Learning word meaning by inferring speakers' intentions:

An incremental approach to socially-guided statistical learning

Michael C. Frank, Molly L. Lewis, Mika Braginsky, & Noah D. Goodman

Department of Psychology, Stanford University

Many thanks to ...

Please address correspondence to Michael C. Frank, Department of Psychology, Stanford University, 450 Serra Mall (Jordan Hall), Stanford, CA, 94305, tel: (650) 724-4003, email: mcfrank@stanford.edu.

Abstract

How do children learn the meanings of words? Some accounts suggest that word learning happens in a single moment, while others privilege the gradual accumulation of information across time. Previous modeling work has attempted to unify these viewpoints in a single framework that allows for both in-the-moment interpretation and gradual statistical accumulation, but at the cost of substantial computational complexity. We describe a new, incremental model of this interaction, in which statistical associations are the product of in-the-moment interpretations. This process-level model successfully captures a number of experimental findings and suggests a number of extensions.

Introduction

Word learning is a foundational part of language acquisition. Starting slowly in late infancy and speeding up after their first birthday, children accumulate a vocabulary of words that they can reliably recognize and produce (Bergelson & Swingley, 2012; Bloom, 2002). The exact timing of this process is highly variable across children, but generally by the end of the second year, children can produce several hundred words and are well on their way to combining these to express complex and novel propositions (R. Brown, 1973; Fenson et al., 1994). How are these early words learned?

In this paper, we elaborate an answer to this question that combines children's statistical learning abilities (Saffran, Newport, & Aslin, 1996; Aslin & Newport, 2012) with their emerging competence in social interaction (Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998). Our proposal is that vocabulary is accumulated via a process in which children attempt to interpret the language they hear. In this proces of interpretation, children make guesses—some more secure than others—about the meanings of words, which they then retain to guide future interpretation. Building on a previous model of joint interpretation and language learning (Frank, Goodman, & Tenenbaum, 2009), we implement this proposal in the language of probabilistic modeling. Our model instantiates the idea that there are two timescales involved in word learning (cf. McMurray, Horst, & Samuelson, 2012): in-the-moment interpretation and cross-situational mapping.

Our model constitutes an incremental approach to integrating across timescales: For each utterance, it proposes a possible referential interpretation, then updates a graded lexicon accordingly. Although the model is incremental, it nevertheless instantiates normative Bayesian inferences; in this respect it is a "rational process model" that makes normative probabilistic inferences under cognitive resource constraints (Griffiths, Vul, & Sanborn, 2012; Sanborn, Griffiths, & Navarro, 2010). We compare this incremental, process-level inference algorithm with

a batch inference algorithm that considers all the available data and find that they are surprisingly similar to one another. One somewhat radical consequence of this approach is that there is no separate process of "word learning"; instead word learning is what falls out of learners remembering their best guesses about what a particular piece of language meant in the contexts of its use.

In what follows, we motivate our proposal with respect to the previous literature on children's word learning. We then review prior computational models of word learning with the goal of grounding and highlighting our conceptual contributions. We next present our model and show that it can be both applied to annotated corpora of child language and can be used to simulate individual experiments. We end by considering open questions for the field of early word learning.

Mechanisms of early word learning

One clue to the mechanisms of early vocabulary learning comes from the makeup of children's early vocabulary. Although children produce words of all different types, certain kinds of words are nevertheless overrepresented in early vocabulary. Names for things make up a substantially larger proportion of children's early vocabulary than they do later (Tardif et al., 2008; Caselli et al., 1995), alongside names for people and and words used in simple social routines (e.g., "hi" and "bye"). Many observers have suggested that these kinds of words—especially basic-level nouns—are learned early because of the ease with which the linguistic form can be mapped to a particular observable referent (Locke, 1700; Bloom, 2002; Clark, 2003). From there, generalizing particular referents to a broader class of referents is relatively trivial, given that the relevant concepts are typically at the basic-level and typically represent whole objects rather than parts (Markman, 1991).

A single learning instance can in principle provide strong evidence for a particular word

meaning. For example, imagine that a parent points to a ball and says "ball" An observer with an understanding of pointing can infer that the speaker is referring to the ball, infer that the concurrently uttered word refers to it, and generalize to the broader concept of balls in general. In practice, however, very young children typically require at least a handful (and often far more) exposures to a word before they recognize it and retain it for future use, even when the context of naming is unambiguous (Woodward, Markman, & Fitzsimmons, 1994).

