A Bayesian Framework for Learning Words From Multiword Utterances

Stephan C. Meylan (smeylan@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Abstract

Current computational models of word learning make use of correspondences between words and observed referents, but as of yet cannot—as human learners do—leverage information regarding the meaning of other words in the lexicon. Here we develop a Bayesian framework for word learning that learns a lexicon from multiword utterances. In a set of three simulations we demonstrate this framework's functionality, consistency with experimental work, and superior performance in certain learning tasks with respect to a Bayesian word leaning model that treats word learning as inferring the meaning of each word independently. This framework represents the first step in modeling the potential synergies between referential and distributional cues in word learning.

Keywords: word learning; Bayesian inference; artificial language learning; distributional learning

Introduction

Among the many feats that comprise first language learning, discovering the meaning of many tens of thousands of words is among the most impressive. Indeed the size and richness of human vocabularies is one of the major points of distinction between the linguistic capacities of humans and those of non-human primates (Pinker & Jackendoff, 2005). Learners start early on this task: long before they utter their first words, toddlers develop a substantive receptive vocabulary (Bergelson & Swingley, 2012). How precisely young learners assemble this knowledge so quickly remains an active area of investigation.

One possibility is that learners are able to concurrently use both correspondences between 1) words and referents and 2) words and other words in order to formulate and assess hypotheses regarding word meaning. For example, consider the two scenes and utterances with five novel words presented in Figure 1. In this example, a learner could use the reliable co-occurrence of *garp* and a particular referent (the depicted animal) across the two scenes to infer its meaning. Having a reasonable hypothesis regarding the meaning of *garp* in turn enables several consequent inferences on the basis of structural regularities in English. Establishing that *garp* is a referential entity (a noun within the adult syntactic system, though the child learner may have somewhat different provisional lexical categories) means that both utterances are consistent with the pattern *<referential entity> is X-ing*, in which *X* describes some activity for that referential entity. Another regularity in English suggests that in *the Y <referential entity>*, *Y* is probably a word that describes that following referential entity; hence, the color of the animal in each scene is a good candidate for the meanings of the words *wub* and *zek*. In this way, learning the meaning of a single word may result in a cascade of further word learning.

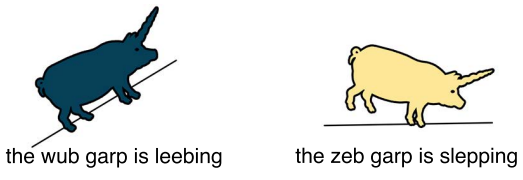

Figure 1: Referential word learning (in this case "garp") helps the learner identify additional regularities, which in turn support further word learning.

Existing word learning models are well-suited to explain how a learner might infer the meaning of the word "garp" in the above scenes. Learners may use hypothesis elimination (Siskind, 1996) or more graded co-occurrence information (Smith & Yu, 2008) to discover the regular mapping from word to referent or concept. Bayesian models are particularly powerful in that they can use implicit negative evidence for this purpose. For example, Xu and Tenenbaum (2007) showed that kids can learn words related by a taxonomic hierarchy in which a hypernym like "animal" is never incorrect for referring to a category member like a cat. Such models also provide a formal framework for the integration of non-linguistic cues in word learning (Frank, Goodman, & Tenenbaum, 2008), as well as additional category information (e.g. a property-vs.-kind distinction) that learners may bring to the problem (Gagliardi, Bennett, Lidz, & Feldman, 2012).

In contrast with the above models, distributional models are naive as to the correspondence between a word like "garp" and entities or states in the world, and instead proceed from the observation that the co-occurrence statistics of words—even in absence of referents—can encode rich information about latent structure in language. As implicated in the above example, a word's immediate context (previous word and following word) constitutes strong evidence of its grammatical category (Mintz, 2003). Other models such as the connectionist network of Elman (1990) and the technique of Latent Semantic Analysis in Landauer and Dumais (1997) show how relationships of synonymy can be extracted from large corpora by means of dimensionality reduction.

In the present work we examine how learners may use information regarding other words in the lexicon to guide the learning of word-to-referent mappings. We begin by outlining a word learning model that learns the referent of a single word, then show how this procedure can be generalized for learning the referents of many different words concurrently from multiword utterances.

