An integrated model of concept learning and word-concept mapping

Molly Lewis
mll@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

To learn the meaning of a new word, children must solve two distinct problems: identify the referent under ambiguity and determine how to generalize that word's meaning to new objects. Traditionally, these two problems have been addressed separately in the literature, despite their tight relationship with one another. We present a hierarchical Bayesian model that jointly infers both the referent of a word in ambiguous contexts and the concept associated with a word. As a first step in testing this model, we provide evidence that our model fits human data in a simple cross-situational concept learning task.

Keywords: cross-situational word learning; Bayesian models

Introduction

Learning a new word requires drawing a link in your mental lexicon between a word and a concept. But, children do not observe associations between words and abstract concepts; they observe associations between words and exemplars of those concepts. Furthermore, the associations between words and objects are ambiguous: a single word uttered in any particular context is consistent with an infinite number of possible interpretations (Quine, 1960). There are thus two problems a child must solve in order to learn the meaning of a new word: Determine which object is referred to by a word in context (the Mapping Problem) and determine the relevant concept of the object (the Generalization Problem; see Figure 1).

To understand these two problems more clearly, suppose you lived in an (impoverished) world with two words, "apple" and "cherry," and three objects, a green apple, a red apple, and a cherry. You hear the word "apple" in the context of a single red apple on the table. You somehow infer that "apple" refers to the red object on the table, and thus correctly solve the Mapping Problem. But you have not yet succeeded in solving the Generalization Problem. To correctly solve the Generalization Problem, you must decide whether "apple" also refers to the green apple, which is similar in shape to your observed apple exemplar, or whether it also refers to the cherry, which is similar in color to your observed apple exemplar. Or, alternatively, whether "apple" refers to neither of these other objects (i.e. a proper name). Thus, to learn the word "apple" in this world, you must infer both that "apple" refers to the red object on the table, and that "apple" should be generalized to other apple-shaped objects.

Separate learning mechanisms and constraints have been proposed to account for each of these problems. In the case of the Mapping Problem, one proposed constraint is cross-situational statistics (Pinker, 1984; Smith & Yu, 2008; Yu & Smith, 2007). Under this account, learners are hypothesized to aggregate the statistics of associations between words and objects across situations. When considered in an isolated situation, the referent of a word may be ambiguous, but when situations are aggregated across, the learner is able to constrain the hypothesis space of likely meanings. There is evidence that children as young as 12-months-old can learn word meanings in this way (Smith & Yu, 2008).

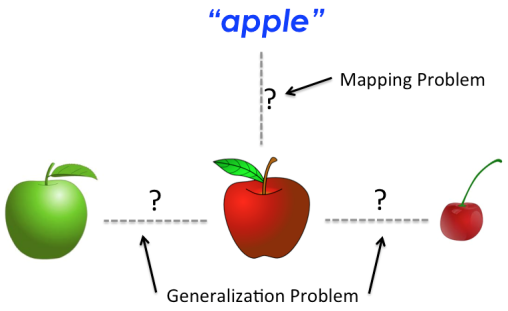

Figure 1: Schema of the two problems associated with learning the meaning of a word. Learning a new word requires that the child both identify which object the word refers to in the referential context (the Mapping Problem) and how to generalize that word to objects of the same kind (the Generalization Problem).

A second class of constraints on the Mapping Problem are accounts of the disambiguation effect. The disambiguation effect refers to children's tendency to select a novel, as opposed to familiar, object as a referent for a novel word. One account of this phenomenon is the principle of mutual exclusivity (Markman & Wachtel, 1988; Markman, Wasow, & Hansen, 2003). Under this proposal, there is a constraint on the types of lexicons considered when learning the meaning of a new word. With this constraint, children are biased to consider only those lexicons that have a one-to-one mapping between words and objects. Thus, when faced with an ambiguous referential context, the child solves the mapping problem by assuming that the novel word refers to the object for which she does not yet have a word in her lexicon. This is the inferred mapping because it is the only referent that allows the learner to maintain a one-to-one structure between words and concepts in the lexicon. Others have proposed that general pragmatic assumptions can also account for this effect (Clark, 1987; Diesendruck & Markson, 2001).