In addition, many learning situations range from slightly to substantially more ambiguous. Imagine that the parent had uttered the phrase "look at that nice, round ball!"—this phrase has a number of competing words that might be candidates for the object name (though perhaps worse candidates by virtue of sentential position, stress, or phonological structure). Or the parent might not have pointed but instead relied on the fact that the child was playing with the ball. Perhaps there might have been other toys present competing for the child's attention. In each of these cases, the child might be able to guess that the word "ball" referred to the ball, but with less certainty than in the simplest case (Gillette, Gleitman, Gleitman, & Lederer, 1999 and Yurovsky, Smith, & Yu, 2013 give concrete instantiations of such differences in certainty in naturalistic play session videos).

Many theorists have noted that learners in principle could accumulate information across many ambiguous naming instances (Pinker, 1984; Gleitman, 1990; Siskind, 1996; Yu & Ballard, 2007). This strategy, dubbed *cross-situational learning*, has now been demonstrated in a number of small-scale laboratory experiments with both adults and children (Yu & Ballard, 2007; L. Smith & Yu, 2008). In addition, a number of computational demonstrations suggest that this kind of strategy could be effective for learning in social situations (Yu & Smith, 2007; Frank et al., 2009; Johnson, Demuth, & Frank, 2012), with realistic vocabulary sizes and levels of ambiguity (Blythe, Smith, & Smith, 2010), and even with more complex propositional meanings

(Siskind, 1996). In fact, a growing literature in natural language processing implements precisely this strategy for a variety of what are known as "grounded language learning" tasks (e.g. Zettlemoyer & Collins, 2005; Wong & Mooney, 2007; Artzi & Zettlemoyer, 2013; Liang, Jordan, & Klein, 2011; Kim & Mooney, 2013).

Nevertheless, the existence of such an uncertainty-reduction strategy does not necessarily imply that learners make use of it. Indeed, there has been substantial debate about the degree to which learners represent cross-situational statistics (Medina, Snedeker, Trueswell, & Gleitman, 2011; Yu & Smith, 2012; Yurovsky et al., 2013; Trueswell, Medina, Hafri, & Gleitman, 2013; K. Smith, Smith, & Blythe, 2011; Yurovsky & Frank, under review). On some accounts, learners encode only a single hypothesis at a time; on others, learners maintain some representation of all of the data that they have access to. But regardless of the specific representation and algorithm that learners use for this process, all accounts of cross-situational learning posit *consistency* across situations, which even single-hypothesis learners exploit by checking their hypotheses across situations (Trueswell et al., 2013; Yu & Smith, 2012). To understand the range of possible theories of cross-situational learning, we next turn to the modeling literature.

Prior modeling work

Although a number of theorists had discussed cross-situational strategies for learning word meanings (Pinker, 1984; Gleitman, 1990), an important early instantiation of this idea came from a model by Siskind (1996). This model used a propositional representation of meaning in combination with a set of deductive rules to infer word-meaning mappings from artificial data. This ambitious model provided a powerful proof-of-concept, but was limited by the assumption that the propositional structure of meanings was observed by the learner.

An alternate, more graded, view of word learning came from the connectionist tradition. Plunkett, Sinha, Møller, and Strandsby (1992) described a graded word-image mapping model

that reproduced a number of category generalization and vocabulary growth phenomena. This basic model type has been followed by a number of related models that capture phenomena like mutual exclusivity (Regier, 2005), the phonological dynamics of the mono- and bilingual lexicon (Li & Farkas, 2002; Li, Farkas, & MacWhinney, 2004; Li, Zhao, & Whinney, 2007), and the emergence of categorization principles (Mayor & Plunkett, 2010). Though these models provide a powerful example of the emergence of a range of phenomena from a simple architecture, none focused specifically on the challenge of disambiguating reference in ambiguous contexts.

The problem of referential uncertainty played a more central role in a number of other models that emerged from the machine learning tradition. In a series of investigations, Yu and colleagues developed a model stemming from classic machine translation systems (P. Brown, Pietra, Pietra, & Mercer, 1993). This model looked for correspondences between words and their referents that was consistent across situations; these mappings could be biased by a number of perceptual and social aspects of the learning scenario (Yu, Ballard, & Aslin, 2005; Yu & Ballard, 2007). Critically, this model could be applied to annotated corpora of child-directed speech, a major advance over previous work that had only been used to learn from artificial corpora.

A complementary model used the Bayesian framework considered how generalization biases could emerge from statistical inferences under ambiguity across situations (Xu & Tenenbaum, 2007). This model was able to predict adults' and children's patterns of taxonomic generalization by reference to general principles of probabilistic inference. Although it did not specifically apply to the problem of referential uncertainty across situations, it nevertheless aggregated information across exposures to make graded inferences about word meaning.

