Using speakers' referential intentions to model early cross-situational word learning

Michael C. Frank, Noah D. Goodman, and Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Text: XXX words (notes = 166, acknowledgements = 24)

Address for correspondence:
Michael C. Frank
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
77 Massachusetts Ave.
Room 46-3037D
Cambridge, MA 02139
tel: (617) 452-2474
email: mcfrank@mit.edu

The authors gratefully acknowledge Roberta Golinkoff, Kathy Hirsch-Pasek, Fei Xu, and Chen Yu for many valuable discussions. The first author was supported by a Jacob Javits Graduate Fellowship.

Abstract

Word learning is a chicken-and-egg problem: if a child understands the meaning of an utterance, it is easy to learn the meanings of individual words; if the child knows what some words mean, it is easy to infer the speaker's intended meaning. We present a Bayesian model that solves these two problems in parallel, rather than learning exclusively from the inferred meanings of utterances or relying only on cross-situational association of words and meanings. Our model infers word-object pairings from CHILDES data with high precision. By using probabilistic inference, it predicts experimental results on mutual exclusivity, one-trial learning, and cross-situational learning; because it directly represents speakers' intentions, it predicts results on object individuation and the use of intentions to disambiguate reference. Our results suggest that theories of early word learning that explicitly represent speakers' communicative intentions and use probabilistic inference to reason about them achieve greater coverage of experimental phenomena.

Abstract: 150 words

Using speakers' referential intentions to model early cross-situational word learning

When children learn their first words, they face a complex chicken-and-egg problem: they are both trying to infer what meaning a speaker is attempting to communicate at the moment a sentence is uttered and trying to learn the more stable mappings between words and referents that constitute the lexicon of their language. Given either of these pieces of information, their task becomes considerably easier. Knowing the meanings of some words, a child can often figure out what a speaker is talking about; on the other hand, inferring the meaning of the speaker's utterance allows the child to work backwards and learn basic-level object names with relative ease. However, for a learner without either of these pieces of information, word learning is a hard computational problem. Following Quine's (1960) metaphor, a young word learner is climbing the inside of a chimney, "supporting himself against each side by pressure against the others" (p. 93).

Many accounts of word learning focus primarily on one aspect of this problem. Social theories suggest that learners rely on a rich understanding of the goals and intentions of speakers and assume that—at least in the case of object nouns—once the child understands what is being talked about, the mapping between words and referents is relatively easy to learn (Augustine, 397/1963; Baldwin, 1993; Bloom, 2002; Tomasello, 2003). These theories must assume some mechanism for making mappings, but it is often taken to be deterministic and its details are rarely specified. In contrast, cross-situational accounts of word learning take advantage of the fact that words often refer to the immediate environment of the speaker, allowing learners to build a lexicon based on consistent associations between words and their referents (Locke, 1690/1964; Siskind, 1996; Smith, 2000; Yu & Smith, 2007).

Computational models of word learning have primarily followed the second, cross-

situational strategy. Models using connectionist (Plunkett, Sinha, Møller, & Strandsby, 1992), deductive (Siskind, 1996), competition-based (Regier, 2005), and probabilistic methods (Yu & Ballard, 2007) have had significant successes in accounting for many phenomena in word-learning. However, speakers often talk about objects that are not visible and actions that are not in progress at the moment of speech (Gleitman, 1990), adding noise to the correlations between words and objects. Thus, cross-situational and associative theories often appeal to external social cues like eye-gaze (Smith, 2000; Yu & Ballard, 2007), but they are used as markers of salience (the "warm glow" of attention), rather than as evidence about internal states of the speaker as in social theories.

More generally, cross-situational theories address only one part of the learners' task—they are able to learn words, but they do not use the words that speakers utter to infer the speakers' intended meanings. By focusing only on the long-term mappings between items in the lexicon and referents in the world, purely cross-situational models effectively treat the complex and variable communicative intentions of speakers as noise to be averaged out via repeated observations or minimized via the use of attentional cues, rather than as an important aspect of communication to be used in the learning task.