There are also a range of proposals about how children might solve the Generalization Problem. One proposal is that children have a bias to generalize by shape (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). With this

bias, a child who has learned that "apple" maps to apple, for example, can generalize "apple" to all apple-shaped things. This bias allows learners to rule out alternative, less probable generalizations strategies, such as generalization along the dimension of color. A second proposal is that children have a bias to generalize to another object of similar kind, rather than to one that is thematically related ("Taxonomic Assumption"; Markman, 1990). For example, upon hearing the word "cherry," a child with this bias would be more likely to generalize the word to another fruit, as opposed to ice cream, despite the fact the ice cream and cherries often go together (see Xu & Tenenbaum, 2007, for a probabilistic view).

Though theoretically distinct, and investigated separately, these the two problems are intimately related. If a child has solved the Generalization Problem for a particular category, the Mapping Problem becomes much easier. For example, suppose a child is faced with a never-before-seen apple and a novel object, and hears the word "dax". If the child has solved the Generalization Problem, the child can identify the apple as an exemplar of the APPLE[1] concept, and determine the correct referent by mutual exclusivity. Conversely, if a learner can easily solve the Mapping Problem, the learner will accumulate more correct exemplars of a category, and thus be more likely to infer the correct concept. Thus, existing proposals about how each of these problems is solved takes the other problem for granted. But, importantly, a child acquiring language begins with neither of these problems solved; both must be solved in parallel. That is, a learner must determine both what object a word refers to, and how to generalize that meaning beyond the particular context. And, critically, she must do both at the same time.

There is limited work exploring how children might solve these two problems in parallel. A study by Akhtar and Montague (1999) begins to address this question by asking whether children might use cross-situational statistics to learn the relevant features for generalization. In their task, 2-4 year old children were presented with three novel objects that all shared a common feature (e.g. color), but varied along two other features (e.g. texture and shape). Children were able to correctly infer that the novel word referred to the shared feature. This result provides important evidence that children can infer word concepts cross-situationally. However, it is unclear whether this type of learning generalizes to the real world because the actual learning environment is not structured in a way that perfectly disambiguates word meanings cross-situationally.

Apart from word learning, the Generalization Problem has been well-studied in adults (Laurence & Margolis, 1999; Rosch & Mervis, 1975; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Medin & Ortony, 1989). However, limited research has attempted to extend this body of literature to work with children. One exception is work by Sloutsky and colleagues which adopts models of similarity to explain how children generalize novel words (e.g. Sloutsky, Lo, & Fisher, 2001).

[1]Small capital letters are used to distinguish concepts from objects.

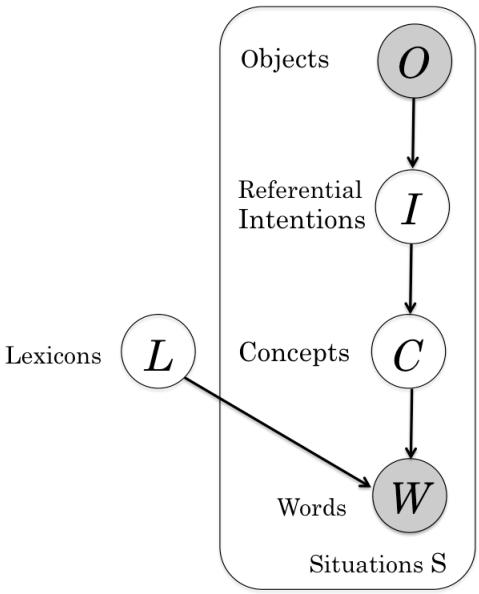

Figure 2: The generative process for our model. Shading indicates observed variables

We present a hierarchical Bayesian model that solves the Mapping and Generalization Problems in parallel. In modeling the Generalization Problem, we draw on the Boolean concept learning framework in which objects are defined by a set of features with a range of values (Shepard, Hovland, & Jenkins, 1961). The goal for the word learner is conceptualized as the task of mapping a word to a set of features that define the relevant concept. In modeling the Mapping Problem, we focus on the role of cross-situational statistics. In particular, we build on the model developed by Frank, Goodman, and Tenenbaum (2009) that takes into account the intentions of the speaker in order to identify the referent in ambiguous contexts.

The plan for the paper is as follows. We first describe the design of this extended model, and then describe the results of an experiment that explores adult performance in a cross-situational Boolean concept learning task.