Building on these two lines of work, we proposed a probabilistic treatment of referential uncertainty in Frank et al. (2009). This model differed from earlier approaches in that it explicitly assumed that in any individual situation, speakers have an intention to refer to a particular object

or objects and that the labels they utter correspond to these words. In contrast, previous models had computed associations directly between words and referents, without considering that some of these associations were not relevant because the speaker might not be trying to refer to some objects. This "intentional" assumption led to an increase in accuracy in learning from corpus data relative to previous work, and also allowed the model to capture a number of experimental findings. A number of related pieces of work have extended this model to incorporate word segmentation (Johnson, Demuth, Frank, & Jones, 2010), social cues (Frank, Ichinco, & Tenenbaum, 2008; Johnson et al., 2012), conceptual generalization (Lewis & Frank, 2013a), lexical constraints (Lewis & Frank, 2013b), bilingual lexica (Zinser, Rolotti, & Li, under review), some basic aspects of grammatical structure (Johnson et al., 2010), and even pragmatic inference (N. J. Smith, Goodman, & Frank, 2013).

The basic framework described in Frank et al. (2009) nevertheless suffered from a number of weaknesses. First, it posited a discrete lexicon represented by a bipartite graph linking words and object concepts. While this representation was technically convenient, it seemed at odds with the conception of graded, uncertain mappings implied by the experimental literature.[1] In addition, the model was posed at Marr's (1982) "computational theory" level; as such, it considered all the available data in its computation. This "batch" inference was of course unrealistic in terms of the memory constraints on leaners; in addition, practically speaking it kept the model from considering phenomena that involved the gradual accumulation of data about a particular inference. Finally, because of technical limitations, the inference scheme in this model was not

[1] Although there is debate about the uncertainty implied by adult cross-situational learning experiments (e.g. K. Smith et al., 2011; Trueswell et al., 2013), many experiments with children imply increases in word recognition accuracy with greater experience (e.g. Fernald, Pinto, Swingley, Weinbergy, & McRoberts, 1998; Bergelson & Swingley, 2012). Some sort of graded conception of the lexicon appears to be an important aspect of models that hope to capture such findings.

fully Bayesian: It did not recover the posterior probability distribution on lexicons given a particular exposure. Instead, it only searched for the single best lexicon.

Two other recent models have made progress in using incremental inference to describe word learning. First, Fazly, Alishahi, and Stevenson (2010) proposed an incremental version of the Yu and Ballard (2007) translation model, which they applied to meaning representations derived from speech in CHILDES. This model reproduced a number of empirical findings in the incremental context. In addition, the meaning representations were graded distributions over the full inventory of lexical elements. Both of these desirable features pointed the way towards further applications, and this basic model has been extended to simulate cognitive constraints (Nematzadeh, Fazly, & Stevenson, 2012), individual differences in acquisition (Nematzadeh, Fazly, & Stevenson, 2011), and the effects of syntactic structure (Alishahi & Fazly, 2010). One weakness of this work, however, is that this model has no obvious route for the accommodation of social information.

Second, McMurray et al. (2012) introduced a connectionist model that—like the Frank et al. (2009) model—explicitly operates at two timescales: the timescale of language interpretation in the moment and the timescale of cross-situational mapping. Because of the incremental nature of this model, however, it was able to simulate the dynamics of reference resolution in the moment. In addition, the exhaustive set of simulations using this model suggests that the combination of graded representation, incremental processing, and multi-timescale inference allows this model to captures most of the relevant data points. Similar to the Fazly et al. (2010) model, however, the one major lacuna in this set of simulations is social learning phenomena.

The current model

Taken together, the models presented by Frank et al. (2009), Fazly et al. (2010), and McMurray et al. (2012) synthesize and unify a striking amount of work on the mechanisms and

dynamics of early word learning. Each of these has some features that could be improved upon, however. In particular, the Frank et al. (2009) model is non-incremental, while the Fazly et al. (2010), and McMurray et al. (2012) models fail to consider social information. Our goal in the current work is thus to provide an incremental interpretation of the Frank et al. (2009) model, which would serve to unify the social framework in that work with the findings of these other incremental models.

Model

In the sections that follow, we describe the formal specification of our model as well as two inference algorithms—a batch Gibbs sampler and an incremental particle filter—that are focused around referential interpretation. As will become clear, these two inference algorithms are deeply related to one another. All code and data for the model and simulations reported below are available at `http://github.com/mcfrank/dmww`.