Here we present a Bayesian model that captures both aspects of the word learning task: it jointly infers what speakers are attempting to communicate and learns a lexicon. We first present the structure of the model and show that it obtains competitive results in learning from mother-child interactions in the CHILDES corpus. We then show how the probabilistic structure of the model allows it to predict experimental results such as mutual exclusivity (Markman & Wachtel, 1988), one-trial word learning (Carey, 1978; Markson & Bloom, 1997), and rapid cross-situational learning (Smith & Yu, in press; Yu & Smith, 2007) while its explicit representation of

intention allows it to predict results on object individuation (Xu, 2002) and word learning using intentional cues (Baldwin, 1993).

Model design

Our goal was to use two observable variables—the words a speaker utters and the context at the time of the utterance (an utterance and its context is collectively referred to as a situation)—to infer two unobservable variables: the lexicon of the speaker's language and the intended meaning the speaker wants to convey with each utterance. To accomplish this task, we constructed a generative model (Figure 1) which formally defines the relationship between each of these variables. We assume that the words uttered in any situation depend on the lexicon; however, the particular words in an utterance also depend on what the speaker intends to say. This intended meaning in turn depends on the observable physical context of the utterance.

While this basic model could be applied to representations of greater complexity, we chose very simple representations of situations, words, and intended meanings. For instance, although the observable context of an utterance could in principle include actions and events, we represent only the mid-size objects that were present at the time of an utterance (in our corpus, toys that mothers and children were playing with). Likewise, the intended meaning of a sentence could be represented by a complex proposition, but here we represent it as only the intended referents of a sentence (a subset of those objects that are present in the context). Finally, rather than representing the full syntactic structure of utterances, our model treats all words independently.

Given these representational simplifications, the graphical model in Figure 1 defines a conditional probability distribution over words W, given the lexicon L and the objects O in the physical context. This distribution can be notated as in equation (1):

$$P(W \mid L,O) = \#_{I} \; P(W \mid I,L)P(I \mid O) \qquad (1)$$

For our purposes, the intention *I* of a sentence is simply a subset of the objects that are present while the sentence is uttered. For instance, if both a ball and a box are present, a speaker could intend to refer to both objects, either one, neither, or something else entirely. We limit the set of intentions the model considers to subsets of the set of objects present in the physical context (though the model has the option of deciding that none of the objects in the context are being talked about) and assume a uniform distribution over these intentions. The model is given no additional information about what a speaker intends to refer to, though in principle we could include other cues to the speaker's intention, such as eye-gaze or pointing. Thus, rather than picking one intention we instead sum over all possibilities, as shown in Equation 1.

We next define the probability of a sentence given a lexicon and an intention by a product over each word in the sentence (because words are assumed to be independent):

$$P(W \mid I,L) = \prod_{w \in W} \left[\gamma P_R(w \mid I,L) + (1-\gamma) P_{NR}(w \mid L) \right] \qquad (2)$$

In this formulation, each word is uttered either because it refers to an object in the intention (with probability) or because it plays some other role in the utterance (with probability 1-). Words in the lexicon pay a penalty ! if they are used non-referentially. This split between referential and non-referential words contrasts with models which rely only on cross-situational association and represent some words (e.g., "the" or "of") only as being less associated with particular objects than the objects' correct labels as opposed to being truly non-referential.

We then define the probability of a particular referential word given some intention, $P_R(W \mid I,L)$. This probability is simply the probability that a particular word would be chosen from the lexicon (represented here as a list of word-object pairings, as in Figure 2) to refer to a

particular object in the intention (if there is no link between the two, the probability is zero; otherwise the probability is proportional to the number of words that correspond to this object). Other words in the sentence are chosen uniformly from the entire vocabulary of the language; but our model assumes that a speaker is *a priori* less likely to use a word for an object non-referentially, so words in the lexicon are slightly less likely to occur when their corresponding object is not present.

Our model defines a generative process over words, objects, intents, and the lexicon. We use Bayes' rule to invert this process and calculate the posterior probability of lexicons given a corpus containing multiple situations:

$$P(L \mid W, O) \propto P(W \mid L, O) \cdot P(L) \qquad (3)$$

This formulation captures the tradeoff between the likelihood of the observed words being generated by a particular lexicon and the prior probability of that lexicon. We chose a simple parsimony prior over lexicons which makes lexicons exponentially less probable as they include more word-object pairings: $P(L) \propto e^{-\alpha|L|}$. In the results below, we employ stochastic search methods using simulated tempering (Marinari & Parisi, 1992) to find the lexicon with the maximum a posteriori probability.