Design of the Model

The goal of our model is to understand how children arrive at an understanding about the meanings of words, on the basis of limited evidence about the associations between words and objects. That is, the goal is to infer a lexicon — a set of word-concept mappings — on the basis of basis of observations of words and objects. To model this, we consider a set of variables relevant to this learning problem, and assume that they are related probabilistically. We assume an identical dependency structure as the model developed by Frank et al. (2009), with the addition of a concept layer to the generative process (see Figure 2). This model is the same underlying

model as presented in Lewis and Frank (2013) but with the addition of a theory of Boolean concepts. For completeness, we present the full model here, but details are identical except where noted.

We model a word learner as performing Bayesian inference to infer a lexicon l, which we represent as a (sparse) bipartite graph connecting words $W = w_1 ... w_n$ to concepts $C = c_1 ... c_m$. Concepts are written as a vector of features with values 1, 2 or *. The $*$ notation denotes a feature that is irrelevant to the definition of a concept. For example, $[1 * *]$ represents a concept that is defined only by the value of the first feature. This hierarchical formulation of concepts is substantially similar to the concept learning model proposed by Goodman, Tenenbaum, Feldman, and Griffiths (2010). The full possible set of lexicons is denoted as L.

The learner infers a distribution over lexicons, given a corpus S of situations (each consisting of sets of words $\bar{w}_s$ and objects $\bar{o}_s$). From Bayes' rule, the posterior probability of a lexicon is given by

$$P(l|S) = \frac{P(S|l)P(l)}{\sum_{l' \in L} P(S|l')P(l')}. \quad (1)$$

The prior $P(L)$ is assumed to be uniform over lexicons that map a concept to at most one word (one word to many concepts). We now define the likelihood term $P(S|L)$.

Using the generative process in Figure 2, we can write the likelihood of a particular situation in terms of the relationship between the objects that were observed in the situation s, the speaker's referential intention i_s (a choice to speak about one of the objects), the concept c_s selected by the speaker to represent the intention, and the referring word w_s. As in our prior work, we assume that referential intentions are unobserved and sum across all possible intentions uniformly:

$$P(s|l) = \sum_{i_s \in \bar{o}_s} p(w_s, c_s, i_s, o_s, |l) \quad (2)$$

By the conditional independence of words and objects, we use the chain rule to expand to:

$$P(s|l) = \sum_{i_s \in \bar{o}_s} P(w_s|c_s, l)P(c_s|i_s)P(i_s|o_s) \quad (3)$$

Finally, we aggregate across situations by taking the product of each independent situation:

$$P(S|l) = \prod_{s \in S} \sum_{i \in \bar{o}_s} P(w_s|c_s, l)P(c_s|i_s)P(i_s|o_s) \quad (4)$$

To find the key term in our concept model, $p(c_s \mid i_s)$, we use a noisy Naive Bayes classifier:

$$P(c_s \mid i_s) = \prod_{j=1...f} \begin{cases} 1-\alpha & \text{if } (c_s^j = i_s^j) \vee (i_s^j = *) \\ \alpha & \text{otherwise} \end{cases} \quad (5)$$

This formulation quantifies the probability of a concept given an intended object in terms of the match between the three features.

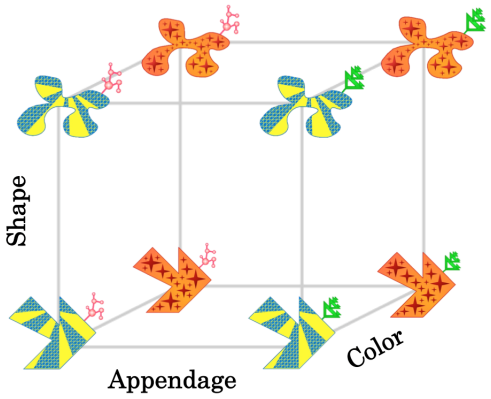

Figure 3: Experimental stimuli. Each object is defined by a binary value for each of three features: shape, appendage, and color.

We assume that there is some level of noise in both the choice of word given intention $P(w_s|i_s, l)$ and the choice of intention given object $P(c_s|i_s)$, such that the speaker could in principle have been mistaken about their referent or misspoken. We implement this decision by assuming a constant probability of random noise for each of these, which we notate α; for simplicity, α is assumed to be the same for both decisions. The particular choice of α values only serves to scale the predictions, and does not influence the relative predictions of the test item types. However, as in nearly all probabilistic models, some level of uncertainty about the individual observations is necessary to be able to make graded predictions.