Modeling Framework

To introduce our modeling framework, we first summarize a previous Bayesian word learning model and then generalize it to multiword utterances.

Bayesian Word Learning

The Bayesian word learning model introduced by Xu and Tenenbaum (2007) focused on the learning of nouns. The learner observes a particular object x being given a word label w, and considers hypotheses h that correspond to the sets of objects that could be given that label. The posterior probability of each h is given by

$$P(h|x,w) = \frac{p(x|h,w)p(h)}{\sum_{h' \in \mathcal{H}} p(x|h,w)p(h')}, \quad (1)$$

corresponding to the normalized product of the *likelihood* $p(x|h,w)$ and the prior $p(h)$.

The likelihood term $p(x|h,w)$ reflects whether the observed concept x is in the set $S_w^{(h)}$ identified by word w given hypothesis h,

$$p(x|h,w) = \begin{cases} \frac{1}{|S_w^{(h)}|} & \text{if } x \in S_w^{(h)} \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

The model's likelihood employs the reciprocal of the set size picked out by the current word – the *size principle* – corresponding to assuming that objects are sampled uniformly at random. The model can accommodate multiple independent observations, in this case an ordered set of objects $\mathbb{X}$ and an ordered set of words $\mathbb{W}$, by modifying the likelihood to become

$$P(\mathbb{X}|h,\mathbb{W}) = \prod_{i=1} p(x_i|h,w_i). \quad (3)$$

The prior reflects the expectations of the learner about which hypotheses are more likely to be true. The simplest prior is one in which each hypothesis regarding the word-to-concept mapping (the power set of concepts) is considered equally likely, $p(h) = 1/2^s$, where s is the size of the hypothesis space which that word could refer to.

Putting these pieces together, the probability that the word applies to a new object is given by

$$P(y \in S_w|x) = \sum_{h: y \in S_w^{(h)}} p(h|x,u), \quad (4)$$

being the sum of the posterior probabilities of those hypotheses under which y would be a member of the corresponding set of objects.

Multiword Utterances

We now generalize this model for learning the individual word-to-referent mappings for nouns to learning several word-to-referent mappings for different classes of words concurrently from a set of utterances. This requires changing two features of the modeling approach. First, rather than individual words referring to sets of objects, we treat each word as referring to a subset of possible states of the world, or *world-states*. In a world in which there is an object that is 1) either red or black and 2) either round or square, there would be exactly four possible world-states. Second, we treat the referential content of an utterance as the set of world-states picked out by some compositional function operating over the relevant word-to-referent mappings in the lexicon. By delineating both word and utterance meaning in terms of sets, the model supports unintuitive—though logically possible—meanings for both. The framework then treats the problem of word learning as one in which the learner must find the best lexicon to explain a set of observed world-states and corresponding utterances.

A lexicon $\mathbb{H}$ consists of one or more word-level hypotheses $\{h_1, \ldots, h_n\}$, each of which is a mapping from a word w to a set of world-states $\{x_n, \ldots, x_m\}$. The posterior probability of a lexicon given a set of utterances $\mathbb{U}$ and a set of observed scenes $\mathbb{X}$ can be calculated according to Bayes' rule:

$$p(\mathbb{H}|\mathbb{X},\mathbb{U}) = \frac{p(\mathbb{X}|\mathbb{H},\mathbb{U})p(\mathbb{H})}{\sum_{\mathbb{H}' \in \mathcal{H}} p(\mathbb{X}|\mathbb{H}',\mathbb{U})p(\mathbb{H}')}. \quad (5)$$

An observation from the above language consists of an utterance u and a world-state x. Assuming the conditional independence of the observed utterance/world-state pairs, the likelihood for a lexicon is the product of the probabilities of observing the world-state x_i for the corresponding utterance u_i for a given hypothesized lexicon $\mathbb{H}$:

$$p(\mathbb{X}|\mathbb{H},\mathbb{U}) = \prod_{i=1} p(x_i|\mathbb{H},u_i). \quad (6)$$