Corpus evaluation

Corpus. We employed two video files of ~10 minutes each from the Rollins section of CHILDES (me06 and di03) in which two preverbal infants and their mothers play with a set of toys. Each line of the transcripts was annotated with a list of all midsize objects judged to be visible to the infant (annotated files available at http://tedlab.mit.edu/~mcfrank/corpus).[1]

Alternate models. For comparison, we implemented several other models of cross-situational word learning using co-occurrence frequency, conditional probability, and point-wise

mutual information. We also implemented IBM Machine Translation Model I (Brown, Pietra, Pietra, & Mercer, 1994), the statistical machine translation model used by Yu and Ballard (2007). We used the translation model to compute association probabilities for both objects given words and words given objects. For each comparison model, we computed the relevant statistic for all possible word-object pairings and then chose a threshold value to maximize the F-score[2] of the resulting lexicon.

Results. We evaluated all models both on the accuracy of the lexicons they learned and on their inferences regarding the speakers' intent. The Bayesian model substantially outperformed the comparison models (all results in Figure 2).[3] We systematically varied the three free parameters of the model (, #, and !) and found that the performance of our model was relatively robust across a range of values (though we report results from the values that produced the best F-score, as with the comparison models). Both the simple statistical models and the translation model found a large number of spurious pairings; the best lexicons found by these models were considerably larger than the best lexicon found by our model.[4] The difference in precision between our model and the comparison models suggests that the split between referential and non-referential words effectively allowed our model to identify and exclude from the lexicon words which were not consistently used referentially. We next examined how well the lexicons learned by each model allowed the model to infer the speaker's intended referents relative to the intended referents found by a human coder. For the comparison models, we assumed that the speaker's intention was the set of objects for which the matching words in the model's lexicon had been uttered. Again we found a substantial increase in precision by the Bayesian model, suggesting that the more precise lexicon found by our model lead to even greater gains in the precision of the model's interpretations.

Prediction of Experimental Results

Rapid cross-situational word learning. Recent work by both Yu & Smith and Vouloumanos has provided strong evidence that both adults and children are able to learn associations between words and objects even in the absence of individually unambiguous trials (Smith & Yu, in press; Vouloumanos, in press; Yu & Smith, 2007). Because the statistics in these experiments so strongly favour the correct lexicon, our model and all of the comparison models successfully found the correct word-object pairings with perfect precision and recall when presented with the artificial lexicons from Yu & Smith (2007).

Mutual exclusivity. In classic demonstrations of mutual exclusivity, a child is presented with two objects, one familiar and one novel. The experimenter asks "can you hand me the dax?" and the child hands over the novel object, indicating that she has correctly inferred that the novel name refers to it (Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; Markman & Wachtel, 1988). Markman and colleagues (Markman, 1989; Markman & Wachtel, 1988; Markman, Wasow, & Hansen, 2003) have suggested that children possess a *principle of mutual exclusivity* which leads them to prefer lexicons with only one label for each object. Other researchers have suggested alternate explanations, including more limited principles that are learned with experience (Golinkoff, Mervis, & Hirsh-Pasek, 1994; Mervis & Bertrand, 1994) or more general pragmatic principles (Clark, 1988; 2002).

Here we examine another possibility: that domain-general principles of probabilistic inference may be sufficient to predict the child's response. The basic intuition driving this hypothesis is that if you consistently hear one label for an object, it is surprising to discover that another label exists (since you have never heard the new label before). For instance, if "dax" meant bird, it would be a very suspicious coincidence that every other time someone wanted to

talk about a bird they chose to say "bird" rather than "dax."