In the simulations reported here, we did inference by exact inference via full combinatoric enumeration of the space of possible lexicons.

Experiment

Our model jointly solves the two problems associated with learning the meaning of a new word, the Mapping and Generalization Problems. As a first step in evaluating the model, we compared human and model performance in a cross-situational Boolean concept learning task. Participants were given a situation in which a word is seen in the context of two objects, but in a way that is ambiguous as to which of these objects (either or both) the label refers to. The learner is then presented with a second such situation. While each of these situations is individually ambiguous, the learner could aggregate information across situations to infer the concept associated with the word. As predicted by the model, we found that participants generalized the meaning of the label in a graded manner: the more features the training objects shared with the test object, the more likely participants were to generalize the label to the test object.

Method and Materials

Participants Two hundred and and sixty-six adults were recruited from Amazon's Mechanical Turk. Twenty-two were

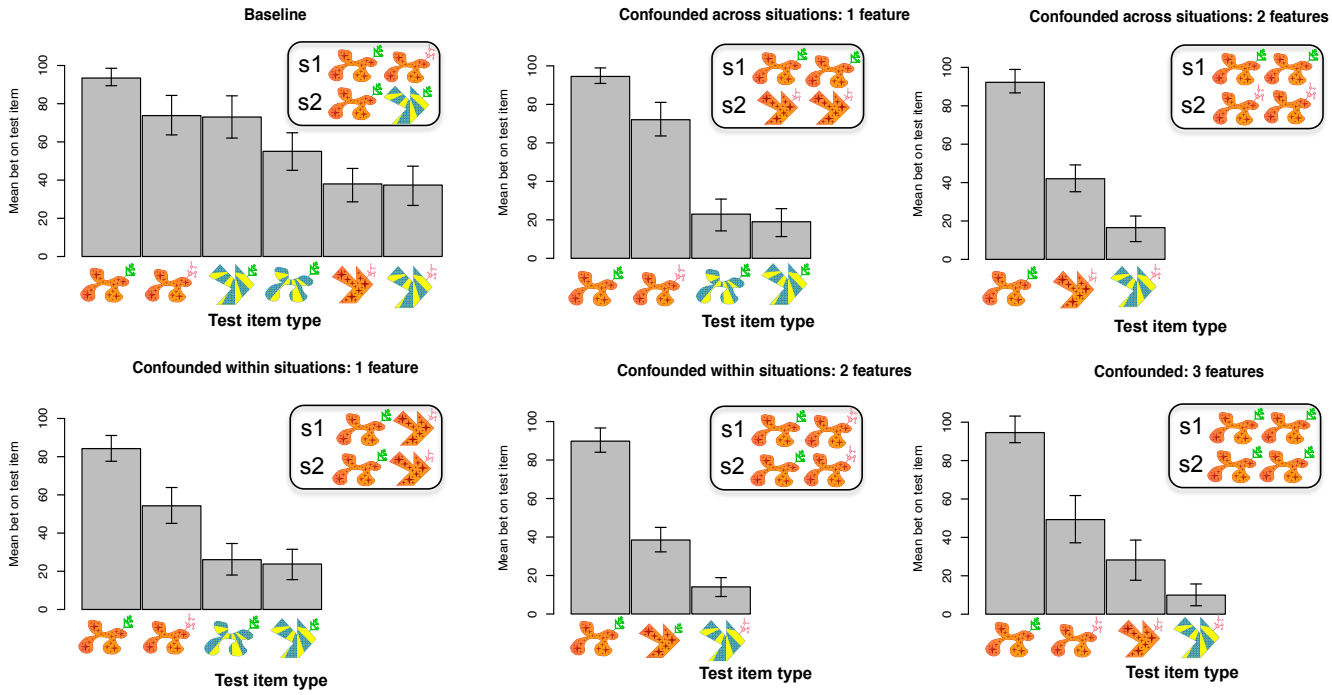

Figure 4: Bets on the probability of "dax bren nes" generalizing to each of the relevant test item types in each condition. Error bars represent 95% confidence intervals as computed via non-parametric bootstrap. Example items seen in the training situations are given in the top-right box of each plot. Given the particular training items shown here, an item in each of the relevant test item types (defined by the number of features shared with the training items) is shown along the x-axis. Actual training items (and thus test items) were counter-balanced across participants.

excluded for either not completing the task appropriately (e.g. by responding with values greater than 100) or failing to provide responses for all 8 test objects. All reported that they were native speakers of English.