The schematic word learning situation is shown in Figure 1. The learner is hypothesized to jointly infer the speaker's referential intention, I, and the lexicon of their language, L. These inferences are informed by the elements of the context, C: the words, W, that the speaker utters, the relevant referents (objects O) that are present in the situation, and potentially auxiliary social cues.

The generative model

A sketch of the generative model is shown in Figure 2. We will write $\vec{W}$ for all utterances across contexts, W^i for the utterance in the ith context, and W^i_j for its jth word; similar for other variables.

Our primary departure from Frank et al. (2009) is to assume that the lexicon places a distribution over all words for each object, rather than being a discrete association from objects to

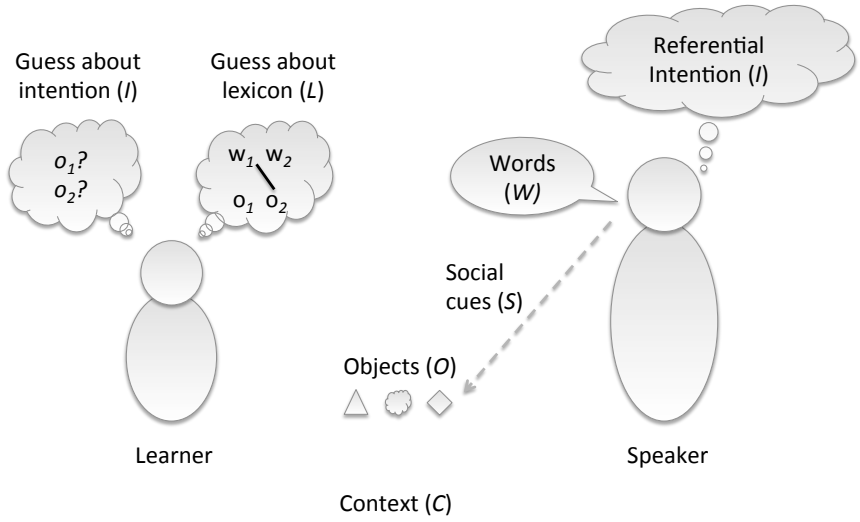

Figure 1. : A schematic depiction of the learning situation, with the relevant variables marked for ease of interpretation. The learner faces a joint inference problem: inferring what the speaker is talking about and learning the consistent meanings of words. Reprinted from ? (?).

words. That is, we smooth the previous model by allowing some probability that an unusual word is used to refer to an object. Formally we assume each the lexical entry for each object (type) is a distribution over words drawn from a symmetric Dirichlet distribution:

$$L_{ob} \sim \text{Dirichlet}(\alpha) \tag{1}$$

Where *ob* can be any object in the world and can also be the special "object" *NR*, which accounts for all non-referential words.[2] The words themselves will be drawn below from multinomial distributions with parameters determined by L_{ob}.

Following Frank et al. (2009) we assume that speakers have simple intentions to refer to

[2]We continue to discuss the model as linking words with *objects*, rather than with *concepts*. Lewis and Frank (2013a) describes the generalization of a related model to learning category generalizations. Since the data we consider here are annotated for objects rather than concepts, the difference is purely rhetorical; but concepts are clearly the right units of generalization for future work.

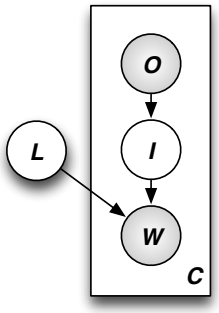

Figure 2. : The generative model assumed in our current work, similar to Frank et al. (2009). The plate over contexts indicates that objects, words, and referential intentions are present in each context, while the lexicon stays constant across contexts. [should update this fig with reference variables... also not currently mentioned in main text.]

objects in the context. For current purposes we simplify by assuming speakers intend to refer to only one object, and that they will refer to each possible object with equal probability. Thus $I^i \in O^i$ and $P(I^i|O^i) = \frac{1}{|O^i|}$. [This is where we put social cues.. either intention directly depends on social cues, $P(I^i|O^i,S^i)$, or social cues depend on intention, which still leads to a distribution on intentions that depends on social cues.]