We tested our model in the classic mutual exclusivity paradigm (Markman & Wachtel, 1988) and found it correctly inferred that the novel word mapped to the novel object. To explore this result further we scored four possible lexicons on both our original CHILDES corpus and the mutual exclusivity scenario (Figure 3). Consistent with the intuition above, hypotheses where "dax" was mapped to the familiar object BIRD (C and D) were unlikely with respect to the original corpus because each sentence where the word "bird" was uttered became less likely due to the (unrealized) possibility of hearing "dax" as well. Learning no new words (hypothesis A) was favoured by the prior because it involved no growth in the size of the lexicon, but gave low likelihood to the experimental scenario (since the word "dax" was not in the lexicon). Thus, our model preferred the correct hypothesis (B) because of the assumption that speakers choose uniformly between the possible names in the lexicon for an object; under this assumption, if an object has two names, hearing either name is less likely.

In fact, this result is not unique to our model: the basic finding of mutual exclusivity is captured by many of the baseline models we tested, including the conditional probability, mutual information, and translation models. The success of all of these models—combined with the demonstration that adults and infants are able to use some sort of statistical information in cross-situational learning tasks (Smith & Yu, in press; Yu & Smith, 2007)—strongly suggests that it is not necessary to posit domain-specific principles to account for findings of mutual exclusivity.

One-trial learning. Another classic result in the literature on word learning is the ability of children to learn a new word from only one or a small number of incidental exposures (Carey, 1978; Carey & Bartlett, 1978; Markson & Bloom, 1997). Our model and the comparison models predict that there are some situations that—in conjunction with the learner's previous

experiences—can provide sufficient evidence for a word to be learned after a single exposure; in fact, the experiment described above provides one such situation. We next turn to a set of experiments which to the best of our knowledge cannot be captured by the comparison models.

Object individuation. Even before their first birthday, infants are able to use the presence of words to help individuate objects (Booth & Waxman, 2002; Waxman & Booth, 2003; Waxman & Markow, 1995; Xu, 2002). In one experiment (Xu, 2002), infants saw first a duck and then a ball emerge and then retreat behind a screen. Infants in the two-word condition heard "look, a duck" and then "look a ball" while infants in the one-word condition heard "look, a toy" twice. At test, the screen dropped, revealing either one or two objects. Infants in the one-word conditioned looked longer at two objects (indicating that they expected only one object), while infants in the two-word condition instead looked slightly longer at the single object (indicating that they expected two objects and were surprised that one had disappeared).

Why would hearing two different labels allow infants to make the inference that two different objects were behind the screen? A prediction of our model is that hearing two different words makes the presence of two different objects more likely, if the meaning of the words is unknown. If the participant hears two words but sees only one object, either A) the second word refers to a second object but gets spoken even though the second object isn't present, B) the second word refers to the first object and both are less likely (as in the mutual exclusivity example above), or C) the second word doesn't refer to either and is just uttered by chance. All of these possibilities are much more likely if there are actually two objects behind the screen; thus the prediction of the model is that infants should be surprised if there is only one object. In the condition where only one word is heard, a similar logic applies. If there are two objects, then sometimes a word is being used even though it does not correspond to the object that is being

seen. Thus, the model predicts that there is more likely to be only one object behind the screen.

To simulate these results in our model, we created two situation sets which corresponded to the habituation stimuli for the two conditions of the experiment (one word / two words). For each set of situations, we created two construals of that set: one construal in which there were actually two objects (though only one was ever seen at any given time) and one in which there was only one object behind the screen. In order to simulate the infant's uncertainty about the meanings of the word (or words) in the experiment, we evaluated each construal on all possible lexicons. We then compared the surprisal of the model—a quantity that has been shown to map model probability to reaction time data (Levy, 2007)—for the two construals of each experimental condition (e.g., two words, one object vs. two words, two objects). This comparison can be interpreted as measuring, for a learner with no knowledge of what the words mean, how much more surprising it would be to find one object as opposed to two behind the screen. We found a cross-over interaction—higher surprisal when the number of words did not match the number of objects—mirroring the results found by Xu (2002) (shown in Figure 4), suggesting that the encoding of intention within our model allowed it to create expectations about the correspondence between words and objects even in the absence of knowledge about the meanings of the words.

