COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 34 (2010) 1017–1063
Copyright © 2010 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2010.01104.x

A Probabilistic Computational Model of Cross-Situational Word Learning

Afsaneh Fazly,[a] Afra Alishahi,[b] Suzanne Stevenson[a]

[a] *Department of Computer Science, University of Toronto*
[b] *Department of Computational Linguistics, Saarland University*

Received 14 November 2008; received in revised form 22 December 2009; accepted 04 January 2010

Abstract

Words are the essence of communication: They are the building blocks of any language. Learning the meaning of words is thus one of the most important aspects of language acquisition: Children must first learn words before they can combine them into complex utterances. Many theories have been developed to explain the impressive efficiency of young children in acquiring the vocabulary of their language, as well as the developmental patterns observed in the course of lexical acquisition. A major source of disagreement among the different theories is whether children are equipped with special mechanisms and biases for word learning, or their general cognitive abilities are adequate for the task. We present a novel computational model of early word learning to shed light on the mechanisms that might be at work in this process. The model learns word meanings as probabilistic associations between words and semantic elements, using an incremental and probabilistic learning mechanism, and drawing only on general cognitive abilities. The results presented here demonstrate that much about word meanings can be learned from naturally occurring child-directed utterances (paired with meaning representations), without using any special biases or constraints, and without any explicit developmental changes in the underlying learning mechanism. Furthermore, our model provides explanations for the occasionally contradictory child experimental data, and offers predictions for the behavior of young word learners in novel situations.

Keywords: Word learning; Child language acqusition; Computational modeling; Cross-situational learning

Correspondence should be sent to Afsaneh Fazly, Department of Computer Science, University of Toronto, 10 King's College Road, Toronto, ON, Canada M5S 3G4. E-mail: afsaneh@cs.toronto.edu, afsaneh.fazly@gmail.com

1. Acquiring a lexicon

An average 6-year-old child knows over 14,000 words, most of which s/he has learned from hearing other people use them in noisy and ambiguous contexts (Carey, 1978). To better appreciate the significance of children's efficiency at such a complex task, let's repeat here the classic example by Quine (1960). A linguist visiting a culture with a language different from her own observes a rabbit scurrying by, while a native says ''gavagai.'' To understand what the word *gavagai* means in the new language, the linguist would have to figure out which part of the scene (if any) is relevant to the meaning of the word. For example, *gavagai* may mean rabbit, it may refer to the action performed by the rabbit, it may have been used to catch the linguist's attention (as in ''Look!''), or may mean something totally irrelevant to what the linguist has observed, for example, ''sky.'' Similarly, children learning their native language need to map the words they hear to their corresponding meanings in a scene they observe. In such a situation, the learner may perceive many aspects of the scene that are unrelated to the utterance they hear (the problem of *referential uncertainty*). Also, the input might be *noisy* due to some error in the perception or interpretation of the heard utterance or the observed scene; for example, not all aspects of the utterance meaning may be directly observable from the scene. In addition, the learner must resolve the *alignment ambiguity*, that is, which word in the utterance refers to which part of the scene.

Clearly, acquiring the meaning of words is an extremely challenging task children encounter early in life. Nonetheless, they eventually learn the words of their language reasonably quickly and effortlessly. Much research has thus focused on trying to better understand what mechanisms and skills underlie children's impressive performance in word learning. Psycholinguistic studies have attempted to explain children's success at this difficult task through examining specific patterns that are observed in the course of lexical acquisition in children. These patterns include the vocabulary spurt (i.e., a slow stage of word learning, followed by a sudden increase in the learning rate), and fast mapping (i.e., the ability to map a novel word to a novel object in a familiar context), among others. Many theories have been proposed to account for these patterns, each suggesting specific word learning mechanisms or dedicated mental biases that help children learn the meanings of words (e.g., Behrend, 1990; Golinkoff, Hirsh-Pasek, Bailey, & Wegner, 1992; Markman & Wachtel, 1988). As a result, the literature contains a variety of such mechanisms and biases, sometimes overlapping or even inconsistent with each other. What is lacking is a unified model of word learning that brings together the suggested mechanisms and biases, and that accounts for the various aspects of the process, including the above-mentioned patterns. Section 1.1 further elaborates on the psycholinguistic theories of early lexical development in children, as well as on our proposed framework for modeling early vocabulary acquisition.

Computational modeling is a powerful tool for the precise investigation of the hypothesized mechanisms of word learning; we can carefully study whether a computational model that is based on a suggested theory or learning mechanism (and is tested on naturalistic data) shows a pattern of behavior similar to those observed in children. Many computational models of word learning have been developed to simulate and account for the observed patterns, such as fast mapping and the vocabulary spurt. Most of the existing models, however, use simplified input data that significantly deviate from the naturalistic input children receive from their environment. Some use data that do not have the properties explained above—-noise, alignment ambiguity, and referential uncertainty (e.g., Li, Farkas, & MacWhinney, 2004; Regier, 2005), whereas others test their models on artificially generated or on very limited input (e.g., Horst, McMurray, & Samuelson, 2006; Siskind, 1996). In addition, not all proposed models incorporate cognitively plausible learning mechanisms (e.g., Frank, Goodman, & Tenenbaum, 2007; Yu, 2005). Section 1.2 provides more detailed descriptions of existing computational models, identifying some of their limitations, and explaining how our proposed model attempts to address these shortcomings.

1.1. Psycholinguistic theories of child lexical development

An important aspect of learning the meaning of a word involves associating a certain mental representation, or concept, with a word form. Some psychologists consider word learning, especially at early stages, to be based on simple associative mechanisms (Smith, 2000): A child hears a word, for example *dog*, while chasing a dog. The child associates the word *dog* with the concept of a ''dog'' after repeatedly being exposed to similar situations. However, not all natural word learning situations are as simple as the one depicted above. As noted by Carey (1978), children learn most of their vocabulary from hearing words used in noisy and ambiguous contexts.[1] In such cases, there are infinitely many possible mappings between words and concepts. Some researchers thus suggest that children use a variety of attention mechanisms to narrow down parts of the scene described by an utterance, and to focus on the referred objects (referential learning). For example, Carpenter, Nagell, Tomasello, Butterworth, and Moore (1998) and Bloom (2000) argue that children use their (innate or acquired) social skills to infer the referent of a word as intended by a speaker. Similarly Smith, Yu, and Pereira (2007) propose the use of embodied cognition in focusing on the intended portion of a scene described by an utterance.

Most of the above mechanisms only apply to cases where a direct and deliberate dialogue is taking place between a child and her caretaker, and do not explain learning from the vast amount of noisy and ambiguous input that children receive from their environment (see Hoff & Naigles, 2002). A powerful and plausible mechanism for dealing with noise and referential uncertainty is cross-situational learning. It has been suggested that children learn the correct mappings between words and their meanings from the huge number of possibilities by observing the regularities across different situations in which a word is used (Gleitman, 1990; Pinker, 1989; Quine, 1960). The cross-situational learning mechanism suggests that the meaning of a word is consistent across different occurrences of it, and can be learned by detecting the set of meaning elements that are common across all usages of the word.

In their original forms, the associative and the cross-situational mechanisms are not precisely specified. Moreover, these hypotheses are not sufficient for explaining the particular developmental patterns (e.g., fast mapping) observed in the experimental data gathered from children. Many researchers thus believe that, in addition to relying on associationist and cross-situational evidence, children are equipped with dedicated mental biases and/or constraints that help them learn the meanings of words (e.g., Behrend, 1990). The literature proposes a variety of such biases and constraints, each accounting for one (or a few) of the observed patterns. For example, the fast mapping ability in children has been suggested to be due to the principle of the mutual exclusivity of word meanings (Markman & Wachtel, 1988), or due to a lexical bias towards finding names for nameless objects/categories (Golinkoff et al., 1992). Other patterns such as vocabulary spurt, or an initial reluctance towards learning a second label for a familiar object (synonymy), are sometimes attributed to a change in the underlying learning mechanism. For instance, it has been suggested that children learn the meaning of their first words through a simple associative process, and later switch to referential learning, which allows them to learn new words at a faster pace and to learn synonyms (e.g., Behrend, 1990; Kamhi, 1986; Reznick & Goldfield, 1992).

Although many specific word learning biases and constraints have been proposed, it has yet to be proven whether and to what extent children depend on them for learning the vocabulary of their language. Indeed, there are many researchers who argue against the necessity of such mechanisms and biases for word learning, and suggest that word meanings are acquired through general cognitive abilities (e.g., Bloom, 2000; Tomasello, 2003). Proponents of this view believe that the patterns of word learning observed in children (such as the vocabulary spurt and fast mapping) are a result of simply receiving more input, and that no developmental changes in the underlying learning mechanisms (e.g., from associative to referential or constraint-based) are necessary (see also Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991; McMurray, 2007; Regier, 2005).

Our goal in the present study is to support this latter view on word learning through computational modeling. We propose a novel model of early vocabulary acquisition that learns word meanings using a general probabilistic approach, without incorporating any specific word learning biases or constraints, and without any explicit developmental changes in the underlying learning mechanisms. Our proposed model learns the meaning of words from naturalistic child-directed data, extracting only very simple probabilistic information to which children have been shown to be sensitive (e.g., Coady & Aslin, 2004). Specifically, the model incorporates a probabilistic interpretation of cross-situational learning and bootstraps its own partially learned knowledge of the previously observed words to accelerate word learning over time. The model exhibits similar behaviours to those observed in children, suggesting that word meanings can be acquired through general cognitive mechanisms.

1.2. Related computational models

The computational model proposed by Siskind (1996) is the first to simulate the process of learning word meanings from ambiguous contexts, and in the presence of noise and referential uncertainty. The model uses cross-situational evidence in conjunction with a set of specific word-learning principles (such as compositionality) to constrain hypotheses about the meaning of each word. In simulations of word learning on artificially generated input, the model exhibits various behavioral patterns observed in children, such as a sudden increase in the rate of vocabulary growth and the acceleration of word learning with exposure to more input. However, the model uses discrete rule-based inference to manipulate sets of ‘‘possible’’ and ‘‘necessary’’ meaning symbols for each word, making it overly sensitive to noise and incomplete data. In particular, Siskind's model incorporates several specific (and at times too-strong) constraints, such as *exclusivity* and *coverage* (to narrow down the set of ''possible'' meanings for a word), and *compositionality* (to handle noise and referential uncertainty). Since these constraints are overly strong, Siskind then needs to devise a specific new mechanism for handling noise and homonymous words.[2] This approach limits the model's adaptability to natural data. For example, it is not possible to revise the meaning of a word once it is considered as ''learned,'' which prevents the model from handling highly noisy data. Moreover, the approach cannot naturally model all the kinds of shifts in meaning that have been observed in children as they gradually glean the full intent of a word (Barrett, 1994), such as moving from a more general meaning to a more specific one (which would require the addition of ''necessary'' meaning primitives that have already been ruled out).