The likelihood term reflects whether the world-state x_i can be referred to by utterance u_i under the lexicon $\mathbb{H}$. If the world-state x is in with the set of world-states picked out by the utterance give the current lexicon, the likelihood is calculated as the reciprocal of the number of world-states that are picked out. Otherwise, the likelihood term is near zero. To prevent overfitting, a small portion of the probability mass (ε) is spread evenly across all hypotheses, yielding

$$p(x_i|\mathbb{H},u_i) = \begin{cases} (1-\varepsilon)\frac{1}{|S_{u_i}^{(\mathbb{H})}|} + \varepsilon\frac{1}{|S|} & \text{if any } x_i \in S_{u_i}^{(\mathbb{H})} \\ \varepsilon\frac{1}{|S|} & \text{otherwise} \end{cases}, \quad (7)$$

where $S_{u_i}^{(\mathbb{H})}$ is the set of world-states picked out by the utterance given the current lexicon. The framework is itself agnostic as to *how* the utterances and the lexicon pick out a particular set of world-states; depending on the assumptions about the semantics, the lexicon may specify different sets of world-states given an utterance. In Simulation 1 we describe one such function that picks out a particular set of world-states given an utterance and a lexicon. Rather that the exact form of this compositional function, the critical contribution of this framework is that of casting the problem of word learning as one in which *all* hypothesized word meanings that comprise a lexicon can be used in the assessment of the likelihood or prior for a particular word-to-referent mapping.

The prior probability of the lexicon, $p(\mathbb{H})$ is the product of the prior probabilities of the hypotheses h that comprise the lexicon $\mathbb{H}$, $\prod_{h \in \mathbb{H}} p(h)$. In the current case, the prior is uninformative: each mapping from a word to a set of world-states is equally likely. Here the prior $p(\mathbb{H}) = 1/2^{s \times n}$, where s is the number of world-states and n is the number of words in the lexicon. A more informative prior, such as a preference for cluster distinctiveness in taxonomic hierarchies (Xu & Tenenbaum, 2007) or a concept prior reflecting higher-level knowledge of word categories (Gagliardi et al., 2012), could also be implemented within this same framework.

The probability that a novel world-state y can be referred to by utterance u (consisting of one or more words) can be computed by generalizing Equation 4,

$$p(y \in S_u|u) = \sum_{\mathbb{H}:y\in S_u^{(\mathbb{H})}} p(\mathbb{H}|\mathbb{X},u), \quad (8)$$

being the sum of the posterior probabilities $p(\mathbb{H}|\mathbb{X},\mathbb{U})$ for all lexicons in which y is in the set of world-states picked out for utterance u.

Simulations

We present three simulations to demonstrate the function and utility of the new modeling framework. In Simulation 1, we show how the model finds the optimal lexicon in a toy world in which an utterance specifies a set of world-states via a simple compositional function. In Simulation 2, we show that the framework generates predictions that are consistent with experimental work in which adults learn the meaning of words from multiword utterances (Kersten & Earles, 2001). Finally, in Simulation 3 we describe a word learning task in which the new framework significantly outperforms the basic Bayesian word learning model.

Simulation 1: Word learning in a simple toy world

First, we demonstrate how the above framework allows us to learn the best lexicon for a simple toy language under the assumption of *intersective* semantics. Whereas the likelihood function in the simple Bayesian word learner only depends on whether a world-state is in the set picked out by a word, some compositional function is needed to pick out the set of world-states given an utterance. Here we assume that this set is specified by the intersection of world-states selected by the words that comprise that utterance:

$$S_{u_i}^{(\mathbb{H})} = \bigcap_{w\in u_i} S_w^{(h)}. \quad (9)$$

Other, more elaborate semantic functions may be substituted (e.g. a fully compositional semantics) within the framework; in the current case we use intersective semantics as the basis for a simple demonstration of the framework that can nonetheless capture aspects of previous work on artificial language learning.

In the toy world, world-states vary along three binary dimensions: an object is either a square or a circle (*pu* or *du*), which is either filled or unfilled (*li* or *ri*), and which moves either side-to-side or up and down (*wag* or *div*). There are thus eight possible states of the depicted in the scene, and eight utterances of three words length (e.g. the utterance *pu li wag* would be accompanied by a world-state of a black square moving side-to-side). A complete set of utterances and world-states are shown in Figure 2 along the vertical and horizontal axes respectively.