Intention-reading. Baldwin (1993) conducted an experiment in which 19-month-old toddlers were shown two opaque containers, each containing a different novel toy. The experimenter opened one container, named the toy inside without showing the child the contents of the container, gave the child the toy from the second container to play with, and then finally gave the child the first (labelled) object. Despite the greater temporal contiguity between the label and the second toy, the children showed evidence of learning that the label corresponded to

the first toy. Baldwin interpreted these results as evidence that the children used the experimenter's referential intention as their preferred guide to the meaning of the novel label. Our model, built around inferring the speaker's intended referents, incorporates this conclusion directly into its structure. To test our model on Baldwin's task, we constructed a situation with two novel objects and a single novel word. If the model was given the information that the speaker intended to refer to the first object, it highly preferred the correct pairing.

This result might not seem surprising: we have coded the referential intention of the experimenter into our simulation. But a model which does not incorporate a representation of referential intent will be unable to predict Baldwin's results: models which rely directly on perceptual salience do not capture this result since the object which must be more salient for the correct mapping to occur is actually out of sight when it is being labelled.

General Discussion

We have presented a Bayesian model which unifies cross-situational and intentional approaches to word learning. This model performs well in learning words from a small, naturalistic corpus; in addition, it predicts a variety of empirical phenomena in word learning. Previous work has examined the experimental coverage of computational models of word learning (Regier, 2005) or their performance on a corpus (Yu & Ballard, 2007), but ours is the first to our knowledge to perform both tasks. Our model operates at the highest of Marr's (1982) levels of explanation: it has an explicit structure in which the learner's assumptions are manifest as a set of relationships between observed and unobserved variables. Thus, in defining our model, we make no claims about the nature of the mechanisms that might instantiate these relationships in the human brain. Instead, our goal is to show that substantial gains in corpus performance and fit to experimental data can be achieved by combining probabilistic, cross-

situational inference with an explicit representation of speakers' intended meanings.

However, this kind of ideal-observer analysis is only one part of a full account of early word learning, and many other computational models provide insights into different aspects of this process (Colunga & Smith, 2005; Gold & Scassellati, 2007; Li, Zhao, & MacWhinney, 2007; Regier, 2005; Xu & Tenenbaum, 2007; Yu & Ballard, 2007). In particular, we wish to highlight the possibility that an algorithmic or process-level approach such as that of Regier (2005) might provide a way to understand how the abstract inferences described by our model could be cashed out in terms of known psychological mechanisms.

The success of our simple model in predicting a wide range of results supports the hypothesis that specialized principles may not be necessary to explain many of the smart inferences that young children are able to make in learning words. However, our results do suggest that theories of early word learning will achieve greater coverage of experimental phenomena if they explicitly represent speakers' communicative intentions. In this respect, our model is only a small first step, and we hope that our work here will inspire future modellers to use the same type of intentional inference to unite the rich variety of information sources available to young word learners.

References

Augustine, S. (397/1963). *The Confessions of St. Augustine* (R. Warner, Trans.). New York: Penguin Books.

Baldwin, D. (1993). Early referential understanding: Infants' ability to recognize acts for what they are. *Developmental Psychology, 29*, 832-843.

Bloom, P. (2002). *How Children Learn the Meanings of Words*: MIT Press.

Booth, A. E., & Waxman, S. R. (2002). Object functions and object names: Effects on categorization in 14-month-old infants. *Developmental Psychology, 38*, 948-957.

Brown, P. F., Pietra, S. D., Pietra, V. J. D., & Mercer, R. L. (1994). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics, 19*, 263-311.

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

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

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language, 15*, p317-335.

Clark, E. V. (2002). *First Language Acquisition*. Cambridge, UK: Cambridge University Press.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review, 112*, p347-38235.

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

Gold, K., & Scassellati, B. (2007). *A robot that uses existing vocabulary to infer non-visual word meanings from observation.* Paper presented at the 29th Annual Meeting of the Cognitive Science Society, Nashville, TN.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental psychology, 28*, 99-108.

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

Levy, R. (2007). Expectation-based syntactic comprehension. *Cognition, in press*.

Li, P., Zhao, X., & MacWhinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science, 31*, 581-612.

Locke, J. (1690/1964). *An Essay Concerning Human Understanding*. Cleveland: Meridian Books.

Marinari, E., & Parisi, G. (1992). Simulated Tempering: A New Monte Carlo Scheme. *Europhysics Letters, 19*, 451-455.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. Cambridge: MIT Press.

