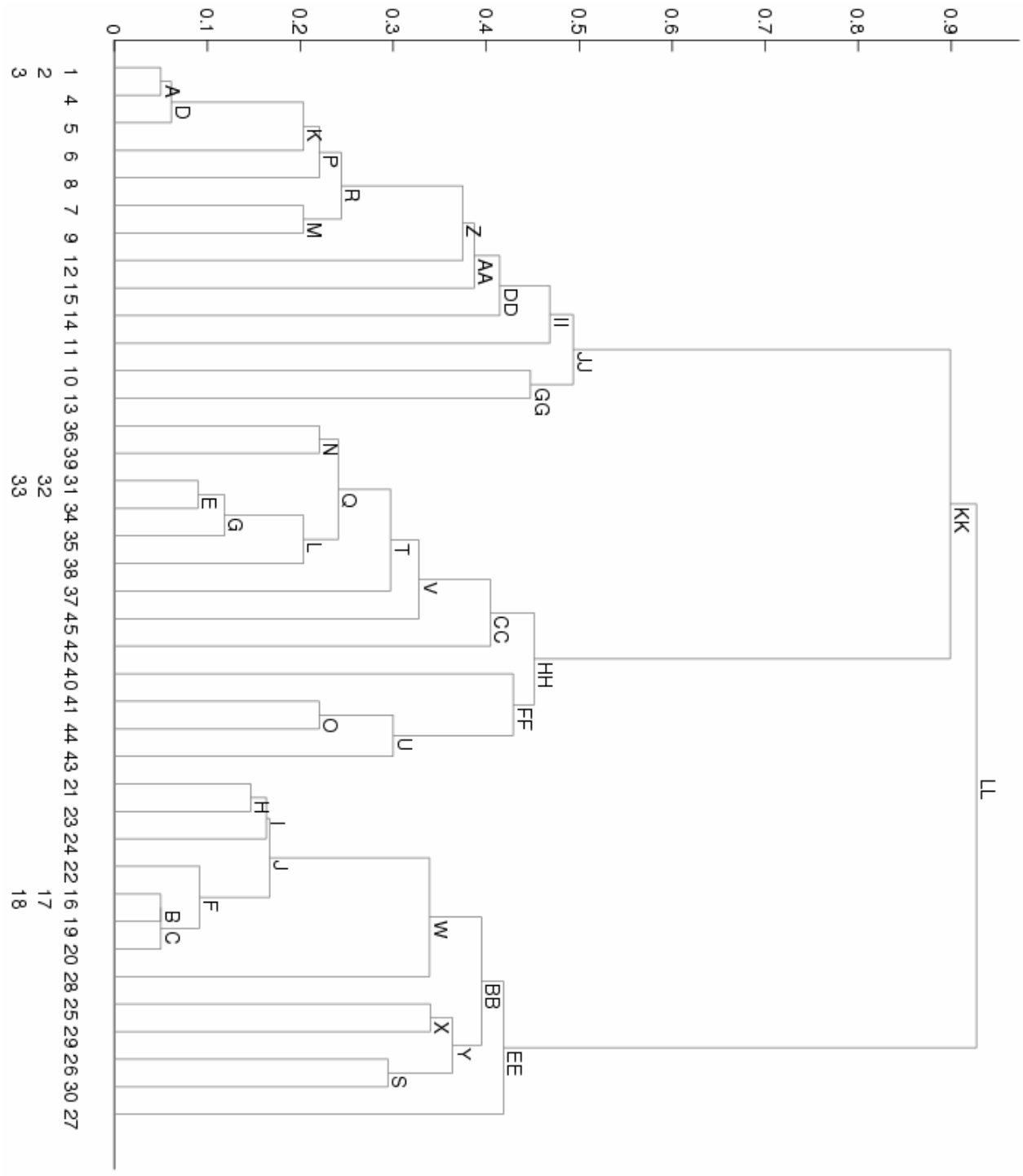

0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1 4 5 6 8 7 9 12 15 14 11 10 13 36 39 31 34 35 38 37 45 42 40 41 44 43 21 23 24 22 16 19 20 28 25 29 26 30 27
2
32
3
33
17
18
A
D
K
P
R
M
Z
AA
DD
II
JJ
GG
N
Q
E
G
L
T
V
CC
HH
FF
O
U
H
I
J
F
B
C
W
BB
X
Y
S
EE
KK
LL