To demonstrate the operation of the model, consider the posterior probability of three different lexicons, each of which maps the word *wag* to a different set of world-states, after seeing eight sentences and corresponding world-states. While each lexicon has the same prior probability under the model, they are distinguished by their likelihood. The lexicon that posits that *wag* refers to objects that move up and down has a likelihood of 0 because that world-state is not seen consistently with that utterance. The lexicon that posits that *wag* refers to things that move side to side *and* those that are black receives a higher probability than the first lexicon because it is consistent with the observed data, but the likelihood is relatively low in that the hypothesis picks out a larger number of world-states. The lexicon that posits that *wag* refers to objects that move side to side receives the highest posterior probability, in that it is the most specific hypothesis that is consistent with the observed data. Probabilities of generalization for each utterance to each world-state are presented in Figure 2.

Simulation 2: Kersten and Earles (2001)

In the second simulation, we show how a model developed within the framework presented here is capable of learning word meanings in an existing artificial language learning paradigm. Kersten and Earles (2001) describe a set of experiments in which participants are presented with one- to three-word utterances coupled with simple visual scenes. Utterances encode some variable aspects of the scenes (e.g. the type and manner of motion of insects depicted in the scene), while many other aspects of each scene vary randomly. To investigate the effects of hearing only partial utterances on language learning, participants in one condition heard a complete set of 72 three-word utterances, while those in the other condition heard 24 single-word utterances, then 24 two-word utterances, and ultimately 24 three-word utterances. All words were marked with a consistent morphological marker, corresponding with the sentence position (e.g. all sentence-final words, which describe the manner of motion, terminate in the particle *–tig*).

After a training period of 72 utterances and scenes, a battery of two alternative forced-choice tests was used to assess the degree to which participants had learned the meanings of words and utterances. In the 12 *isolated* test trials, each participant chose between two scenes which was the better example of the single-word utterance test item. In 12 *embedded* test trials, each participant chose between two scenes which was the better example of a three-word utterance.

This study is an appealing task to model within our new framework for two reasons. First, it involves learning correspondences between words and many possible candidate features in each scene. For example, participants must infer that the background of a scene is *not* encoded by any words in the lexicon. Second, Kersten and Earles assumed an intersective semantics for their artificial language, making their experiment straightforward to model.

Memory Noise We use a noise model to simulate a learner's imperfect memory or limited attention in observing which words were said. Each word in the set of observed utterances $\mathbb{U}$ is switched with an alternative word that appears in the same sentence position at rate η, between 0 and 1. Edits can be attributed to any mixture of attentional deficit (the learner did not attend to a feature, resulting in an edit) or noisy

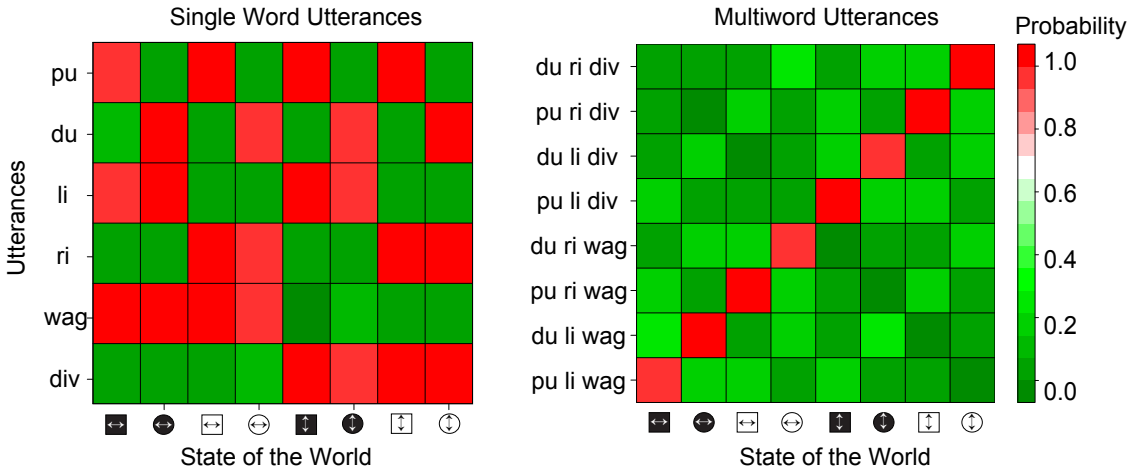

Figure 2: Probability that each single word utterance (left) or multiword utterance (right) can refer to each of eight world-states given the lexicon learned by the model in Simulation 2. Colors represent the probability of generalization, or the probability that a given world-state can referred to by an utterance.

memory. Inference then proceeds over the set of utterances with noise imposed, $\mathbb{U}'$.