Other computational models incorporate probabilistic interpretations of the cross-situational inference mechanism, enabling them to address some of the shortcomings of a discrete approach to manipulating sets of meaning symbols. Specifically, the flexibility of a probabilistic framework lets a model capture more nuanced associations of meanings with a word and also makes it robust to noisy and incomplete data. For example, the word learning model of Yu (2005) uses an existing algorithm (Brown, Della Pietra, Della Pietra, & Mercer, 1993) to model word–meaning mapping as a probabilistic language translation problem. Variations of this model are used to examine the role of different factors in word learning, such as social cues (Yu & Ballard, 2008) and syntax (Yu, 2006). However, the models proposed by Yu and colleagues (2005, 2006, 2008) are all tested on limited experimental data containing a very small vocabulary, and with no referential uncertainty. Frank et al. (2007) propose a Bayesian model of cross-situational word learning that can also learn which social cues are relevant to determining references of words. Using only domain-general probabilistic learning mechanisms, their model can explain various phenomena such as fast mapping and social generalization. However, their experiments are also performed on a small corpus containing a very limited vocabulary. Moreover, all these models (those used by Frank et al., 2007; Yu, 2005, 2006; Yu & Ballard, 2008) are nonincremental and learn through an intensive iterative batch processing of a corpus.

The Bayesian model of Xu and Tenenbaum (2007) provides insight into how humans learn to generalize category meanings from examples of word usages. Assuming as prior knowledge a probabilistic version of the basic-level category bias (Markman, 1989; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), Xu and Tenenbaum's model learns appropriate category names for exemplar objects by revising the prior bias through incorporating the statistical structure of the observed examples. Although their model shows similar behavior to that of humans performing the same task, the model is tested only in a very specific word learning situation, and on a small sample of object exemplars.

Many connectionist models have also been proposed for learning associations between a word form and its meaning, and for investigating various patterns in the process of learning. An important shortcoming of existing connectionist models is that they rely on a simplified (unnaturalistic) input consisting of pairings of a semantic representation with a single word form (or its phonological representation)—as opposed to full utterances. Regier

(2005), for example, proposes an associative exemplar-based model that accounts for the developmental changes observed in children's word learning, such as fast mapping and learning synonymy, without a change in the underlying learning mechanism. The simulations are performed on small artificially created training and test data in highly controlled conditions. Li et al. (2004, 2007) simulate vocabulary spurt and age of acquisition effects in an incremental associative model. To reduce the interference effect often observed in connectionist models, they specifically incorporate two modes of learning: an initial map organization mode and a second incremental clustering mode to account for vocabulary growth. Horst et al. (2006) focus on fast mapping within a connectionist model of word learning and show that the behavior of their computational model matches child experimental data (as reported in a study by the same authors, Horst & Samuelson, 2008). However, the learning capacity of their model is limited, and the fast mapping experiments are performed on a very small vocabulary. While each of these models investigates an interesting aspect of word learning, they do so using artificial and clean data, which contain no noise or alignment ambiguity or referential uncertainty.

Our proposed computational model of word learning seeks to build on the strengths of earlier approaches, while addressing some of the shortcomings mentioned above. Specifically, we are the first to propose a model that achieves all of the following:

- Our model is founded on a general and cognitively plausible probabilistic learning mechanism.
- The model can handle both alignment ambiguity (i.e., the mapping between words and meanings is not indicated in the input) and referential uncertainty (i.e., many meaning elements are included in the input that are not associated with words in the utterance).
- A single learning mechanism incrementally refines word–meaning associations without getting misled by substantial noise.
- Our model successfully learns word–meaning mappings from large-scale, naturalistic data that more closely resemble the learning environment of children.
- Our model exhibits behavior analogous to that of children in a range of word learning tasks.

The following sections (Sections 2 and 3) explain our proposed computational model in more detail.

2. Overview of our computational model

2.1. Basic assumptions about the learning environment

The main focus of our model is to study word learning in a naturalistic context. We assume that the learner/child is watching a scene while hearing an utterance describing the scene. In realistic situations, this is not always the case: as noted by Gleitman (1990), ''caretaker speech is not a running commentary on scene and events in view'' (see also Bloom, 2000). It is nonetheless reasonable to assume that very young children starting to learn the meanings of words are exposed to many utterances that refer to things and situations in the perceptible scene (Veneziano, 2001). We also assume that when a child hears an utterance while observing a scene, he or she can establish a link between the full utterance and the set of meaning elements inferred from the scene through observation or other means. We thus use pairings of a complete utterance and a set of semantic elements (or a scene representation) as the basic input to our model.

Specifically, we use naturalistic input pairs with properties similar to those of the input children receive from their learning environment. That is, utterance–scene pairs contain alignment ambiguity, referential uncertainty, and noise, as explained here:

- Alignment ambiguity: the mappings between specific words in an utterance and specific meaning elements in the corresponding scene representation are not explicitly marked. (We simply use the term *ambiguity* to refer to the alignment ambiguity in word learning. To refer to lexical ambiguity—that a word type may have more than one meaning in a lexicon—we use the term *homonymy*.)
- Referential uncertainty: the representation of a scene may contain meaning elements that are not relevant to the corresponding utterance.
- Noise: an utterance may contain words whose appropriate meanings are not included in the representation of the corresponding scene. (Note that this models only one type of noise, in which the child is unable to perceive the meaning of the word in the scene. In particular, we do not assume noise in perception of the utterance, that is, every word is assumed to be perceived clearly.)

In summary, it is not explicitly indicated in the input which word refers to which meaning element (alignment ambiguity). Furthermore, although the child is assumed to hear each word in the utterance, the scene representation may contain ''extra'' meaning elements that do not correspond to words in the utterance (referential uncertainty), and the scene representation may be missing meaning elements for some words in the utterance (noise). Fig. 1 presents such an input, where a child hears the utterance *Joe is happily eating an apple*, while perceiving that ''Joe is quickly eating a big red apple with his hands.''

In modeling learning in the presence of referential uncertainty, we assume that the potentially huge space of possible meanings for each utterance has been considerably reduced through some attentional mechanism. Many such mechanisms have been shown to be used by children in order to focus on a small subsection of the complex scenes in the real world, such as embodied cognition (e.g., Smith et al., 2007), using social cues such as eye gaze and gesture (e.g., Baldwin, et al., 1996; Kalagher & Yu, 2006), or incorporating skills of social cognition and theory of mind for understanding the intention of the speaker (Bloom,

Utterance: *Joe is happily eating an apple*
Scene: {joe, quickly, eat, a, big, red, apple, hand}

Fig. 1. A sample input utterance–scene pair.

2000; Carpenter et al., 1998). Although we assume that such a mechanism is in play prior to the selection of the scene representation in our input data, we do not make any claims on the nature of this attention mechanism. Moreover, we assume that although the use of such a mechanism helps the learner to focus on a set of possibly relevant concepts or objects or events in the scene, much uncertainty still remains. Section 4 provides details on how we simulate referential uncertainty in the input.

To disentangle the problem of word learning from other acquisition problems, we make several simplifying assumptions in our model. Learning the meaning of a word in our model is restricted to the acquisition of associations between a word form (e.g., *ball*) and a symbol (`ball`) specifying either a concept or the referent of the word in the real world. Currently in our model, we do not distinguish between the referent of a word, which is an object or an event in the real world, and a concept that is an internal mental representation of the word's meaning. We thus use the terms *meaning* and *referent* (or *object*) interchangeably throughout the paper, and we use the same symbol (e.g., `ball`) for both. Although syntactic and morphological properties of a word (such as its part of speech or case marking), as well as its relation to other words, are also considered as part of the word's meaning (Carey, 1978; Gleitman, 1990; Gleitman & Gillette, 1994), here we do not address the acquisition of such properties.

We also assume that the (nontrivial) task of word segmentation is performed prior to word learning (Aslin, Saffran, & Newport, 1998; Johnson & Jusczyk, 2001; Jusczyk & Aslin, 1995; Mattys Jusczyk, Luce, & Morgan, 1999).[3] In addition, we assume that by the time children start to learn word meanings, they can form conceptual representations from the perceived scenes (Golinkoff, Hirsh-Pasek, Mervis, Frawley, & Parillo, 1995; Mandler, 1992). That is, both the input utterance and the scene representation are broken down into appropriate units (i.e., words and meaning elements). Both of these tasks are most likely interleaved with word learning: It has been shown that partial knowledge of word meaning is used in speech segmentation (Brent, 1996) and that learning word meanings contributes to the formation of concept categories (Bowerman & Choi, 2003; Choi & McDonough, 2007). However, in this paper, we study word learning as an isolated process of mapping words to their meanings.

Finally, in processing utterance–scene pairs, we represent words in their root form and ignore the syntactic properties of the sentence. Morphology and syntax are valuable sources of knowledge in word learning, and it has been shown that children are sensitive to morphological and syntactic cues from an early age (Fisher, 1996; Gertner, Fisher, & Eisengart, 2006; Naigles, 1990; Naigles & Kako, 1993). In fact, it has been argued that the meaning of some verbs cannot be learned through cross-situational learning only, and the knowledge of syntax is vital for their acquisition (Gentner, 1978; Gleitman, 1990). For example, many verbs describe a particular perspective on events that cannot be inferred merely by cross-situational analysis (e.g., ''buying'' and ''selling'' almost always happen at the same time). Future work will need to integrate these information sources into the model.

2.2. Overview of the learning algorithm

We define word meaning as a probabilistic association between a word form and a concept. These associations (or word meanings) are learned based on a probabilistic interpretation of cross-situational learning. Experimental data on children suggest that they are sensitive to cross-situational statistics, and that they use such information in word learning (Forbes & Farrar, 1995; Smith & Yu, 2007).

We attempt to find the best mapping between each word and each meaning element from a sequence of utterance–scene pairs similar to the pair presented in Fig. 1. We view this task as analogous to learning a bilingual word-list that contains the equivalences between words in two different languages. The word learning algorithm we propose here is thus an adaptation of an existing model for automatic translation between two languages: the IBM Translation Model 1, originally proposed by Brown et al. (1993). Unlike the original model (and the version used by Yu, 2005 as a computational model of word learning), our adaptation is incremental and does not require an iterative batch process over an entire set of input pairs.