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

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

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature, 385*, 813-815.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*: Henry Holt and Co., Inc. New York, NY, USA.

Mervis, C. B., & Bertrand, J. (1994). Acquisition of the Novel Name-Nameless Category (N3C) Principle. *Child Development, 65*, 1646-1662.

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

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

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

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

Smith, L. (2000). Learning how to learn words: An associative crane. *Becoming a word learner: A debate on lexical acquisition*, 51–80.

Smith, L., & Yu, C. (in press). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*: Harvard University Press.

Vouloumanos, A. (in press). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*.

Waxman, S. R., & Booth, A. E. (2003). The origins and evolution of links between word learning and conceptual organization: New evidence from 11-month-olds. *Developmental Science, 6*, 130-137.

Waxman, S. R., & Markow, D. R. (1995). Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cognitive Psychology, 29*, 257-302.

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

Xu, F., & Tenenbaum, J. B. (2007). Word Learning as Bayesian Inference. *Psychological Review, 114*, 245-272.

Yu, C., & Ballard, D. (2007). A unified model of word learning: Integrating statistical and social cues. *Neurocomputing, 70*, 2149-2165.

Yu, C., & Smith, L. (2007). Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. *Psychological Science, 18*, 414-420.

Footnotes

1. These videos are the same as those used by Yu and Ballard (2007) although the annotations are our own.

2. F-score is the harmonic mean of precision (proportion of the lexicon that was correct) and recall (proportion of total correct pairings included in the lexicon).

3. Precision and recall scores were computed relative to a *gold-standard* lexicon created by a human coder; this lexicon incorporated all standard word-object pairings (for a lamb toy, "lamb"), plurals ("lambs"), babytalk ("lambie"), and onomatopoeia ("ba ba").

4. The performance we report for the translation model is considerably lower than that reported by Yu and Ballard (2007). Several factors may have contributed to this difference: the speech transcripts used in our study were taken from CHILDES, while those in the Yu & Ballard study were automatically extracted and may have contained shorter utterances; our corpus coding may have included more objects in each situation; and our gold-standard lexicon differs from Yu and Ballard's. All of these factors should penalize all models equally, however.

Figure Captions

Figure 1.

The graphical model representing dependence relations in our model. *O*, *I,* and *W* represent the objects present in the context, the objects that the speaker intends to refer to, and the words that the speaker utters, respectively. These variables are related within each situation *S* (shown by the plate under these variables), and the words that the speaker utters are additionally determined by the lexicon of their language, *L*, which does not change from situation to situation (and hence lies outside of the plate).

Figure 2.

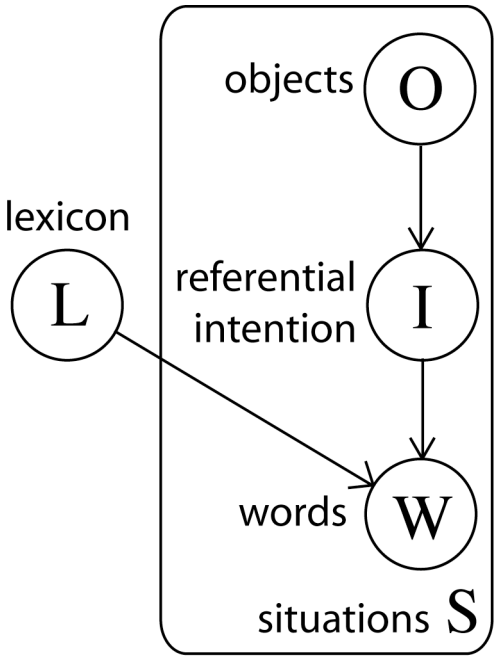
The best lexicon found by the model, a comparison of the performance of different models on the CHILDES corpus, a comparison of the performance of different models on their fit to the referential intents found by a human coder, and two examples of situations with correctly inferred intentions. For the best lexicon, entries judged to be correct according to the gold-standard are in bold. In the model comparison, reported scores are for the best lexicon found by each model.

Figure 3.