We next assume that speakers realize their intention by deciding which word in the utterance should be the referring word, generating this word appropriately using the lexicon, and then generating "non-referring" words for the remainder of the utterance. [3] That is, for an utterance of N_i words there is a variable $r^i \in \{1,\ldots,N_i,\emptyset\}$ that determines which word realizes the referential intention. The probability that $r^i = \emptyset$ (i.e. no word realizes the referential intention) is a fixed constant, and r^i has equal probability for the other words. Words are then drawn from either

[3]This is a departure from Frank et al. (2009) where it was assumed that each word might refer to an object independent of whether anther word also referred to that object. That is, here we assume that the pressure to refer can be "satisfied" by one word, leaving the other words to play other (non-referential) roles.

the distribution specified by the lexicon for the intention object or the *NR* non-referential item:

$$W_j^i \sim \begin{cases} \text{Multinomial}(L_{I^i}) & \text{if } r^i = j \\ \text{Multinomial}(L_{NR}) & \text{otherwise} \end{cases} \tag{2}$$

Note that for any j, $\emptyset \neq j$ and as a result *no* word in the utterance is referring when $r^i = \emptyset$.

The words in the utterance are independent given r^i, I^i, and L, hence:

$$P(W^i|r^i,I^i,L) = \prod_{j=1..N} P(W_j^i|r^i,I^i,L) \tag{3}$$

Learning at the computational level

We assume that the words, $\vec{W}$, and objects, $\vec{O}$, are fully observed in the learning contexts, and hence the problem for the learner is to infer the lexicon, L, and the intentions and referential words, $\vec{I}$ and $\vec{r}$. By Bayes' rule:

$$P(\vec{I},\vec{r},L|\vec{W},\vec{O}) \propto P(\vec{W}|\vec{I},\vec{r},L,\vec{O})P(\vec{I},\vec{r},L|\vec{O}).$$

By assumption $\vec{W}$ depends on the context only through the intention, $\vec{I}$, while $\vec{R}$ and L are *a priori* independent of the other variables. In addition, inferences in different contexts are independent, given L, hence:

$$P(\vec{I},\vec{r},L|\vec{W},\vec{O}) \propto P(\vec{W}|\vec{I},\vec{r},L)P(\vec{I}|\vec{O})P(\vec{r})P(L) \tag{4}$$

$$\propto P(L)\prod_i P(W^i|I^i,r^i,L)P(I^i|O^i)P(r^i) \tag{5}$$

While the learner must jointly consider intentions and lexical meanings, it is only the posterior on lexica, $P(L|\vec{W},\vec{O})$, that impacts future language use or learning (that is, L screens offprevious contextual variables from future instances of those variables). Thus, the task of the learner in the moment can be seen as integrating out the contextual variables (and then updating lexicon beliefs): $P(L|\vec{W},\vec{O}) = \sum_{\vec{I},\vec{r}} P(\vec{I},\vec{r},L|\vec{W},\vec{O})$. We next describe two algorithmic, sampling-based strategies for the context variables.

Sequential learning

Imagine that we know the posterior on lexica for some amount of learning data, $P(L|\vec{W},\vec{O})$, and then observe an additional context, w and o. The full posterior can be written as:

$$P(L|\vec{W},\vec{O},w,o) \propto P(w|o,L)P(L|\vec{W},\vec{O}) \quad (6)$$

$$\propto \sum_{I,r} P(w,I,r|o,L)P(L|\vec{W},\vec{O}) \quad (7)$$

We take the policy of approximating the summand of Equation 7 with a single sample from the distribution on I,r.[4] This simplifies the calculation required. The initial prior, Equation 1, on the lexicon entries L_{ob} is Dirichlet; because Dirichlet is the conjugate prior to the multinomial distribution on words, the posterior, given an I and r, will also be Dirichlet (?, ?). If we approximate the posterior by a single sample of I and r in each context then the distribution on lexica remains Dirichet throughout.

That is, if we assume $P(L_{ob}|\vec{W},\vec{O})$ is distributed as Dirichlet with pseudo-count of word k from object ob being c_k^{ob}, then the posterior will be Dirichlet with counts updated appropriately: for true objects

$$c_k'^{ob} = c_k^{ob} + \delta_{w_r=k},$$

and for the non-reference "object"

$$c_k'^{NR} = c_k^{NR} + \sum_{j\in\{1,\ldots,|w|\}\setminus r} \delta_{w_j=k}.$$

This closed form update means that we can easily sample I and r in the next context (by doing this update for each possible I,r and normalizing) and then update the posterior on lexica; iterating through all learning contexts in this way yields a sample from the distribution on lexica

[4] Iteratively sampling the context variables also yields a sample from the distribution on these variables across contexts. Since this is a form of sequential importance sampling the proof is standard. [I kind of punted here... do we need to show that this is a sample from the posterior explicitly?]

conditioned on all the learning evidence. Thus, we construct learning algorithm that focusses on interpreting the intention, I, and it's realization, r, in the moment, but uses these inferences to update beliefs about the lexicon, L, and it is these updated lexical beliefs that get carried forward into the future. [should walk through an example ala figure 3.]