The model maintains a meaning representation for each word as a probability distribution over all of the possible meaning elements. We refer to this distribution as the *meaning probability* of the word, and we refer to the probability of an individual meaning element in this distribution as the meaning probability of that element for the word. In the absence of any prior knowledge, all meaning elements are equally likely to be the meaning of a word. Hence, prior to receiving any usages of a given word, the model assumes a uniform distribution over meaning elements as its meaning. The input pairs are processed one by one and discarded after being processed. After processing each input pair, the meaning probabilities for all the words in the current utterance are updated.

As the first step in processing an input pair, the meaning/referent of each word in the utterance must be determined from the corresponding scene—that is, words in the utterance must be *aligned* with the meaning elements in the scene. Our model does so through calculating an *alignment probability* for each word in an utterance and each meaning element in the corresponding scene. Fig. 2 depicts some hypothetical alignments established between words and meaning elements in the utterance–scene pair of Fig. 1. Each alignment between a word and a meaning symbol is shown as a line whose thickness indicates the strength of the alignment (i.e., the value of the alignment probability).

To calculate the alignment probabilities, we use the partially learned knowledge of the model about the meanings of words (reflected in their meaning probabilities). That is, the probability of aligning a meaning element and a word is proportional to the meaning probability of that meaning element for the word. In addition, we assume that words in an utterance tend to contribute nonoverlapping elements in the corresponding scene. In other words, if there is evidence in the meaning probabilities (prior to receiving the current input pair)

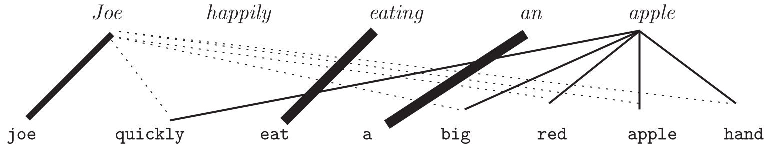


Fig. 2. Sample alignments between words in an utterance and meaning elements in the corresponding scene representation. Thickness of a line indicates the strength of the established alignment; dashed lines represent very weak alignments. For readability, only a subset of the alignments are shown.

that a meaning element in the current scene is strongly associated with a word in the current utterance, it is less likely for the meaning element to be (strongly) aligned with another word in the same utterance. Fig. 2 presents a situation where the model encounters an utterance including some familiar words (e.g., *Joe*, *an*, *eating*) and some novel ones (e.g., *apple*). For two of the familiar words, *an* and *eating*, the model has learned strong associations between the word and its correct meaning, and hence establishes high-confidence alignments between the two (shown as very thick lines). For a word whose meaning is not learned yet (e.g., *apple*), uniform (and weak) alignments are established between the word and those meaning elements that are not strongly aligned to any other word in the utterance (here, quickly, big, red, apple, and hand). Intuitively, the model assumes that all five of these elements are equally likely to be the meaning of the novel word *apple*, and that it is not very likely that the other elements (e.g., joe, eat, a) are the meaning of this word. Even though the model has previously seen the word *Joe* co-occurring with its meaning, it has not yet established a reliable association between the two. Thus, the model establishes a somewhat strong alignment between *Joe* and its meaning, but also some weaker alignments between the word and the novel meaning elements in the scene representation (shown as dashed lines).[4]

As the second step of processing an input pair, the meaning probabilities of the words in the current utterance are updated according to the accumulated (probabilistic) evidence from prior co-occurrences of words and meaning elements (reflected in the alignment probabilities). This evidence is collected by maintaining a running total of the alignment probabilities over all input pairs encountered so far. The running total for a word and a meaning element—referred to as the association score between the two—is increased by their alignment probability (a value between 0 and 1) every time the two appear together in an input pair. In other words, each time a word and a meaning element appear in an input pair together, we add to their association score a probability that reflects the confidence of the model that their co-occurrence is indeed because the meaning element is associated with the word. In summary, in this step the model updates the association scores for all of the words and meaning elements in an input pair based on the calculated alignment probabilities for that pair, and then revises the meaning probabilities of the words in the utterance accordingly. Fig. 3 shows two sample meaning probability distributions after processing the input pair presented in Fig. 2. For the word *eating* whose meaning has already been learned by the model, the meaning probability distribution is skewed towards the correct meaning element eat. The meaning probability distribution for the novel word *apple* shows that its meaning is not

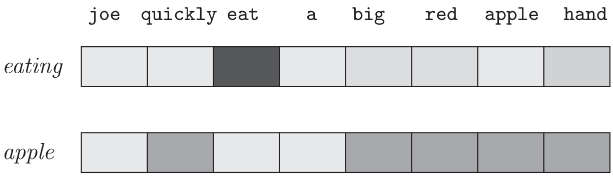


Fig. 3. A grayscale representation of the meaning probability distributions for the words *eating* and *apple*. Intensity of the color of each small box shows the confidence of the model in mapping a specific symbol to a specific word in the learned lexicon.

learned yet, but also that the model has formed a probabilistic assumption about the possible meanings of the word.

The two steps explained above are repeated for all input pairs, one at a time. Fig. 4 presents an example of how the model learns the meaning of a word by processing several input pairs containing usages of the word. The figure depicts the change in the meaning probability distribution for the word *ball* after processing each of the six utterance–scene pairs given in the top portion of the figure. Utterances are all taken from the CHILDES database (MacWhinney, 2000); see Section 4.1 for more details. (Note that the scene representations contain irrelevant meaning symbols, simulating referential uncertainty. Also, noise is added to the fourth input pair by removing the meaning element `ball` from the scene representation.) At first, all symbols are equally likely to be the meaning of *ball*, albeit with a very small probability (*t*=0, not shown in the figure). After receiving the first input pair (*t*=1), the meaning probability of *ball* slightly increases for those symbols appearing in the scene and slightly decreases for other (unseen) symbols (note the difference in the intensity of the colors for the observed symbols and for the unseen ones). Processing the second input pair causes an increase in the probability of symbols that are common between the first and the second input pairs (i.e., `a, ball, be`) and a decrease in the probability of the other symbols. After receiving an input in which *ball* co-occurs with `ball`, but not the other symbols initially in common, the meaning probability of *ball* becomes more skewed towards its correct meaning `ball` (*t*=3). Note that receiving a noisy input pair (*t*=4) does not overly mislead the learner: The learning process may be slowed down (the meaning probabilities do not change substantially between *t*=3 and *t*=4), but with additional input in which `ball` and *ball* co-occur, the meaning of `ball` for *ball* becomes stronger (*t*=5 and *t*=6).

Utterance	**Scene**
shall we find you a ball	{`shall, we, find, you, a, ball, oh, here, be`}
with a ball	{`with, a, ball, that, be, right`}
the ball there	{`the, ball, there, and, what, about, boat`}
get your other ball under there look	{`get, your, other, under, there, look, cooker`}
the ball what	{`the, ball, what, touch, it`}
do you kick the ball	{`do, you, kick, the, ball, what, else`}

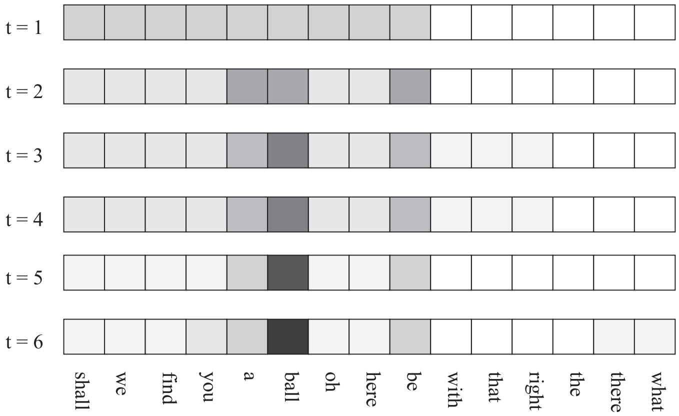


Fig. 4. A trace over time of the meaning probability distribution for the word *ball*.

Note that the ability to recover from a noisy input holds even if the very first usage of a word is noisy—that is, it does not contain the correct meaning symbol for that word. Although not shown, if the fourth (noisy) input pair in the above example occurred first in the sequence, the model would still eventually learn the correct meaning of *ball*. In such a case, the model would initially assign for *ball* a relatively high probability to the (irrelevant) meaning elements observed in the corresponding scene. However, this does not rule out the possibility of the previously unobserved correct meaning gaining in probability later. Further exposure to *ball* in the presence of `ball` will cause the model to adjust the probabilities and overcome the initial noise. Note further that this situation requires no special processing mechanism or recognition of an ''error'' on the part of the model (compare Siskind, 1996). Indeed, this situation is completely analogous to the behavior in the example above, in which the model gradually decreases the probability of the irrelevant meanings that *ball* is initially associated with, and it increases its probability with the more consistently associated (correct) meaning.

3. Details of the probabilistic model

3.1. Utterance–scene input pairs

The input to our word learning model consists of a sequence of utterance–scene pairs that link a scene representation (what the child perceives or conceptualizes) to the utterance that describes it (what the child hears). We represent each utterance as a set of words, and the corresponding scene as a set of meaning symbols, as in:

1. $\mathrm{U}^{(t)}$: *Joe is quickly rolling a ball*
   $\mathrm{S}^{(t)}$: { `joe, happy, roll, a, red, ball, hand, mommy, talk` }

where the superscript t stands for the time at which the current input pair is received—that is, t uniquely identifies the current input pair. $\mathrm{U}^{(t)}$ stands for the current utterance, and $\mathrm{S}^{(t)}$ for the current scene. The above pair represents a situation where a child hears the utterance *Joe is quickly rolling a ball*, while perceiving that ''Joe is happily rolling a red ball with his hand while talking to his mom.'' (Note that the word *quickly* has no correct meaning element in the scene representation (noise), and there are a number of meaning elements that do not correspond to words in the utterance (referential uncertainty).) Section 4 provides details on how the utterances and the corresponding meaning symbols are selected to form the input pairs.

3.2. Word–meaning associations

Given a corpus of pairings between utterances and their corresponding scene representations, our model learns the meaning of each word w as a probability distribution $p(.|w)$ over all the meaning symbols appearing in the corpus. In this representation, $p(m|w)$ is the probability of a symbol m being the meaning of a word w, reflecting the strength of the association between m and w. As the learning proceeds, the meaning probability distribution for a word w is expected to become skewed towards the symbol m_w that is the ''correct'' meaning of w. For example, if the model has learned the correct meaning of the word *ball*, we expect $p(\texttt{ball}|ball)$ to be very high (close to 1), and $p(m|ball)$ for every m other than `ball` to be very low (close to 0). The final grayscale diagram in Fig. 4 (at t=6) depicts the meaning probability for the word *ball* when the meaning of the word is considered to be learned by the model.

3.3. The algorithm

Step 1: Calculating the alignment probabilities.