Schematic depiction of possible hypotheses in a mutual exclusivity experiment. If the experimenter utters the novel word "dax" in the presence of a novel object (a DAX) and a known object (a BIRD), the learner can decide the word refers to both, one or the other, or neither. Each panel represents one of these options, with a line between a word and an object signifying that the link is represented in the lexicon. The corpus likelihood, the likelihood of the experimental situation, the prior probability of the lexicon, and the posterior (total) probability, normalized across the four lexicons, are shown for each hypothesis.

Figure 4.

Data from Xu (2002), Experiment 1 (on the use of labels to individuate objects) and model surprisal (negative log probability) in the four conditions of Xu's experiment. Our model predicts the results of infants in these studies, mirroring the interaction in looking times between the number of objects seen and number of labels heard.


objects
O
lexicon
L
referential
intention
I
words
W
situations
S

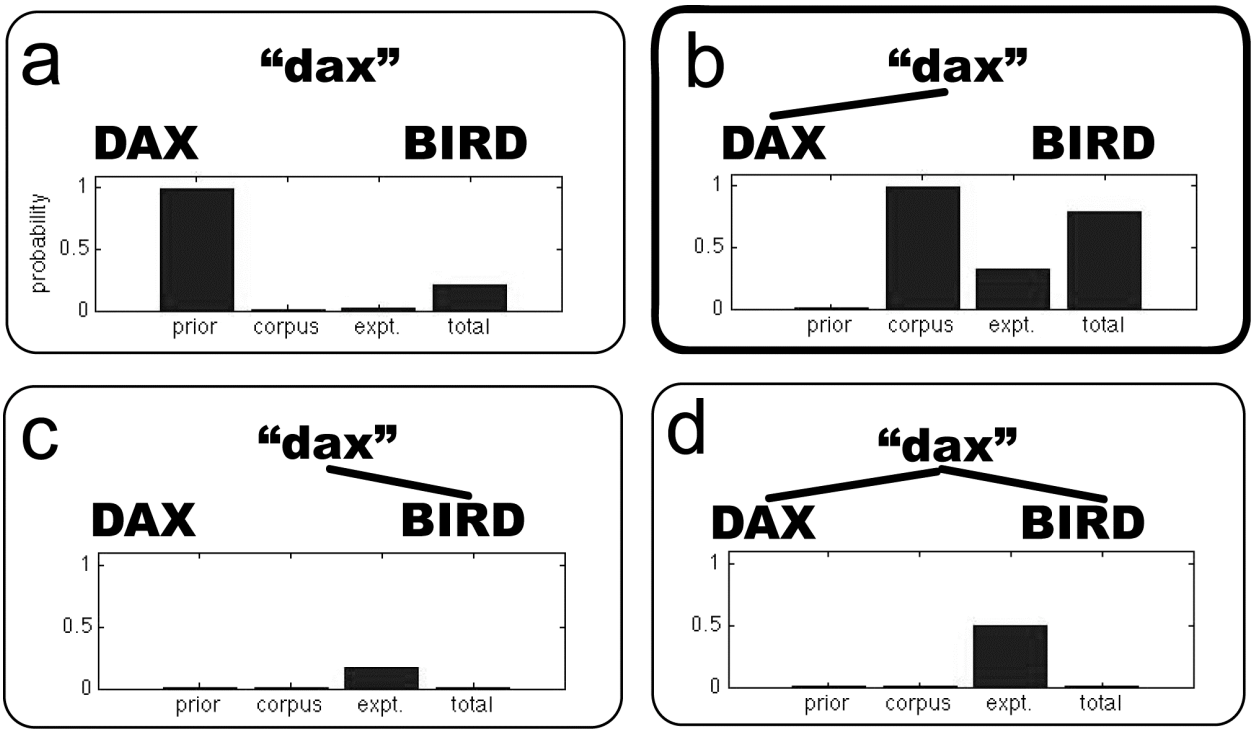
Best lexicon	
Word	*Object*
bear	**bear**
bigbird	**bird**
bird	**duck**
birdie	**duck**
book	**book**
bottle	bear
bunnies	**bunny**
bunnyrabbit	**bunny**
hand	**hand**
hat	**hat**
hiphop	mirror
kittycat	**kitty**
lamb	**lamb**
laugh	cow
meow	baby
mhmm	hand
mirror	**mirror**
moocow	**cow**
oink	**pig**
on	ring
pig	**pig**
put	ring
ring	**ring**
sheep	**sheep**