Batch learning

In the full generative model the contexts are exchangeable: the order of contexts doesn't impact the overall probability. This means that we could always pretend a particular context was the last one. This idea gives rise to the *Gibbs sampling* strategy: rather than building a sample incrementally as above, start with any assignment of $\vec{I}$ and $\vec{r}$ and then update by selecting a context, pretending it is the last, removing its counts and then re-sampling using the update described above. Sweeping through contexts updating in this way yields a Markov chain Monte Carlo sampling algorithm—independent of were we start the context variables, they will converge to a sample from the correct posterior.

[if we want to make the point that batch and incremental aren't so different, we should point out that the key property that let us derive both is that contexts (i.e. I and r) are independent given lexica, L. the conjugacy business just gives us a clean closed form for the, evolving, posterior on L. where should this discussion go?]

Simulations

Corpus simulations

We provide proof of concept that. We make use of the Rollins Corpus from CHILDES (MacWhinney, 2000),

Cross-situational word learning with adults

A key test of the model is its ability to learn word-object mappings across many individually ambiguous contexts. To evaluate the model's performance, we tested the model in a design identical to Yu and Smith (2007). We presented the model with 18 novel words across many trials that were individually referentially ambiguous. A trial consisted of n words presented with n objects. As in Yu and Smith (2007), we tested three values for n: 2, 3, 4. For example, in an $n = 2$ condition, the model was presented with two words and two possible referents. Across conditions, the number of co-occurences of each word-object pairing was held constant. We evaluated the posterior lexicon using a luce choice rule for each word with four referential alternatives (one correct and three foils).

Our model was able to successfully recover the correct mappings for most words. Because the lexicon is a continuous representation unlike the original model (Frank et al., 2009), the model produced a graded pattern in performance in the test trials. As in the behavioral data (Yu & Smith, 2007), accuracy declined as the number of referential alternatives in the training trials increased: the model was most accurate in the $n = 2$ condition, and least accurate in the $n = 4$ condition (Fig. 4). The particle filter inference algorithm performed overall better than the Gibbs sampler.

Experiments with children

Disambiguation. Disambiguation is the behavioral bias for children to map novel words on In the classic demonstrations of this phenomenon, children are presented with two objects, one familiar and one novel, and asked to This phenomenona is a particular

To simulate this result, we asked our model to determine the meaning of a

Dewar & Xu (2007).

Discussion

Open questions for the cross-situational, social viewpoint

Synergies with other problems.

Extensions to other parts of the vocabulary.

Representations supporting cross-situational word learning. (Yurovsky & Frank, under review)

Conclusions

References

Alishahi, A., & Fazly, A. (2010). Integrating syntactic knowledge into a model of cross-situational word learning. In *Proc. of cogsci* (Vol. 10).

Artzi, Y., & Zettlemoyer, L. S. (2013). Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, *1*, 49–62.

Aslin, R. N., & Newport, E. L. (2012). Statistical learning from acquiring specific items to forming general rules. *Current directions in psychological science*, *21*, 170–176.

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.

Bloom, P. (2002). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Blythe, R. A., Smith, K., & Smith, A. D. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*, 620–642.

Brown, P., Pietra, V., Pietra, S., & Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, *19*, 263–311.

Brown, R. (1973). *A first language: The early stages.* Harvard U. Press.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, *63*.

Caselli, M., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., et al. (1995). A cross-linguistic study of early lexical development. *Cognitive Development*, *10*, 159–199.

Clark, E. (2003). *First language acquisition*. Cambridge, UK: Cambridge University Press.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*, 1017–1063.

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D., Pethick, S., et al. (1994). Variability in early communicative development. *Monographs of the society for research in child development*, *59*.

Fernald, A., Pinto, J., Swingley, D., Weinbergy, A., & McRoberts, G. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*, 228.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

Frank, M. C., Ichinco, D., & Tenenbaum, J. B. (2008). Principles of generalization for learning sequential structure in language. In *Proceedings of the 30th annual meeting of the cognitive science society.*

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*, 135–176.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 3–55.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*, 263–268.

Johnson, M., Demuth, K., & Frank, M. (2012). Exploiting social information in grounded language learning via grammatical reduction. In *Acl.* Retrieved from `http://www.aclweb.org/anthology/P12-1093`

Johnson, M., Demuth, K., Frank, M., & Jones, B. (2010). Synergies in learning words and their referents. *Advances in Neural Information Processing Systems*.