Recall from Section 2 that for a given utterance–scene pair, $\mathrm{U}^{(t)}$–$\mathrm{S}^{(t)}$, the likelihood of aligning a symbol in the scene with a word in the utterance is proportional to the meaning probability of the given symbol for the word. In addition, we assume that the words in $\mathrm{U}^{(t)}$ are more likely to contribute nonoverlapping portions of the meaning represented in $\mathrm{S}^{(t)}$: A meaning symbol in the scene is likely to be *strongly* aligned with no more than one of the words in the corresponding utterance.[5] More formally, for a symbol $m \in \mathrm{S}^{(t)}$ and a word $w \in \mathrm{U}^{(t)}$, the higher the probability of m being the meaning of w (according to $p(m|w)$ at the time of receiving the current input pair), the more likely it is that m and w are aligned in the current input. In other words, the likelihood of aligning w with m in the current input pair, $a(w|m,\mathrm{U}^{(t)},\mathrm{S}^{(t)})$, is proportional to $p^{(t-1)}(m|w)$. Moreover, if there is strong evidence that m is the meaning of another word in $\mathrm{U}^{(t)}$—that is, if $p^{(t-1)}(m|w')$ is high for some $w' \in \mathrm{U}^{(t)}$ other than w—the likelihood of aligning m to w should decrease. Combining these two requirements:

$$a(w|m,\mathrm{U}^{(t)},\mathrm{S}^{(t)}) = \frac{p^{(t-1)}(m|w)}{\sum\limits_{w' \in \mathrm{U}^{(t)} \cup \{d\}} p^{(t-1)}(m|w')} \tag{1}$$

where $a(w|m,\mathrm{U}^{(t)},\mathrm{S}^{(t)})$ stands for the probability of aligning w and m in the current utterance–scene pair, and d represents a dummy word that is added to the utterance as a smoothing factor, prior to calculating the alignment probabilities. The denominator is a normalizing factor (to get valid probabilities) that also has the effect of decreasing the alignment probability for w if other words w' have a high probability for m.

By adding the dummy word, we do not require that each meaning element from the scene be aligned with a word from the utterance. Recall that a scene representation contains symbols that are irrelevant to the meaning of the words in the corresponding utterance. The irrelevant meaning symbols (which do not have a counterpart in the utterance) may thus be aligned with the dummy word. Since the dummy word is added to every utterance–scene pair, over time its meaning probabilities reflect the relative frequency of the meaning elements encountered in the input. Due to this accumulated probabilistic knowledge, if a previously observed (familiar) meaning element appears in an input pair *without* its associated word, the meaning element is likely to be aligned with the dummy word rather than a new word in the input. By contrast, a novel meaning is more likely to be aligned with a new word in the utterance, since it has not been linked to the dummy word earlier. We investigate one of the interesting effects of this informed smoothing on the acquisition of second labels (synonyms) in Section 8.

Step 2: Updating the word meanings.

On the basis of the evidence from the alignment probabilities calculated for the current input pair, we update the probabilities $p(.|w)$ for each word $w \in \mathrm{U}^{(t)}$. We add the current alignment probabilities for w and the symbols $m \in \mathrm{S}^{(t)}$ to the accumulated evidence from prior co-occurrences of w and m. We summarize this cross-situational evidence in the form of an association score, which is updated incrementally:

$$\mathrm{assoc}^{(t)}(w,m) = \mathrm{assoc}^{(t-1)}(w,m) + a(w|m,\mathrm{U}^{(t)},\mathrm{S}^{(t)}) \tag{2}$$

where $\mathrm{assoc}^{(t-1)}(w,m)$ is zero if w and m have not co-occurred prior to receiving the current input pair. The association score of a word and a symbol is basically a weighted sum of their co-occurrence counts: Instead of adding one each time the two have appeared in an utterance–scene pair together, we add a probability that reflects the confidence of the model that their co-occurrence is because m is the meaning of w.

The model then uses these association scores to update the meaning of the words in the current utterance:

$$p^{(t)}(m|w) = \frac{\mathrm{assoc}^{(t)}(m,w) + \lambda}{\sum_{m' \in \mathcal{M}} \mathrm{assoc}^{(t)}(m',w) + \beta \times \lambda} \tag{3}$$

where $\mathcal{M}$ is the set of all symbols encountered prior to or at time t, β is an upperbound on the expected number of symbol types, and λ is a small smoothing factor.[6] Basically, the meaning probability of a symbol m for a word w is proportional to the association score between the two. The denominator is simply a normalization factor to get valid probabilities for $p(.|w)$.

Our model updates the meaning of a word every time the word appears in an utterance. For a learned word w, we expect the probability distribution $p(.|w)$ to be highly skewed towards its correct meaning m_w. (An input-generation lexicon contains the correct meaning for each word, as described in Section 4. Note that the model does not have access to this lexicon for learning; it is used only for input generation and evaluation.) In other words, $p^{(t)}(m_w|w)$—which indicates the strength with which w has been learned at time t—should be reasonably high for a learned word. For ease of reference, we refer to $p^{(t)}(m_w|w)$ as the comprehension score of w at time t:

$$\mathrm{comprehension\_score}^{(t)}(w) = p^{(t)}(m_w|w) \tag{4}$$

and consider a word w to be accurately learned if its comprehension score exceeds a predefined threshold, θ. Also, from this point on, we may simply use $p(m|w)$ (omitting the superscript (t)) to refer to the meaning probability of m for w at the present time of learning.

4. Experimental setup

We perform a variety of experiments (presented in Sections 5–8), in which we train our model on input resembling what children receive, and then compare its word learning behaviors to those observed in children. Specifically, we perform two groups of experiments. In one group, we let the model process a large number of input pairs one by one (incrementally) and examine its lexical acquisition behavior over time—where time is measured as the number of input utterance–scene pairs processed. These experiments simulate word learning by children in a naturalistic setting. In these, we use a subset of the full corpus as training data, containing 20,000 (or fewer) input pairs (see Section 4.1 below for details on the creation of the corpus). As specified in each particular experiment, the training pairs may or may not contain noise and/or referential uncertainty.

A second group of experiments simulate specific word learning tasks performed by children in a laboratory setting. In these experiments, we first train our model on a small random subset of the full corpus (typically containing 1,000 pairs), and then present the model with contrived test pairs, each simulating a particular experimental condition. The initial training data are used to simulate some amount of learning in the model prior to being exposed to the test pairs. In our experiments, we found that the exact number of training pairs was not important. In such cases, we report results of 20 random simulations of the same experiment, either by taking their averages or by showing some representative sample, in order to avoid behavior that is specific to a particular sequence of input pairs.

Next, we elaborate on the properties and sources of the data we use in our experiments (Section 4.1) and discuss the values we choose for the parameters of the learning algorithm (Section 4.2).

4.1. Input data

We train our model on naturalistic utterances paired with automatically generated scene representations corresponding to the utterances. The utterances are taken from the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 2001) in the CHILDES database (MacWhinney, 2000). The Manchester corpus contains transcripts of caretakers' conversations with 12 children between the ages of 1;8 and 3;0 (years;months). The original corpus contains a number of recording sessions for each child. In order to maintain the chronological order of the data (with respect to the children's age), we concatenate the first sessions from all children, then the next sessions, and so forth. We then preprocess the transcripts by removing punctuation and lemmatizing nouns and verbs.

There is no semantic representation of the corresponding scenes available from CHILDES. Therefore, we automatically construct a scene representation for each utterance, as a set containing the correct meanings of the words in that utterance. We get these from an input-generation lexicon that contains a symbol associated with each word as its meaning. An excerpt from the input-generation lexicon is shown in Fig. 5(a); sample utterances from CHILDES and their scene representations are given in Fig. 5(b). Note that, in the input-generation lexicon, each word has one meaning. That is, in most of our experiments, we assume

(a) An excerpt from the input-generation lexicon:

Word	**Meaning symbol**
but	`but`
very	`very`
boring	`boring`
now	`now`
mommy	`mommy`
...	...

(b) Sample utterances from the Manchester corpus, along with their scene representations (without noise and uncertainty):

Utterance: *but it is very boring*
Scene: {`but`, `it`, `is`, `very`, `boring`}

Utterance: *are we going to play now*
Scene: {`are`, `we`, `going`, `to`, `play`, `now`}

Utterance: *did you get fed up of mommy chitchatting*
Scene: {`did`, `you`, `get`, `fed`, `up`, `of`, `mommy`, `chitchatting`}

Utterance: *was I a bit boring*
Scene: {`was`, `i`, `a`, `bit`, `boring`}

Utterance: *what do you want to play*
Scene: {`what`, `do`, `you`, `want`, `to`, `play`}

Utterance: *let I play a game*
Scene: {`let`, `i`, `play`, `a`, `game`}

... ...

(c) A noisy input pair:

Utterance: *did you get fed up of mommy chitchatting*
Scene: {`did`, `you`, `get`, `fed`, `up`, `of`, `mommy`}

(d) Sample utterance–scene pairs with referential uncertainty; the second pair also contains noise:

Utterance: *but it is very boring*
Scene: {`but`, `it`, `is`, `very`, `boring`, `are`, `we`, `going`, `to`, `play`, `now`}

Utterance: *did you get fed up of mommy chitchatting*
Scene: {`did`, `you`, `get`, `fed`, `up`, `of`, `mommy`, `was`, `i`, `a`, `bit`, `boring`}

Utterance: *what do you want to play*
Scene: {`what`, `do`, `you`, `want`, `to`, `play`, `let`, `i`, `a`, `game`}

... ...

Fig. 5. Excerpts from the input-generation lexicon, and sample utterance–scene pairs with or without noise and referential uncertainty.

the input does not contain any homonymous words. We return to the acquisition of homonyms in the experiments presented in Section 8.

Recall that we do not assume that children always form complete semantic representations of the scene they perceive. To simulate such noise in our input, we pair a proportion of the utterances with noisy scene representations, where we do not include the meaning of one word (at random) from the utterance. A sample noisy input pair can be found in Fig. 5(c), in which the scene representation is missing the meaning of *chitchatting*. The experiments reported in this article are performed on a corpus with 20% noisy pairs, unless stated otherwise. (Note that even though only one word in each noisy pair is missing its corresponding meaning symbol, this affects the updated meaning probabilities for all words in the utterance, due to the impact of a missing meaning symbol on the alignment probabilities for all the meaning symbols in the scene.)