Inference To provide for maximum generality in possible word meanings, the framework specifies that any word in the model can refer to the power set of world-states. The hypothesis space for the lexicon is thus a very large discrete space even for the small language presented in Kersten and Earles (2001): there are $2^{6\times 8192}$ possible lexicons (a binary can refer/can't refer indicator for 8,192 possible world-states, for each of 6 words.) We use a hybrid approximation strategy to approximate the posterior in this large space by both sampling from a subset of "structured" hypotheses using Gibbs sampling as well as taking likelihood-weighted samples from the full space.

Structured hypotheses are those that consistently refer to a feature that is shared across states of the world (e.g. all insects that have square bodies). Unstructured hypotheses additionally include heterogenous combinations of world states as potential word meanings, including complex meanings like "bugs traveling upwards so long as the legs move back and forth, and also bugs with oval bodies." Treating meaning as denotation—a mapping from a word to a set of states of the world—permits the representation of both kinds of hypotheses within the same formalism.

Even the structured set alone contains $2^{6\times 26}$ hypotheses. Consequently we use Gibbs sampling (Gelman et al., 2013) to approximate the posterior on the structured set by sampling from the full conditional distribution according to a Markov chain on hypotheses. We use a burn-in period of 2500 samples, then collect 5000 samples and thin to every fifth sample. Convergence was assessed by assessing the log likelihood on repeated simulations. The posterior over the full hypothesis space was estimated using likelihood-weighted samples from the prior. Likelihood weighting is a special case of importance sampling in which the importance distribution is the prior. To compensate for sampling from the prior distribution rather than the posterior, probabilities are adjusted by weighting by the likelihood and normalizing.

The two sampling techniques outlined above have complementary weaknesses: Markov chain Monte Carlo over the structured hypotheses omits the unstructured hypotheses, while the likelihood weighting—in that is sampled from the prior—finds relatively few high-value hypotheses. We mix samples from the two distribution with weights $1-\alpha$ and α. Including the inference procedure, the model for Simulation 2 thus has three free parameters: the per-word error rate η for the stored utterances, noise in the likelihood function ε, and mixing weight α. For the simulations reported here, we take 1000 samples from the prior and set α to .1 and under the assumption that these hypotheses constitute a relatively small proportion of the overall mass, ε to .05, and test a range of η values between 0 and 1 and intervals of .01. We randomly generate 50 experimental setups of the sort described in Kersten and Earles (e.g. different training data and test data in each case) for each level of noise, collect 2 sets of samples using MCMC and likelihood weighting for each setup, and assess each set of samples against 10 instances of the testing battery.

Table 1: Example utterances and scene descriptions from the artificial language learning paradigm in Kersten and Earles (2001), modeled in Simulation 2. Scene vary randomly along five additional dimensions.

Utterance	Body and Legs ("Object")	Path of Movement	Manner of Movement
"geseju elnugop doochatig"	light oval	towards stationary character	legs angled forward and back
"mogaju ontigop neematig"	dark rectangle	away from stationary character	side-to-side movement
"geseju elnugop neematig"	light oval	towards stationary character	side-to-side movement