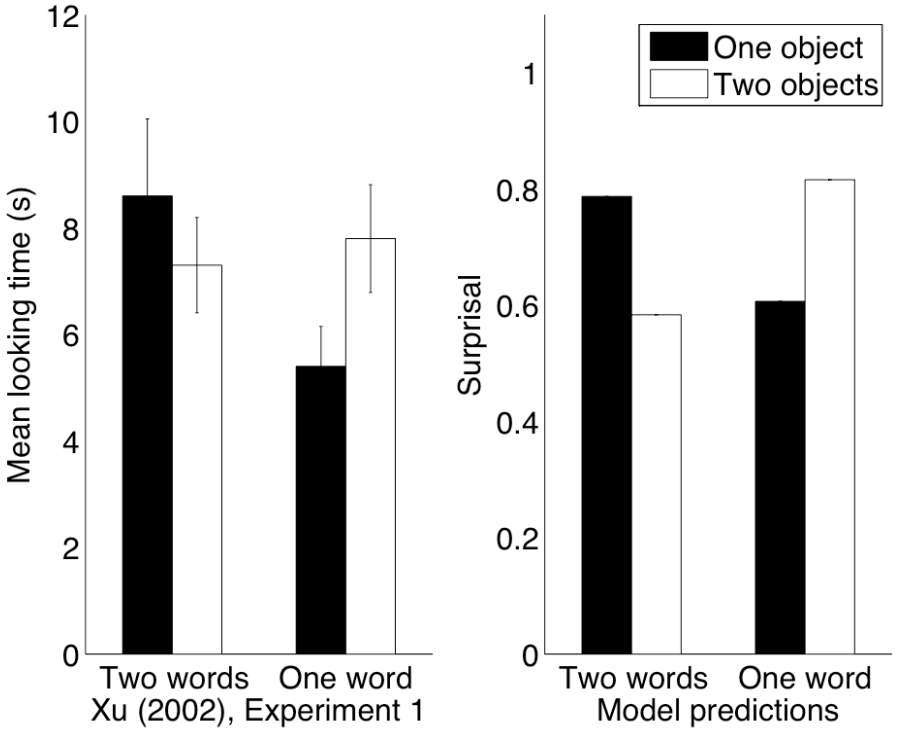
Lexicon	**Precision**	**Recall**	**F-score**
Association frequency	0.07	0.30	0.12
CP (object | word)	0.07	0.19	0.10
CP (word | object)	0.08	0.24	0.12
Mutual information	0.15	0.32	0.21
Yu & Ballard (object | word)	0.08	0.30	0.13
Yu & Ballard (word | object)	0.19	0.43	0.26
Bayesian model	**0.71**	**0.46**	**0.56**

Inferred intentions	**Precision**	**Recall**	**F-score**
Association frequency	0.27	0.52	0.36
CP (object | word)	0.32	0.43	0.37
CP (word | object)	0.33	0.49	0.40
Mutual information	0.37	0.03	0.06
Yu & Ballard (object | word)	0.57	0.52	0.54
Yu & Ballard (word | object)	0.43	**0.54**	0.48
Bayesian model	**0.88**	0.42	**0.57**

Example situations and inferred intentions

Words	"Do bunnies go jumping through the forest?"
Objects	BOOK, BIRD, RATTLE, MIRROR, BUNNY
Intent	BUNNY
Words	"You're not much into sitting up in that chair ..."
Objects	BIRD
Intent	NO OBJECT IN CURRENT SITUATION


a
"dax"
DAX
BIRD
probability
1
0.5
0
prior
corpus
expt.
total
b
"dax"
DAX
BIRD
1
0.5
0
prior
corpus
expt.
total
c
"dax"
DAX
BIRD
1
0.5
0
prior
corpus
expt.
total
d
"dax"
DAX
BIRD
1
0.5
0
prior
corpus
expt.
total


12
10
8
6
4
2
0
Mean looking time (s)
Two words
One word
Xu (2002), Experiment 1
One object
Two objects
1
0.8
0.6
0.4
0.2
0
Surprisal
Two words
One word
Model predictions