To simulate referential uncertainty, we use every other sentence from the original corpus, preserving their chronological order. We then pair each sentence with its own scene representation as well as that of the following sentence in the original corpus. Note that the latter sentence is not used as an utterance in our input; see Fig. 5(d) for a sample set of utterances with referential uncertainty generated from the six utterance–scene pairs from 5(b). Our assumption here is that consecutive child-directed utterances taken from a recorded session of parent–child conversations are likely to be talking about different aspects of the same scene. Thus, the extra semantic symbols that are added to each utterance correspond to meaningful and possibly relevant semantic representations, as opposed to randomly selected symbols (as in, e.g., Siskind, 1996). In the full resulting corpus containing 173,939 input pairs, each utterance is, on average, paired with 78% extra meaning symbols, reflecting a high degree of referential uncertainty.

4.2. Parameters

We set the parameters of our learning algorithm using a development data set, a portion of the full corpus set aside for development purposes only and not used as part of the training or test data in our experiments. The upperbound on the expected number of symbols, β in Eq. 3, is set to 8,500 based on the total number of distinct symbols extracted for the development data. Therefore, the initial probability of a symbol for a novel word before any input is processed containing that word (referred to as the default probability) will be $1/8{,}500 \approx 10^{-4}$. Of course, children do not know the precise number of meaning symbols they will be exposed to. Thus, we set this parameter to a large value, to reflect an upperbound on the expected number of possible meaning symbols that a child (the model) may be exposed to. As noted, λ in Eq. 3 is a smoothing parameter (to avoid zero counts). Intuitively, a smoothed (and normalized) assoc score for a meaning symbol previously unseen with a familiar word should be less than the default probability of $1/\beta$, since the unseen symbol has not occurred with the familiar word in the previous usages. We thus set λ to be 10^{-5} (i.e., a value less than $1/\beta$, which is $\approx 10^{-4}$). Generally, λ should be very small; in our early experiments we found that the model works better when λ is less than $1/\beta$.[7]

In our experiments, we often need to decide whether the meaning of a particular word is learned by our model. Recall from Section 3.3 that we assume a word is learned if its comprehension score exceeds a threshold θ. In the experiments reported here, we set θ to 0.7 (unless stated otherwise), a value determined by examining the performance of the model over the development data. Given the large number of meaning elements in total, a substantial portion of the probability mass for a word is assigned to irrelevant meaning elements (those other than the correct meaning of the word), even if each individual probability is very small. Thus, we consider 0.7 a reasonably large portion of the probability mass to assign to the single correct meaning element.

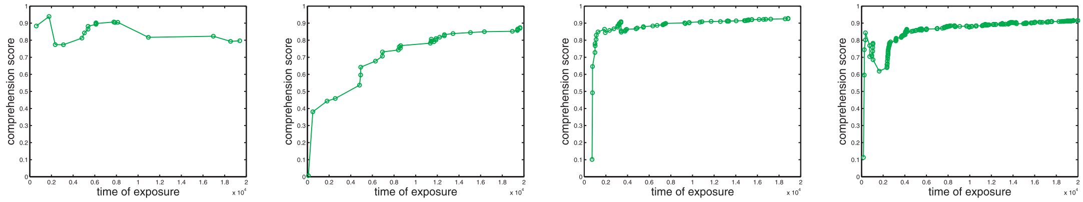
5. Overall learning patterns

This section examines the overall learning behavior of our model. First, we investigate the ability of the model in learning mappings between words and their meanings (Section 5.1) and how this ability is affected by noise and referential uncertainty in the input (Section 5.2). Next, we look into the role of frequency in the acquisition of word meanings in the model (Section 5.3).

5.1. Convergence and learning stability

Our learning algorithm revises the meaning of a word every time it is heard in an utterance; thus, the model can handle noise by revising an incorrectly learned meaning. It is, however, important to ensure that the learning is stable despite this constant revision—that is, the meaning of earlier-learned words is not corrupted as a result of learning new words (the problem of catastrophic interference often observed in connectionist models). If learning is stable, we expect the comprehension scores for words generally to increase over time as more and more examples of the word usages are encountered in the input. To verify this, we train our model on 20,000 input pairs with noise and referential uncertainty (as explained in Section 4) and look at the patterns of change in the comprehension scores of words over time.

Fig. 6 shows the change in the comprehension scores of four sample words over time. The words are chosen from different frequency ranges, from *kiss* having a low frequency of 18 (in 20,000 utterances), to *car* having a high frequency of 236. For all four words, the comprehension scores show some fluctuation at the beginning, but they converge on a high value as more examples of the word are observed. Fig. 7 depicts the change in the average comprehension score of all words, as well as of those which have been learned at some point (i.e., their comprehension score has surpassed the threshold θ). The average comprehension score of all words increases rapidly and becomes stable at around 0.7 after processing around 6,000 input pairs, reflecting the stability in learning. Not surprisingly, the average



kiss (f=18) *fish* (f=33) *book* (f=63) *car* (f=236)

Fig. 6. Change in the comprehension scores of four sample words as more usages of the words are processed.

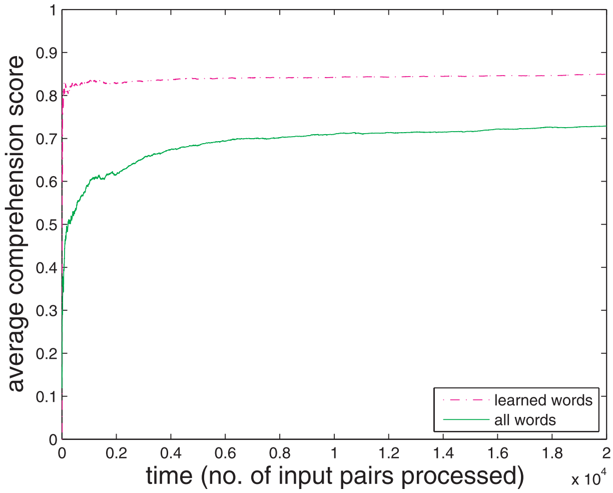


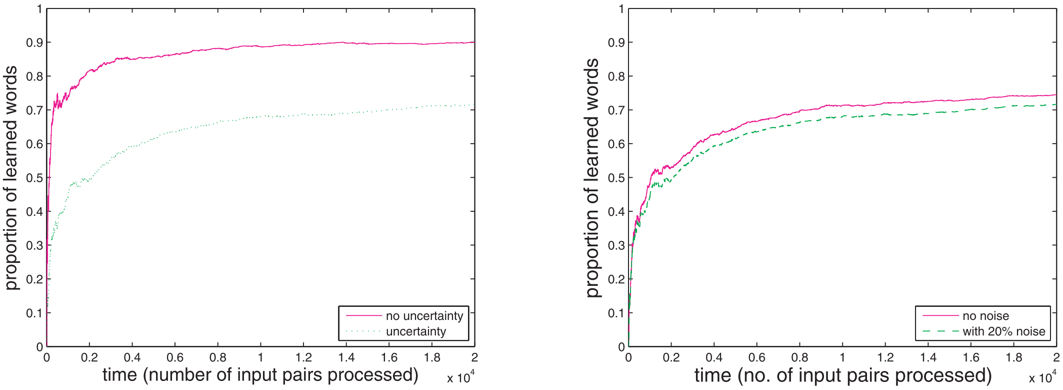
Fig. 7. Change over time in the average comprehension scores of all words, as well as for learned words (i.e., words whose comprehension score exceeds a predefined threshold). Time is measured as the number of input utterance–scene pairs processed.

comprehension score of the learned words increases more quickly (almost instantaneously) and reaches a higher value (around 0.85). This difference is expected since the learned words all have comprehension scores exceeding 0.7.

The stability in the comprehension scores reveals that, in general, after the model has observed a word in a variety of contexts and has converged on some meaning for it, it becomes less and less likely that the word has a completely different meaning. Nonetheless, our model does not fix the meaning of a word—even after a strong association between the word and a meaning element is acquired—giving the model the ability to revise an incorrect meaning learned due to noisy input, as well as the ability to learn the secondary meaning of a homonymous word (see Section 8 for more details on the latter).

5.2. Effects of noise and referential uncertainty

Here, we look into how the learning process in our model is affected by the noise and uncertainty in the input. First, we examine the effect of referential uncertainty: We train our model on 20,000 input pairs, both with and without uncertainty, and look at the difference in the rate of word learning over time in the two conditions. (In both conditions, the input contains 20% noisy pairs since our analysis presented later shows the effect of noise to be constant.) Fig. 8(a) depicts the learning rates, measured as the proportion of learned words over time. The bottom curve shows the learning pattern for input with referential uncertainty, and the top one shows the results for data without uncertainty. In both cases, the proportion of learned words increases over time, with a rapid pace at early stages of learning, and a more gradual pace later. The plots show that the task of word learning is much easier in the absence of referential uncertainty, reflected in the sharp vocabulary growth, as well as in the high proportion of learned words in this condition (90% compared to 70%).[8]



(a) Effect of referential uncertainty on learning

(b) Effect of noise on learning

Fig. 8. Difference in the learning rates: (a) for input with and without referential uncertainty; (b) for input with and without noise.

Next, let's examine the effect of noise on learning. Fig. 8(b) depicts the learning rates on input (with referential uncertainty), with and without noise. The curves show that noise has a constant (though minimal) effect on the learning rates: Even in the presence of a substantial rate of noise in the input (20% of the pairs are noisy), the model learns the meaning of most words. Moreover, the difference in the learning rates in the absence and presence of noise is not substantial, reinforcing the robustness of the probabilistic model.

We observe that the adverse effect of referential uncertainty on word learning is much more pronounced than that of noise. This difference can be attributed to a corresponding difference in the proportions of uncertainty and noise in our data. On average, each utterance is paired with 78% irrelevant meaning symbols, whereas only 20% of our input pairs are noisy, and even these are missing only one meaning symbol. Although these precise proportions are arbitrary, we believe the difference in them is justified since it is much more likely that the learner/child perceives aspects of a scene that are irrelevant to the corresponding utterance, as opposed to not being able to observe or conceptualize the meaning of a word from the utterance.

Overall, our model is robust to noise and referential uncertainty in the input, but learning gets slower with data that contain these. The observed patterns suggest that cleaner data make word learning easier. These results are consistent with the findings of Brent and Siskind (2001) that children's access to words in isolation (used with their referents specified clearly and unambiguously) helps them acquire the words faster. Psycholinguistic studies have shown that the socioeconomic and literacy status of mothers affects the quantity and the properties of the mothers' speech directed to their children (Ninio, 1980; Pan et al., 2005; Schachter, 1979), and this in turn affects the pattern of vocabulary production in the children. For example, Pan et al.'s experiments show that nonverbal input (e.g., pointing) has a positive effect on children's vocabulary growth, reinforcing that cleaner data (with less referential uncertainty) accelerates vocabulary acquisition. Nonetheless, our model is capable of learning the meanings of words, even in the presence of a substantial degree of noise and referential uncertainty, which is congruent with the fact that all (normal) children eventually learn the vocabulary of their language. (Note that Pan et al., 2005 also find that some differences in maternal input mainly affect vocabulary growth at earlier stages of learning.)