Stimuli Each object in our stimulus set was constructed to have three binary features. The features of interest were shape, appendage and color. When fully permuted, this defines a space of eight possible objects (see Figure 3).

Procedure Participants viewed a webpage that showed two situations with two objects each. In the first situation, they were instructed: "Suppose you saw these two objects and heard 'dax bren nes.'" A multi-word novel label was used to avoid biases towards meanings consistent with the grammatical class of the word. In other words, we wanted to avoid participants inferring that because the word was an adjective, it was more likely to refer to a property (e.g. color) than a particular object (i.e. a proper noun), for example. Two more objects were presented below and participants were asked to "Now suppose you saw these two new objects and heard 'dax bren nes' again." They were then asked to "bet whether or not you think each of the objects below could also be called 'dax bren nes.'" Images of all eight objects (including the training items) were then presented, and participants were asked to provide a bet 0–100 indicating their judgement.

Across participants, we manipulated the number of features shared within and across situations.[2] We tested an unambiguous baseline condition in which the same object was paired with a different object in each situation and 5 ambiguous conditions in which the features of the objects were confounded either within or across situations. For the ambiguous conditions, we tested cases in which 1 or 2 features were shared within situations ("confounded within" conditions), 1 or 2 features shared across situations ("confounded across" conditions), and a case in which 3 features were shared both within and across situations (Figure 4).

[2]This manipulation was motivated by the observation that different types of ambiguity license different inferences. To illustrate this, imagine a learner in a confounded across context. The learner observes a situation with two apples and a situation with two oranges. In each situation, she hears "dax bren nes". The referent is clear in each individual situation — apple and orange, respectively — and the learner might infer that this phrase corresponds to a superordinate category, such as FRUIT. In the confounded within context, the learner observes two situations, both containing an apple and an orange, and again hears "dax bren nes" in each. Unlike in the across case, a learner in this context would have no information about how to correctly map the meaning of this phrase, since the context is consistent with both a subordinate and superordinate interpretation. Different generalization patterns are thus predicted in the confounded across and within conditions.

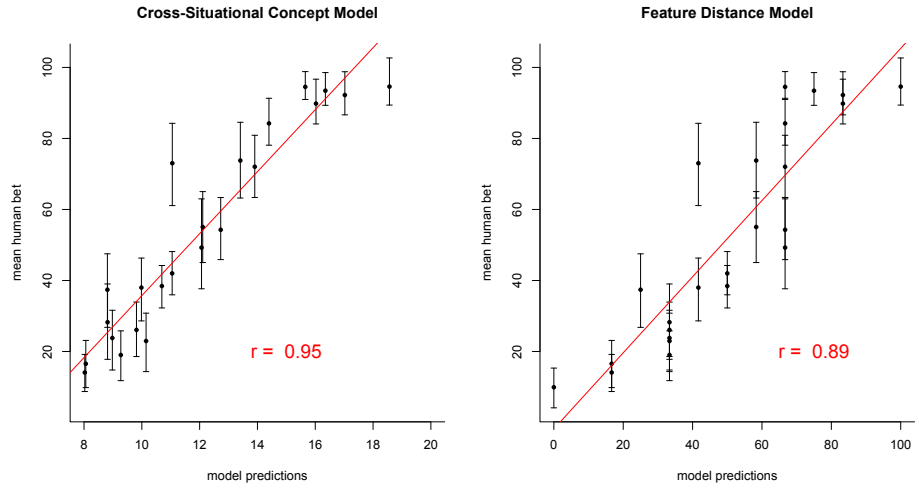

Figure 5: Mean bets in each experiment condition, plotted by predictions for each of the models (model predictions are scaled on the horizontal axis). Error bars represent 95% confidence intervals, as computed via non-parametric bootstrap. The line of best fit is plotted in red.

Results and Model Fits

Participants showed a consistent gradient of generalization such that greater number of distinct features resulted in lower bets (weaker generalizations), consistent with previous experiments (Figure 4).

Model fits are shown in Figure 5. Our model fits the data with a correlation of $r = .95$. We compared this fit against a null model in which we calculated the target's total feature distance from objects in the situations. This was calculated by counting the number of features for which the target differed from each situation object (e.g. the feature distance $[111]$ and $[122]$ is 2) and summing across all four objects in the situation. This standard exemplar style model fits the data relatively well ($r = .89$). Nevertheless, our model provides a substantial gain in fit. Using non-parametric bootstrap, the cross-situational concept-learning model fits the data significantly better than the feature distance model ($p < .05$).