Figure 8

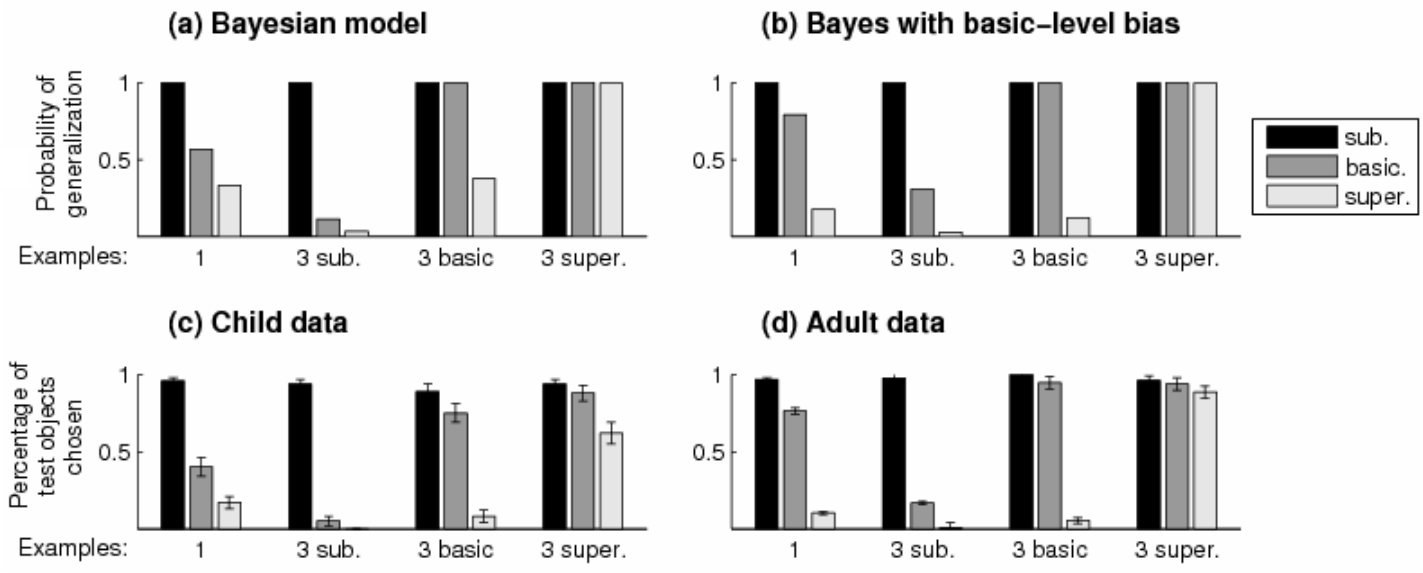

(a) Bayesian model
(b) Bayes with basic−level bias
(c) Child data
(d) Adult data
Probability of generalization
Percentage of test objects chosen
1
0.5
Examples:
1
3 sub.
3 basic
3 super.
sub.
basic.
super.