Kim, J., & Mooney, R. J. (2013). Adapting discriminative reranking to grounded language learning. In *Acl.* Retrieved from `http://www.cs.utexas.edu/users/ml/papers/kim.acl13.pdf`

Lewis, M., & Frank, M. C. (2013a). An integrated model of concept learning and word-concept

mapping. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society.*

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

Li, P., & Farkas, I. (2002). A self-organizing connectionist model of bilingual processing. *Bilingual sentence processing*, 59–85.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, *17*, 1345–1362.

Li, P., Zhao, X., & Whinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science: A Multidisciplinary Journal*, *31*, 581–612.

Liang, P., Jordan, M. I., & Klein, D. (2011). Learning dependency-based compositional semantics. In *Association for computational linguistics (acl).*

Locke, J. (1700). *An essay concerning human understanding*.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Markman, E. M. (1991). *Categorization and naming in children: Problems of induction*. The MIT Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: Henry Holt and Co.

Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological review*, *117*, 1.

McMurray, B., Horst, J., & Samuelson, L. (2012). Using your lexicon at two timescales: Investigating the interplay of word learning and recognition. *Psychological Review*, *119*, 831–877.

Medina, T., Snedeker, J., Trueswell, J., & Gleitman, L. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*, 9014.

Nematzadeh, A., Fazly, A., & Stevenson, S. (2011). A computational study of late talking in word-meaning acquisition. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 705–710).

Nematzadeh, A., Fazly, A., & Stevenson, S. (2012). A computational model of memory, attention, and word learning. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics* (pp. 80–89).

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

Plunkett, K., Sinha, C., Møller, M., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, *4*, 293–312.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science: A Multidisciplinary Journal*, *29*, 819–865.

Saffran, J. R., Newport, E., & Aslin, R. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*, 606–621.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, *117*, 1144.

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

Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*, 480–498.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational

statistics. *Cognition*, *106*, 1558–1568.

Smith, N. J., Goodman, N., & Frank, M. (2013). Learning and using language via recursive pragmatic reasoning about other agents. In *Advances in neural information processing systems* (pp. 3039–3047).

Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N., & Marchman, V. A. (2008). Baby's first 10 words. *Developmental Psychology*, *44*, 929.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, *66*, 126–156.

Wong, Y. W., & Mooney, R. J. (2007). Learning synchronous grammars for semantic parsing with lambda calculus. In *Acl.* Retrieved from `http://www.cs.utexas.edu/users/ai-lab/?wong:acl07`

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

Xu, F., & Tenenbaum, J. (2007). Word Learning as Bayesian Inference. *Psychological Review*, *114*, 245.

Yu, C., & Ballard, D. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*, 2149–2165. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S092523120600508X`

Yu, C., Ballard, D., & Aslin, R. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science: A Multidisciplinary Journal*, *29*, 961–1005.

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

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word–referent learning: Prior questions. *Psychological review*, *119*, 21.

Yurovsky, D., & Frank, M. C. (under review). An integrative account of constraints on cross-situational word learning.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: the baby's view is better. *Developmental science*, *16*, 959–966.

Zettlemoyer, L. S., & Collins, M. (2005). Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars. In *Uai.*

Zinser, B., Rolotti, S., & Li, P. (under review). Bayesian word learning in multiple language environments.