General Discussion

In this cross-situational Boolean concept learning task, our model performed competitively with a simple feature distance model. Critically, however, our model has the machinery to solve not only this simple concept-mapping problem under minimal ambiguity, but can also deal with more complex worlds in which multiple words are present. Given that no existing model is able to jointly account for both the Mapping and Generalization Problems, this model provides a fruitful theoretical tool for future work to explore how children might solve these problems.

For example, this experiment could be straight-forwardly extended to introduce a more complex Mapping Problem component to the task. This could be done by adding additional words to the cross-situational learning context. In a minimal version of this experiment, the learner could observe w_1 with $[11]$ and $[12]$ and w_2 with $[22]$ and $[12]$. A learner who assumes that the speaker refers to both objects within each situation, might infer a mapping between w_1 with $[1*]$ and a mapping between w_2 with $[*2]$, given this referential context. Using situations such as these, this paradigm can be extended to directly explore joint inference of both the Mapping and Generalization problems.

An important underlying assumption of this model is that features are given *a priori*. This seems like an extreme position given that it is implausible that children acquiring language have an innate "appendage" feature, for example. It is, in a sense, the very goal of this model to explain how children acquire such abstract concepts as APPENDAGE. That is, features are themselves concepts that can be considered as primitives in the construction of more complex concepts. This problem, however, is not specific to the word learning problem, but rather is a challenge more generally to the Boolean concept learning framework. Nonetheless, a complete theory of how children acquire word concepts will need to provide an account for the origin of features.

Given this theoretical point, our model should be understood as a computational level description of the problem of acquiring word-concepts, given some set of concepts (i.e. features). Our model remains agnostic about the origins and nature of these initial concepts but, given some primitive set of concepts, our model describes how a learner might bootstrap from these primitives to infer more and more complex concepts. While it seems unlikely that children have an innate APPENDAGE feature, there is evidence that children may have certain perceptual categories, such as color, very early in development (Bornstein, Kessen, & Weiskopf, 1976). Primitive perceptual features like color categories may provide the

initial building blocks for the construction of more complex concepts, given experience with the environment.

In sum, our model provides a rich framework for studying the word learning problem at the computational level. Previous research has explored how children might solve the two subproblems associated with word learning — the Generalization and Mapping Problems — separately. Our model contributes to this area by providing a unifed account for both of these problems. The experiment reported here suggests that our model is able to account for participants' behavior in solving one of these problems — the Generalization Problem — in a simple cross-situational task. Importantly, our model's contribution to theories of the Generalization Problem is to provide an account of the generalization inferences, given an initial set of primitive concepts. This account, coupled with the ability to explore the Mapping Problem, lays the groundwork for a more cohesive understanding of how children learn the meanings of words.

Acknowledgements

We thank Mia Kirkendoll for her assistance in data collection.

References

Akhtar, N., & Montague, L. (1999). Early lexical acquisition: The role of cross-situational learning. *First Language*, *19*(57), 347–358.

Bornstein, M., Kessen, W., & Weiskopf, S. (1976). Color vision and hue categorization in young human infants. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(1), 115.

Clark, E. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition. Hillsdale, NJ: Erlbaum*.

Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. *Developmental Psychology*, *37*(5), 630.

Frank, M. C., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2010). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Laurence, S., & Margolis, E. (1999). Concepts: Core readings. In E. Margolis & S. Laurence (Eds.), (chap. 1). MIT Press.

Lewis, M., & Frank, M. C. (2013). Modeling disambiguation in word learning via multiple probabilistic constraints. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.*

Markman, E. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*(1), 57–77.

Markman, E., & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Markman, E., Wasow, J., & Hansen, M. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*(3), 241–275.

Medin, D., & Ortony, A. (1989). Psychological essentialism. *Similarity and Analogical Reasoning*, 179–195.

Pinker, S. (1984). *Language learnability and language development, with new commentary by the author* (Vol. 7). Harvard University Press.

Quine, W. (1960). *Word and object* (Vol. 4). The MIT Press.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Shepard, R., Hovland, C., & Jenkins, H. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Sloutsky, Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference. *Child Development*, *72*(6), 1695-1709.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13-19.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.