Figure 9

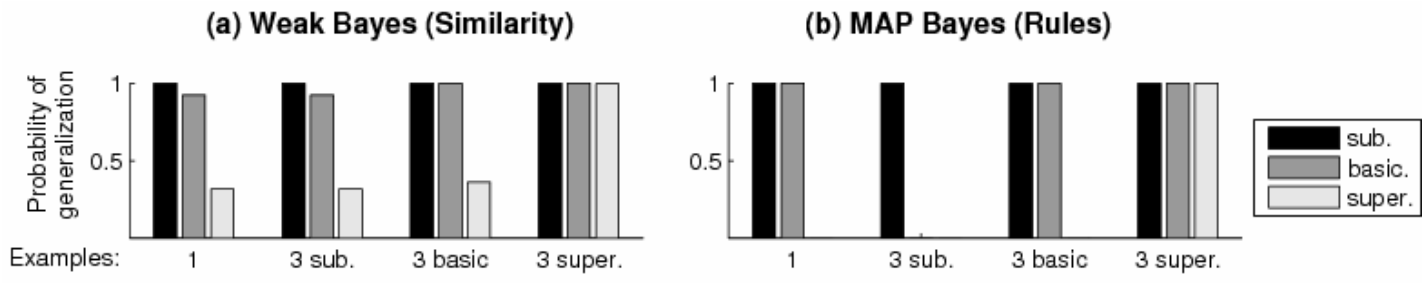
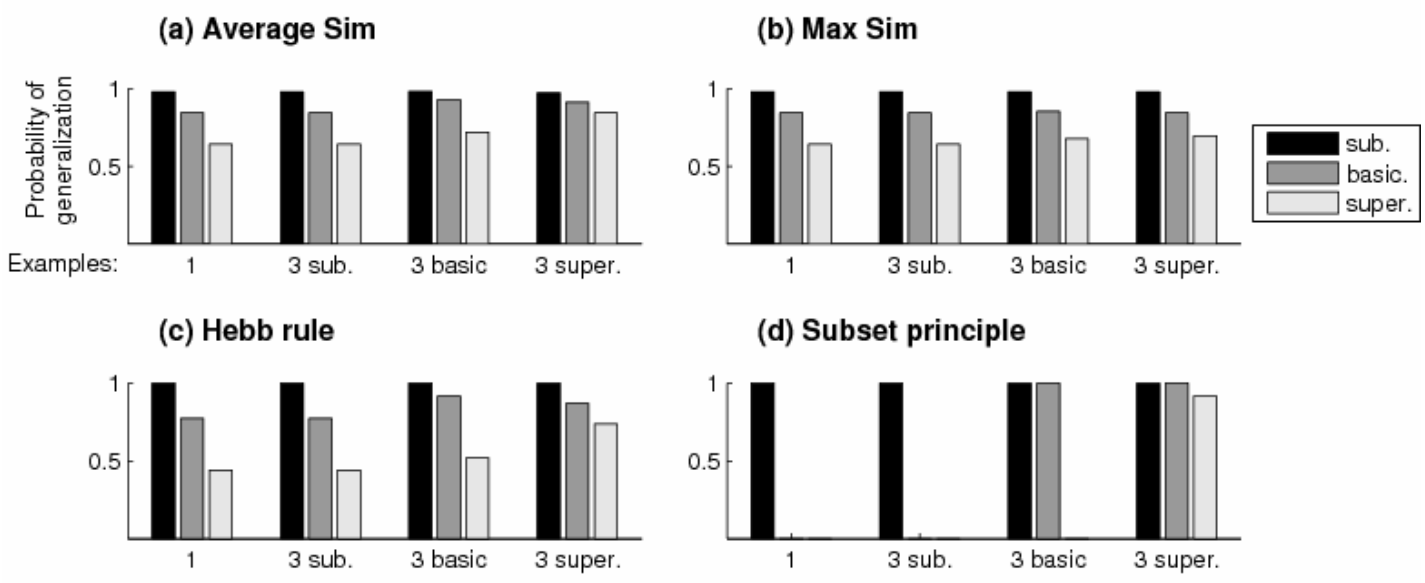

(a) Weak Bayes (Similarity)
(b) MAP Bayes (Rules)
Probability of generalization
1
0.5
Examples:
1
3 sub.
3 basic
3 super.
sub.
basic.
super.

Figure 10


(a) Average Sim
(b) Max Sim
(c) Hebb rule
(d) Subset principle
Probability of generalization
1
0.5
Examples:
1
3 sub.
3 basic
3 super.
sub.
basic.
super.