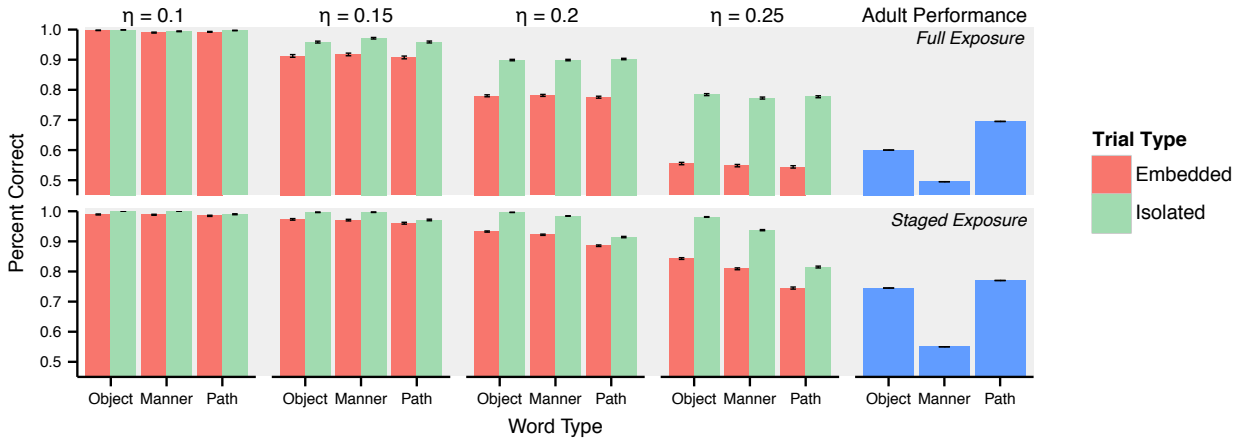

Figure 3: Model performance at four levels of noise compared with adult performance found in Experiment 1 of Kersten and Earles (2001). Participants/models choose between two scenes for a given word (isolated test trials) or a given utterance (embedded test trials). Error bars indicate standard error of the mean.

Results Test scores from Simulation 2 (Figure 3) indicate that the Bayesian word learning model presented here, like human participants, is fully capable of learning correspondences between words and world-states from multiword utterances. The model performs at or near ceiling at low levels of memory noise ($\eta = 0$ to $\eta = .1$), while it demonstrates levels of performance in the range achieved by human participants at moderate levels ($\eta = .1$ to $\eta = .3$). Memory noise levels beyond $\eta = .3$ result in performance near chance.

To further explore the effects of staged vs. full exposure, memory noise, word type, and test trial type we constructed a logistic regression model to predict the outcome of individual forced-choice trials (predicting correct vs. incorrect choices). Manner words, embedded trials, and complete exposure are treated as the reference levels for the categorical predictors. The model scores consistently *higher* on the testing battery when trained on the partial training utterances ($\beta = .118$; $z < 0.001$; intercept = 7.04). Furthermore there is an interaction with memory noise such that the model's performance given partial exposure is higher at higher levels of noise ($\beta = 5.584$; $z < 0.001$). Like participants in Experiment 1 in Kersten and Earles (2001), the Bayesian word learning model presented here performs better when trained on staged exposure. This result, like Kersten and Earles's observation of the empirical phenomenon, is intriguing in that a participant/model in the full exposure condition should be able to achieve the same results as one in the partial exposure condition by selectively attending to just a subset of the data. It appears that the model entertains more inclusive hypotheses for individual words and two-word phrases than three-word phrases, which consequently helps the model to avoid overfitting the lexical hypotheses. In effect, memory noise leads the model to prefer lower-complexity lexicons, which then generalize better upon exposure to novel test data. This conclusion leads to the empirically-testable prediction that the same higher performance for partial exposure would be observed among human participants if the order of staged presentation were reversed—starting with three-word utterances and ending with single word utterances.

The model performance diverges from human behavior in two notable ways. The model performs substantially better on isolated test trials—in which utterances consist of a single word—than on embedded ones ($\beta = .774$, SE = .032, $z < .001$). At higher levels of noise, isolated test trials exhibit higher levels of performance than embedded test trials, as evinced by the trial type $\times$ memory noise interaction term in the model ($\beta = 0.918$, $z < .001$). In contrast, Kersten and Earles found no significant difference in people's performance on the two test trial types ($p > .1$). The model also predicts only minor differences in performance across word types (object, manner, and path), whereas Kersten and Earles found that participants who saw staged input learned object and path words significantly better (74.5% and 77% for those who saw partial input, and 60% and 69.5% for those who saw complete) than manner words (55% for partial and 49.5% for complete exposure). The explanation for this dissociation is straightforward: the model presented here has no information that would substantively distinguish word types from one another. Performance for manner and path are lower under staged exposure because the model observes a manner word in 2/3 of cases and a path word in just 1/3 of cases.