5.3. Effect of frequency in word learning

Here, we examine the role of frequency in word learning by looking into the relation between a word's frequency and how easily the model learns it. Specifically, we train our model on input that contains noise and referential uncertainty, and we examine the difference in the learning rates for words from different frequency ranges. Fig. 9 displays four learning curves: one for all words in the input, and three others, each for words which have appeared in the input at least twice, three times, or five times, respectively. (Note that low-frequency words are only removed from the evaluations, and not from the input data.) A comparison of the curves shows that the more frequent a word is, the more likely it is to be learned. In particular, when only considering the learning rate of words with a minimum frequency of five, learning is as easy as when there is no referential uncertainty in the input (cf. the top curves in Figs. 8(a) and 9). These observations conform with the findings of Huttenlocher et al. (1991) who show that there is a high correlation between the frequency of usage of a word in mothers' speech and the age of acquisition of the word. Results of experiments by Schachter (1979), Naigles and Hoff-Ginsberg (1998), and Hoff and Naigles (2002) also suggest that the frequency of words has a positive effect on their acquisition.[9]

6. Vocabulary growth

Examining the patterns of children's vocabulary growth over the course of lexical development has provided researchers with insight on the mechanisms that might be at work for

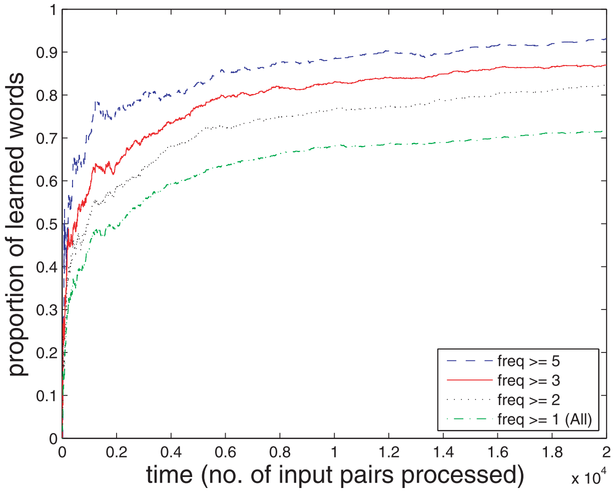


Fig. 9. Effect of frequency: difference in the rate of learning words from different frequency ranges.

word learning, as well as on whether and how these mechanisms change over time. We thus look at the change in the pattern and rate of word learning over time in our model (Section 6.1) and accordingly suggest some possible sources for the patterns we observe (Section 6.2).

6.1. The developmental pattern of word learning

Longitudinal studies of early vocabulary growth in children have sometimes shown that vocabulary learning is slow at the very early stages of learning, then proceeds to a rapid pace, and finally becomes less active (e.g., Gopnik & Meltzoff, 1987; Kamhi, 1986; Reznick & Goldfield, 1992). The middle stage of such a progression is often referred to as the *vocabulary spurt*. The vocabulary spurt has been suggested to arise from qualitative changes in the nature of lexical acquisition over time, for example, a shift from an associationist to a referential word learning mechanism (Nazzi & Bertoncini, 2003), a sudden realization that objects have names, or the naming insight (Kamhi, 1986; Reznick & Goldfield, 1992), the development of categorization abilities (Gopnik & Meltzoff, 1987), or the onset of word learning constraints (Behrend, 1990). The common belief among the proponents of this view is that children's early words (those learned prior to the spurt) are learned through a slow associative process, whereas for learning later words children need to make use of biases and/or constraints such as those mentioned above.

Psycholinguistic experiments examining patterns of vocabulary growth have often shown substantial individual differences among children, both with respect to whether they show a vocabulary spurt, and with regard to the age at which the spurt is observed, if at all (Ganger & Brent, 2004; Huttenlocher et al., 1991; Pan et al., 2005; Reznick & Goldfield, 1992). Moreover, there is no agreed-upon method for identifying a true spurt in the course of lexical development of a child. Thus, what might be viewed as a spurt by one researcher may be considered as a gradual increase by another. For these reasons, another group of researchers have argued against the existence of a sudden spurt; instead, they suggested that the rate of word learning increases in a more linear and gradual fashion (e.g., Bates & Carnevale, 1993; Bloom, 2000; Ganger & Brent, 2004). Proponents of this view believe that the vocabulary growth rate is faster at early stages of word learning largely due to the properties of the input children receive from their environment (McMurray, 2007). Huttenlocher et al. (1991), for example, suggest that the acceleration in word learning during early stages might be in part due to an *indirect* effect of exposure, as reflected in the current levels of lexical knowledge in the learner.

We have already seen that our model, like those of Siskind (1996) and Yu (2008), learns more quickly as the model is exposed to more input; indeed, its learning rate increases very quickly at earlier stages and then gradually stabilizes. Here we further explore this aspect of the model's behavior by examining the data separated by individual children, and by looking at behavior in terms of the number of word types heard thus far (rather than in terms of the number of input pairs, as in our analysis in the previous section). This enables us to consider patterns across individual children in terms of finer grained properties of the input they have been exposed to (rather than just shear amount).

In this more detailed analysis, we examine the pattern of vocabulary growth (i.e., the rate of word learning) to see whether we observe a sudden or a gradual increase in the learning rate. Whatever pattern we observe in the behavior of our model emerges in the absence of any particular developmental change or shift in the underlying learning mechanism, since our model incorporates a single mechanism of vocabulary acquisition at all stages of learning. Such an analysis can help us better understand possible causes of a (sudden or gradual) increase in the rate of learning words in the course of lexical acquisition, and the extent to which the changes in the vocabulary growth correlate with the input. To examine the pattern of vocabulary growth in individual children, here we train our model separately on data from each of the 12 children in our corpus (instead of the usual training on a subset of the corpus containing 20,000 input pairs).

Fig. 10 depicts the change in the proportion of learned words as a function of the number of word types received at each point in time. The figure plots the vocabulary growth curve for each child as the model processes the corresponding training pairs for that child. The number of training pairs for the different children varies from around 10,000 to just above 18,000, and the total number of word types in the input ranges from 1,387 to 2,556. The general pattern of growth is similar for all children: Growth rate is higher at the early stages but gradually decreases as more input is processed. The observed pattern can be attributed to the property of our model that uses its own learned knowledge of word meanings to facilitate the learning of new words. Learning is slow at the beginning because the model has no knowledge of word meanings. As the model learns some words, it can bootstrap on this knowledge to acquire new words. This observation is in line with those studies suggesting that the more words the word learner (a child or a computational model) acquires, the easier it gets for it to learn the meaning of novel words (Huttenlocher et al., 1991; Yu, 2008). However, the learning rate eventually slows over time; this is expected because, with

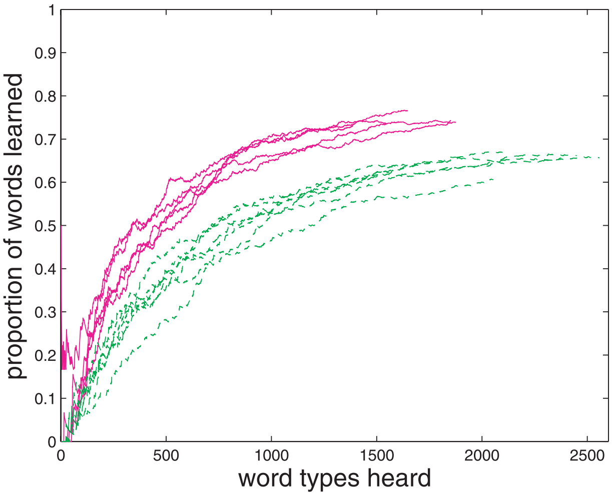


Fig. 10. The patterns of change in the rate of word learning as new words are received.

realistic data such as what we use, there will be a decrease over time in the rate of hearing words that are new to the model.

Similar to the results of experiments on children, here we observe individual differences with respect to the rate of increase in the vocabulary size. Indeed, we can identify two groups of children: one with a sharp vocabulary growth at early stages (first group, shown as solid curves), and the other one with a less steep increase in growth rate (second group, depicted as dashed curves). The learning curves of the children in the first group are all higher than those of the second group, suggesting a faster vocabulary growth in the former group. A closer look at the different training pairs for the two groups of children reveals that the second group—for whom learning appears to be harder—receives utterances that are on average longer than those received by the first group. This observation is based on the values of the mean length utterance (MLU) calculated over the first 100 utterances: the average MLU of data for children in the first group is 3.66, whereas that of the second group is 4.32.[10] In a related study, Brent and Siskind (2001) show the accelerating effect of isolated words—utterances of length one—on early word learning. Our findings, however, are more general and predict that children receiving longer utterances (involving higher degrees of alignment ambiguity) may have a harder time learning the meanings of words. In contrast to this prediction, results of experiments by Bornstein, Haynes, and Painter (1998) and Hoff and Naigles (2002) suggest that children of mothers with higher MLU show better vocabulary competence. In both studies, however, MLU is found to be positively correlated with the number of word types in the input, which in turn positively correlates with children's vocabulary growth. (In a preliminary investigation of our input data which are created based on the Manchester corpus, we also found that the number of word types in the child-directed utterances were positively correlated with the number of types produced by the children. More research into this matter requires a careful examination of the speech produced by children, which is outside the scope of this study.) Thus, it is not clear whether the observed effect in the studies of Bornstein et al. and Hoff and Naigles is directly due to higher MLU (interpreted as syntactic complexity of the input utterances) or indirectly due to a larger number of word types in the child-directed speech. More psycholinguistic studies are needed to further investigate the direct effect of MLU on word learning in young children.

The observed individual differences in children with respect to the rate of word learning and/or vocabulary size are sometimes associated with variation in children's language learning abilities (see Huttenlocher et al., 1991 and the references therein). For example, some children may be more conservative than others in using a learned word in their produced speech. Similarly, the behaviors we observe in our model are all dependent on the value we choose for the parameter θ, which is the confidence of the model in whether a word is learned. (Recall that we consider a word learned when its comprehension score exceeds the threshold θ.) If we assume that children also use a probabilistic representation of their knowledge of word meanings, it is possible that, as in our model, children also need to reach a certain level of confidence about a word's meaning before they can accurately comprehend or produce it. Considering θ as a confidence factor that enables the model (learner) to comprehend or use a word, we examine the variation in the pattern of vocabulary growth in our model by varying the value of this parameter. We train the model on the same input containing 20,000 pairs and plot the vocabulary growth over time for five different values of θ, ranging from 0.5 to 0.9 by steps of 0.1 (see Fig. 11).