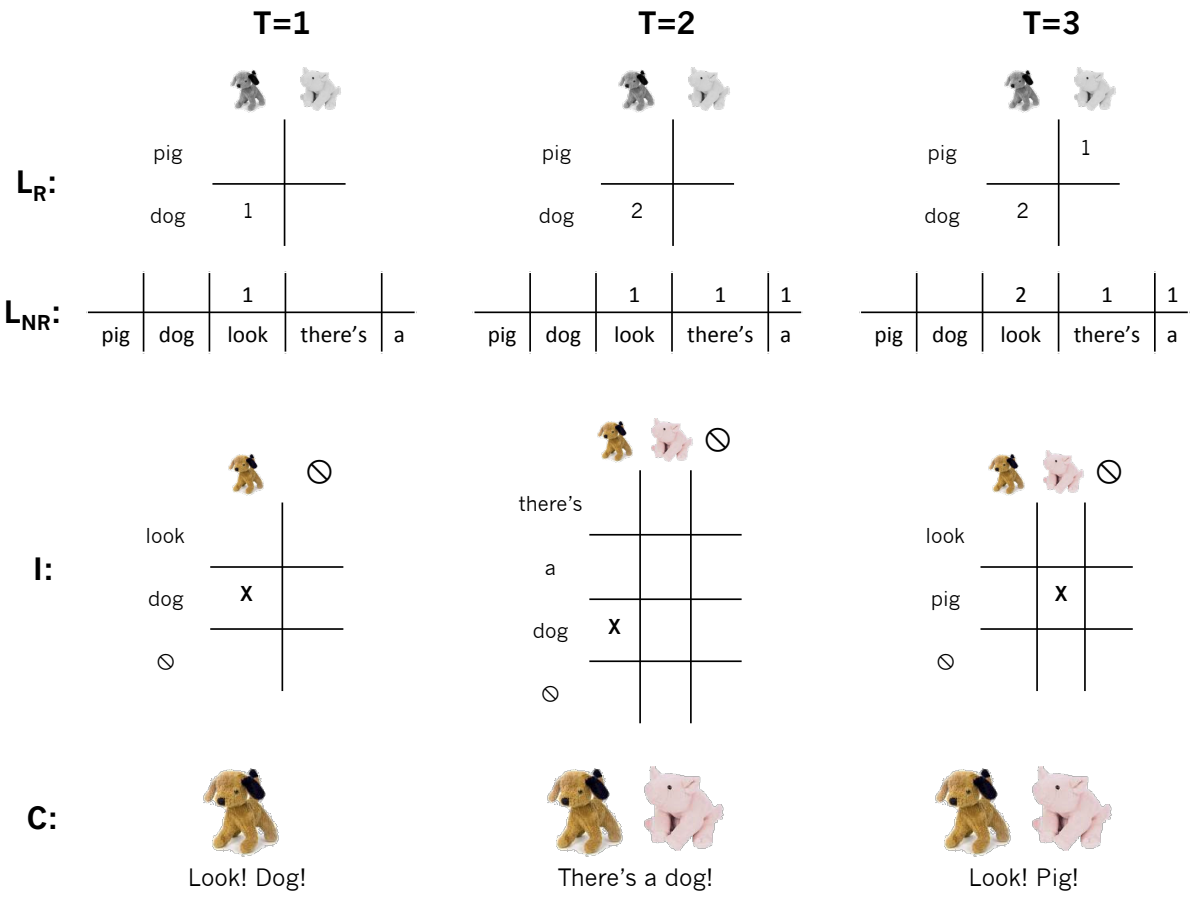

Figure 3. : A schematic of the model's evolving internal states when applied to a simple corpus with three contexts (marked $T = 1, 2, 3$). States are shown for the particle filter inference algorithm with a single particle. Show at the top are the referential and non-referential lexicons (for simplicity, only the words "pig" and "dog" are shown in the referential lexicon). In the middle, the model's guesses about the referential correspondence between words and objects are shown for each situation.

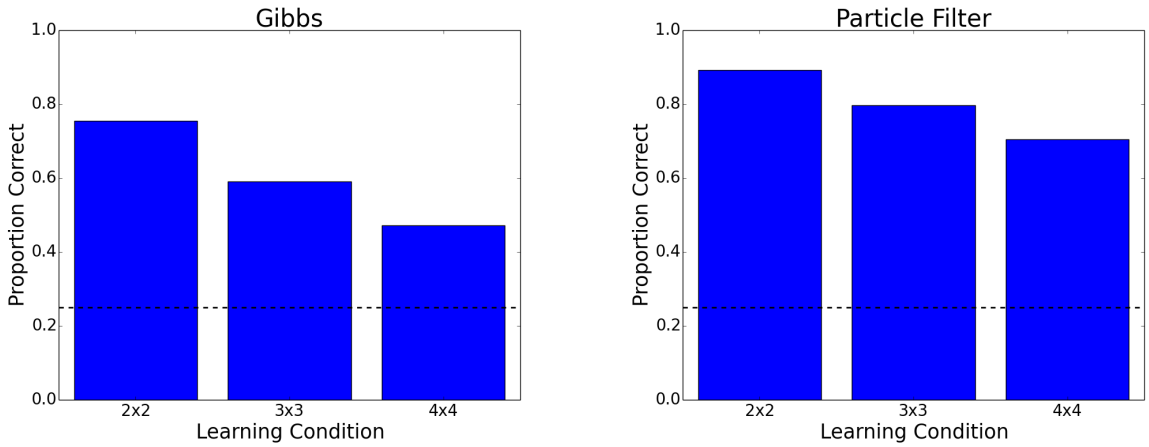

Figure 4. : Model predictions for the Yu and Smith (2007) stimuli. Proportion correct in a four alternative forced choice task is plotted as a function of the referential ambiguity condition. Performance is plotted for the Gibbs sampler (left) and incremental particle filter (right). For both inference algorithms, the model performs above chance on all conditions, and displays a graded decline in performance as the number of referential alternatives increases.