Simulation 3: Multiple Objects Per Scene

A simple Bayesian word learning model that learns the meaning of words independently performs equally well on the first two simulations. In the final simulation, we demonstrate a case in which a model using intersective semantics within the new framework significantly outperforms the simple Bayesian model. Inference, testing procedure, and set of observed utterances are the same as Simulation 2, though we set $\eta = 0$ for simplicity. The critical change is that rather than a single world-state, the model observes *four* different world-states along with each utterance. The likelihood functions in Equations 2 and 7 are altered such that they assess whether *any* of the observed world-states is in the set picked out by the utterance, following from the possibility the utterance could refer to any of the world-states depicted. Additionally, the

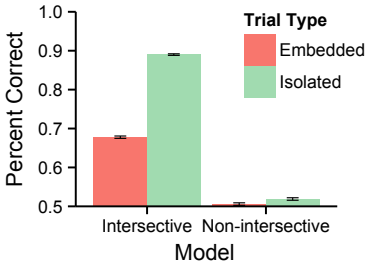

Figure 4: Model performance for the language learning experiment presented in Simulation 3.

four world-states presented in each learning trial are chosen to be very similar to one another: instead of being drawn at random from the entire set of world-states, an observation consists of a world-state and a corresponding veridical utterance in the language, as well as three world-states that are consistent with utterances that differ by only one word from the veridical utterance.

Results In this case, the intersective model significantly outperforms the base model. In the standard Bayesian model, the set of world-states consistent with a given utterance are those that are identified by each independent word-level hypothesis. The intersective learner is more choosy: it only considers world-states that are picked out as the intersection of all word-level hypotheses. In this way, the intersective learner leverages information about other words in the lexicon to identify a single world-state consistent with the entire utterance. Both models perform well if the objects in a scene are highly dissimilar because alternative word-level hypotheses receive little support from the data. However, if the set of observed world-states in each scene are all very similar to one another, the base model entertains many hypotheses as consistent with the data that the intersective model avoids because they do not describe any one world-state in the scene. Performance drops to near chance on the test set for the simple model, while the intersective word learner is still able to infer much of the lexicon (Figure 4).

Discussion

We demonstrate a powerful, extensible, and versatile Bayesian framework for learning word-to-referent mappings from multiword utterances. By assuming an underlying simple compositional semantics, an utterance can be treated as more than a collection of words with independent denotations. Instead, as we demonstrate in Simulation 3, the rich information contained in multiword utterances can be leveraged to guide the word learning process.

The model presented here makes use of strong simplifying assumptions regarding the nature of word meanings and the formalism underlying semantic composition. Word meaning is treated here as denotation, or the selection of world-states, and leaves the matter of connotation unaddressed. For English, this is analogous to saying that word "cat" means the set of things in the world that are cats, whereas criteria like "four-legged," "predatory," and "mammal" are taken as implicitly defining this set. We make the additional simplifying assumption of intersective semantics: "black cats" would refer to the set of things in the intersection of things that are cats and things that are black. Rich compositional semantics, rather than intersective semantics as demonstrated here, will better approximate real-world word meanings.

Despite the shortcomings, we believe that this work is an essential first step in understanding how learners flexibly use information form the entire lexicon in the process of word learning. Future work will require 1) a more elaborate model of semantics 2) the formulation of priors that constrain the space of preferred lexicons 3) the development of inference methods that operate over this large hypothesis space. Fully integrating lexical distributional information will require non-trivial formal machinery for identifying structural categories of words and relating them to dimensions of similarity among world-states. However, by recasting the problem of word learning as one of lexicon inference–and one in which the whole utterance can be used—we take the necessary first steps in bridging the gap between referential and distributional models of word learning.

Acknowledgments. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under grant number DGE-1106400 and by grant number FA9550-13-1-0170 from the Air Force Office of Scientific Research.

References

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*, 3253-3258.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, *14*, 179–211.

Frank, M. C., Goodman, N., & Tenenbaum, J. B. (2008). A Bayesian framework for cross-situational word-learning. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems 20* (pp. 457–464). Curran Associates, Inc.

Gagliardi, A., Bennett, E., Lidz, J., & Feldman, N. (2012). Children's inferences in generalizing novel nouns and adjectives. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 354–359).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. London: Chapman and Hall.

Kersten, A. W., & Earles, J. L. (2001). Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language*, *44*, 250–273.

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

Mintz, T. H. (2003, November). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91–117.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition*, *95*, 201–236.

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

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–68.

Xu, F., & Tenenbaum, J. B. (2007, April). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–72.