The plots show that a learner who can comprehend or use a word only if it is associated with a meaning with a very high confidence (bottom curve, with θ=0.9) has a much slower and a more gradual vocabulary growth. In contrast, for a learner who uses a word even if it has been learned with a low confidence (top curve, with θ=0.5), we observe a very sharp increase in the rate of vocabulary growth at a very early stage in learning. The above results suggest that, in addition to the variation in the input, other factors relating to the learning abilities of children might influence the rate of vocabulary growth, especially in earlier stages of word learning.

6.2. Context familiarity and word learning

The observed shift from slow to fast word learning suggests that children become more efficient word learners later in time (e.g., Woodward, Markman, & Fitzsimmons, 1994). Whereas some researchers attribute this to a change in the nature of learning, others assume this is a natural consequence of being exposed to more input (as noted above). The latter view states that once children have learned a repository of words, they can easily link novel words to their meanings based only on a few exposures. We examine this effect in our model by looking at how its ability to learn novel words changes over time. That is, we look at the relation between the time of first exposure to a word (its ‘‘age of exposure’’ in terms of number of input pairs processed thus far), and the number of usages that the model needs for learning that word (similar effects have been observed in the computational models of Horst et al., 2006; Regier, 2005; Siskind, 1996). Fig. 12 plots this relation for words that have been learned at some point in time. We can see that, generally, words received later in

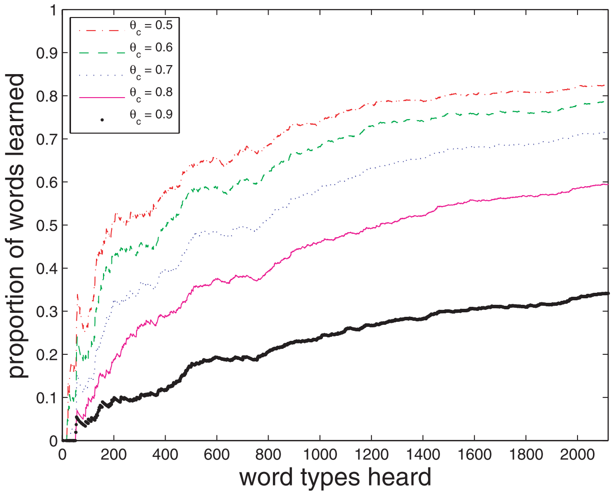


Fig. 11. Variation in the pattern of vocabulary growth as a function of θ.

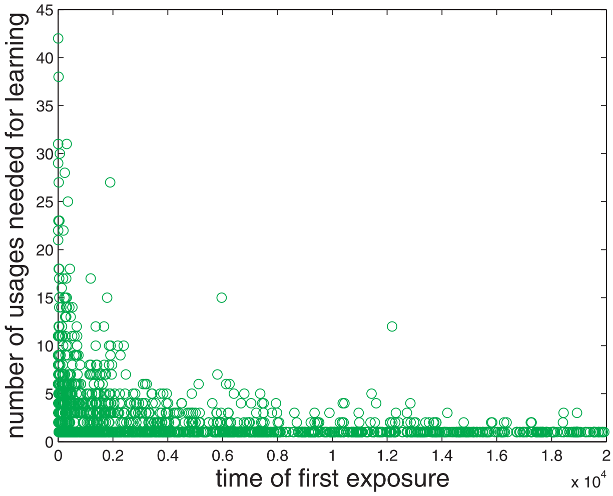


Fig. 12. Number of usages needed to learn a word, as a function of the word's age of exposure.

time require fewer usages to be learned. Similar to the vocabulary growth pattern discussed above, the change in the ability to learn novel words in our model can also be attributed to the bootstrapping mechanism.

The effect of exposure to more input on the acquisition of novel words can be described in terms of context familiarity: The more input the model has processed so far, the more likely it is that a novel word's context (the other words in the sentence and the objects in the scene) is familiar to the model. Note that having more familiar words in an input pair in turn results in a decrease in the degree of alignment ambiguity of the pair. This hypothesis is congruent with the results of a study done by Gershkoff-Stowe and Hahn (2007), who showed that extended familiarization with a novel set of words (used as context) led a group of 16- to 18-month-old children to more rapidly acquire a second set of (target) novel words. (See Alishahi, Fazly, & Stevenson, 2008 for a computational simulation of Gershkoff-Stowe & Hahn's experiment using the same word learning model.)

7. Fast mapping

One interesting ability of children as young as 2 years of age is that of correctly and immediately mapping a novel word to a novel object in the presence of other familiar objects—a phenomenon referred to as *fast mapping* (Carey & Bartlett, 1978). Children's success at selecting the referent of a novel word in such a situation has raised the question of whether and to what extent they actually learn and retain the meaning of a fast-mapped word from a few such exposures. Experiments performed on children have consistently shown that they are generally good at referent selection for a novel word. But the evidence for retention is rather inconsistent; for example, whereas the children in the experiments of Golinkoff et al. (1992) and Halberda (2006) showed signs of nearly perfect retention of the fast-mapped words, those in the studies reported by Horst and Samuelson (2008) did not (all participating children were close in age range). In experiments on children, retention is tested either by making children generalize a fast-mapped novel word to other similar exemplars of the referent object (comprehension) or by having them produce the novel word in response to the referent (production).

The relation between fast mapping and word learning has thus been a matter of debate. Some researchers consider fast mapping as a sign of a specialized (learned or innate) mechanism for word learning. Markman and Wachtel (1988), for example, argue that children fast map because they expect each object to have only one name (mutual exclusivity). Golinkoff et al. (1992) attribute fast mapping to a bias towards mapping novel names to nameless object categories. Some even suggest a change in children's learning mechanisms at around the time they start to show evidence of fast mapping (which coincides with the vocabulary spurt), for example, from associative to referential word learning (Gopnik & Meltzoff, 1987; Reznick & Goldfield, 1992). In contrast, others see fast mapping as a phenomenon that arises from more general processes of learning and/or communication, which also underlie the impressive rate of lexical acquisition in children (e.g., Clark, 1990; Diesendruck & Markson, 2001; Halberda, 2006; Horst et al., 2006; Markson & Bloom, 1997; Regier, 2005).

We investigate fast mapping and its relation to word learning in the context of our computational model. We take a close look at the onset of fast mapping in our word learning model by simulating some of the psychological experiments mentioned above. Specifically, we examine the behavior of our model in various tasks of referent selection (Section 7.1) and retention (Section 7.2), and provide explanations for the (occasionally contradictory) experimental results reported in the literature.

To preview the discussion below, we suggest that fast mapping can be explained as an induction process over the acquired associations between words and meanings. Our model learns these associations in the form of probabilities within a unified framework; however, we argue that different interpretations of such probabilities may be involved in choosing the referent of a familiar as opposed to a novel target word (as suggested by Halberda, 2006). Moreover, the overall behavior of our model confirms that the probabilistic bootstrapping approach to word learning naturally leads to the onset of fast mapping in the course of lexical development, without hard-coding any specialized learning mechanism into the model to account for this phenomenon.

7.1. Referent selection

In a typical word learning scenario, a child faces a scene where a number of familiar and unfamiliar objects are present. The child then hears a sentence, which describes (some part of) the scene, and is composed of familiar and novel words (e.g., hearing *Joe is eating a cheem*, where *cheem* is a previously unseen fruit). In such a setting, our model aligns the objects in the scene with the words in the utterance based on its acquired knowledge of word meanings, and then updates the meanings of the words accordingly. The model can align a familiar word with its referent with high confidence since the previously learned meaning probability of the familiar object given the familiar word, or $p(m|w)$, is much higher than the meaning probability of the same object given any other word in the sentence. In a similar fashion, the model can easily align a novel word in the sentence with a novel object in the scene because the meaning probability of the novel object given the novel word ($1/\beta$, according to Eq. 3, Section 4.2) is higher than the meaning probability of that object for any previously heard word in the sentence (since a novel object is unseen for a familiar word, its probability is less than $1/\beta$).

Earlier fast mapping experiments on children assumed that it is such a contrast between the familiar and novel words in the same sentence that helps children select the correct target object in a referent selection task. For example, in Carey and Bartlett's (1978) experiment, to introduce a novel word–meaning association (e.g., *chromium*–`olive`), the authors used both the familiar and the novel words in one sentence (*bring me the chromium tray, not the blue one.*). However, further experiments showed that children can successfully select the correct referent even if such a contrast is not explicitly mentioned in the sentence. Many researchers have performed experiments where young subjects are forced to choose between a novel and a familiar object upon hearing a request, such as *give me the ball* (familiar target) or *give me the dax* (novel target). In all of the reported experimental results, children could readily pick the correct referent for a familiar or a novel target word in such a setting (Golinkoff et al., 1992; Halberda, 2006; Halberda & Goldman, 2008; Horst & Samuelson, 2008).

Halberda's eye-tracking experiments on both adults and preschoolers suggest that the processes involved for referent selection in the familiar target situation (*give me the ball*) may be different from those in the novel target situation (*give me the dax*). In the latter situation, subjects systematically reject the familiar object as the referent of the novel name before mapping the novel object to the novel name. In the familiar target situation, however, there is no need to reject the novel distractor object because the subject already knows the referent of the target. The difference between these two conditions can be explained in terms of two different uses of the probabilistic knowledge in our model. In the familiar target condition, the meaning probabilities are used directly. In the novel target condition, however, the learner has no previously learned associations between the word and its correct meaning (i.e., the meaning probabilities for the novel word are uniform over all meaning symbols). In this case, the learner needs to reject the unlikely referent by performing some reasoning over the probabilities (as further explained below).

In a typical referent selection experiment, the child is asked to *get the ball* while facing a `ball` and a novel object (*dax*). We assume that the child knows the meaning of verbs and determiners such as *get* and *the*, therefore we simplify the familiar target condition in the form of the following utterance (U) and scene (S) pair:

2. U: *ball* (Familiar Target)
   S: { `ball, dax` }

As described before, the model maintains a meaning probability $p(.|w)$ for each word w over time. A familiar word such as *ball* has a meaning probability highly skewed towards its correct meaning. That is, upon hearing *ball*, the model can confidently retrieve its meaning `ball`, which is the one with the highest probability $p(m|ball)$ among all possible meanings m. In such a case, if `ball` is present in the scene, the model can easily pick it as the referent of the familiar target name, without processing the other objects in the scene.

Now consider the condition where a novel name is used in the presence of a familiar and a previously unseen object:

3. U: *dax* (Novel Target)
   S: { `ball, dax` }

Since this is the first time the model has heard the word *dax*, both meanings `ball` and `dax` are equally likely because $p(.|dax)$ is uniform. Therefore the meaning probabilities are uninformative and cannot be solely used for selecting the referent of *dax*. In other words, the model/learner has no previously learned knowledge of the correct meaning of the novel word *dax*, and hence any object is a potential referent for it. In this case, the model has to perform some kind of induction on the potential referents in the scene based on what it has learned about each of them, in order to accept or reject each of the hypotheses of *dax* referring to `ball` and *dax* referring to `dax`. To achieve this, the model needs to consider the likelihood of a particular word w referring to each of the two meanings m; we call this the *referent probability*. This probability is calculated by drawing on the model's previous knowledge about the mapping between m and w (i.e., $p(m|w)$), as well as the mapping between m and other words in the (learned) lexicon. More specifically, the likelihood of using a particular name w to refer to a given object m is calculated as:

$$\begin{aligned} rf(w|m) &= p(w|m) \\ &= \frac{p(m|w)\cdot p(w)}{p(m)} \\ &= \frac{p(m|w)\cdot p(w)}{\sum_{w'\in\mathcal{V}} p(m|w')\cdot p(w')} \end{aligned} \tag{5}$$

where $\mathcal{V}$ is the set of all words that the model has seen so far, and $p(w)$ is simply the relative frequency of w, as in:

$$p(w) = \frac{\text{freq}(w)}{\sum\limits_{w'\in\mathcal{V}} \text{freq}(w')} \tag{6}$$

What the model is now faced with in Example 3 above are the two probabilities $rf(dax|$`ball`$)$ and $rf(dax|$`dax`$)$. Note that these probabilities cannot be directly compared (because they are conditioned on different meanings), but we can use them to accept or reject the use of *dax* to refer to each of the objects. Given the formulation of *rf* in Eq. 5, the denominator has a high value if the object *m* is strongly associated with a word in $\mathscr{V}$ (as is the case for ball), and the denominator has a low value if *m* is a novel object not yet strongly associated with any word (as is the case for dax). In general, this means that *rf*(*w*|*m*) for a novel target word *w* will be low for a familiar object, and high for a novel object. Specifically in this example, the symbol ball has strong associations with another word *ball* and the referent probability of the novel name *dax* for ball is very low, while the symbol dax does not have strong associations with any of the words in the lexicon and its referent probability is very high. The model will thus reject the use of *dax* to refer to ball and accept the use of *dax* to refer to dax.

To simulate the process of referent selection in our model, we first train it on 1,000 input pairs containing noise and referential uncertainty (as described in Section 4), and then present the model with one more input pair representing either the Familiar Target or the Novel Target condition (Examples 2 and 3 above, respectively). Results reported here are averages over 20 such random simulations. In the Familiar Target condition (shown in the top panel of Table 1), the model demonstrates a strong preference towards choosing the familiar object as the referent. In the Novel Target condition (shown in the bottom panel of Table 1), we look at the referent probabilities *rf*(*target*|object) for both objects after processing the input pair as a training pair, simulating the induction process that humans go through to select the referent in such cases. The model will now reject the use of the target (novel) word to refer to the familiar object and accept the novel object as the referent of the target word. Our results confirm that in both conditions, the model consistently selects the correct referent for the target word across all the simulations, by looking at two different interpretations of the same learned associations between words and meaning. It is important to emphasize, however, that the learning mechanism has not changed: The model still acquires a single set of probabilities by processing the input data, but it can interpret these probabilities in different ways according to the task at hand.

The following section examines the relation between fast mapping and word learning through a series of experiments.

Table 1
Referent selection in Familiar and Novel Target conditions

Upon hearing the target word

Condition	*p*(ball|*target*)	*p*(dax|*target*)
Familiar Target	0.830 ±0.099	≪0.0001

After performing induction

Condition	*rf*(*target*|ball)	*rf*(*target*|dax)
Novel Target	0.116 ±0.139	0.992 ±0.002

7.2. Retention

As discussed above, results from human experiments as well as our computational simulations show that the referent of a novel target word can be selected based on the previous knowledge about the present objects and their names. However, the success of a subject in a referent selection task does not necessarily mean that the child/model has *learned* the meaning of the novel word based on that one trial. In order to better understand what and how much children learn about a novel word from a single ambiguous exposure, some studies have performed retention trials after the referent selection experiments. Often, various referent selection trials are performed in one session, where in each trial a novel object–name pair is introduced among familiar objects. Some of the recently introduced objects are then put together in one last trial, and the subjects are asked to choose the correct referent for one of the (recently heard) novel target words. The majority of the reported experiments show that children can successfully perform the retention task (Golinkoff et al., 1992; Halberda, 2006; Halberda & Goldman, 2008).

We simulate a similar retention experiment by first training the model on 1000 input pairs containing noise and referential uncertainty (as explained in Section 4). We then present the model with two experimental training pairs similar to the one used in the Novel Target condition in the previous section, with different familiar and novel objects and words in each input:

4. U: *dax* (Referent Selection Trial 1)
   S: { `ball, dax` }

5. U: *cheem* (Referent Selection Trial 2)
   S: { `pen, cheem` }

These two additional training pairs are followed by a retention test trial, where the two novel objects used in the previous experimental inputs are paired with one of the novel target words:

6. U: *dax* (2-Object Retention Trial)
   S: { `cheem, dax` }

After processing the retention input, we examine the referent probabilities *rf*(*dax*|`cheem`) and *rf*(*dax*|`dax`) to see if the model can choose the correct novel object in response to the target word *dax*.[11] The top panel in Table 2 presents the average results over 20 such simulations: The model will strongly accept the target word as referring to the correct novel object (this is also the case in all individual simulations).

Recall that, unlike for referent selection, experimental results on retention have not been consistent across various studies. Horst and Samuelson (2008) perform experiments with 2-year-old children involving both referent selection and retention, and report that their sub-

Table 2
Retention of a novel target word from a set of novel objects

2-OBJECT RETENTION TRIAL	
rf(*dax*|`dax`)	*rf*(*dax*|`cheem`)
0.995 ±0.001	0.473 ±0.079

3-OBJECT RETENTION TRIAL		
rf(*dax*|`dax`)	*rf*(*dax*|`cheem`)	*rf*(*dax*|`lukk`)
0.994 ±0.001	0.435 ±0.063	0.988 ±0.001

jects perform very poorly at the retention task. One factor that discriminates the experimental setup of Horst and Samuelson from others (e.g., Halberda, 2006) is that, in their retention trials, they put together two recently observed novel objects with a third novel object that has not been seen in any of the prior experimental sessions. The authors do not attribute their contrasting results to the presence of this third object, but this factor can in fact affect the performance considerably, since this presents the child with an additional, truly novel object (the third unseen object) in the presence of a low-frequency word (the word for one of the recently observed novel objects).

To explore this situation in our model, we perform two referent selection trials such as those presented above (in 4 and 5), but present the model with a different retention test trial, as in:

7. U: *dax* (3-OBJECT RETENTION TRIAL)
   S: { `cheem, dax, lukk` }

The third object, `lukk`, is being seen for the first time in the retention trial. We perform the simulation under the new condition, and collect the appropriate referent probabilities after processing the training and test pairs. Results (averages over 20 random simulations) are reported in the bottom panel of Table 2. As can be seen, the model again shows a high *rf* for using the novel target *dax* to refer to the correct novel object `dax`, and a relatively low *rf* for using the novel target *dax* to refer to the other recently seen novel object `cheem`. However, the *rf* probability for *dax* to refer to the unseen object `lukk` is also very high.

Thus, while the model can confidently reject `cheem` as the referent of *dax*, it cannot similarly reject `lukk` as a referent based on the *rf* values. Essentially, the model has no previously acquired knowledge about `lukk`—that is, an association of it with another word—to rule it out as a referent for *dax*. In addition, since the model has seen the target word *dax* together with its correct referent `dax` only once, the model has not yet learned a strong association between the two, and is thus open to the possibility of `lukk` being the referent of *dax*. These results show that introducing a new object for the first time in a retention trial considerably increases the difficulty of the task. This can explain the contradictory results reported by Horst and Samuelson (2008): When both the meaning probabilities and the referent probabilities are not informative in determining reference, other factors (not modeled here) might influence the outcome of the experiment, such as the amount of training received for a novel word–object, or a possible delay between training and test sessions.

We have shown that the probabilistic association between words and meanings can be used to simulate various fast mapping experiments performed on children, such as referent selection and retention. Our experimental results suggest that fast mapping can be explained as an induction process over the acquired word–meaning associations. In that sense, fast mapping is a general cognitive ability, and not a hard-coded, specialized mechanism of word learning. In addition, our results confirm that the onset of fast mapping is a natural consequence of learning more words, which in turn accelerates the learning of new words. This bootstrapping approach results in a rapid pace of vocabulary acquisition in children, without requiring a developmental change in the underlying learning mechanism.

8. Learning synonyms and homonyms

Children's ability at fast mapping is sometimes taken as evidence for a bias towards a one-to-one mapping between words and meanings (e.g., Markman & Wachtel, 1988). Support for the bias often comes from experiments which show that even though children are generally very good at mapping novel words to novel meanings, they exhibit difficulty in learning homonymous and synonymous words—which require the acquisition of one-to-many and many-to-one mappings, respectively (Casenhiser, 2005; Doherty, 2000; Doherty, 2004; Liittschwager & Markman, 1994; Markman, Wasow, & Hansen, 2003; Mazzocco, 1997).[12] The existence of such a bias as a cognitive mechanism for word learning is sometimes also supported by a common belief that language is primarily a means of communication, and hence cases that cause ambiguity should be dispreferred (Casenhiser, 2005). Clearly, however, children are able to eventually learn homonyms and synonyms of their language.

One explanation is that children are equipped with a bias towards a one-to-one mapping (a bias which may be innate or acquired), and that they need to overcome this bias in order to learn synonyms and homonyms, hence exhibiting difficulty at earlier stages of learning (Markman & Wachtel, 1988; Merriman & Bowman, 1989). Another explanation is that children may early on resist learning the secondary meaning of a homonymous word, simply because an existing mapping between the word and its primary meaning is triggered every time the word is heard, resulting in a conflict (Doherty, 2004). (A similar conflict exists between the primary and secondary labels in the case of synonymy.) The learner thus needs more evidence in order to establish a mapping between the homonymous word and its secondary meaning (and similarly for learning synonyms). In contrast, in a situation where a child hears a novel word in the presence of a novel meaning, neither the word nor the meaning triggers any existing word–meaning mappings, and hence the child can easily learn a mapping between the novel word and the novel meaning.

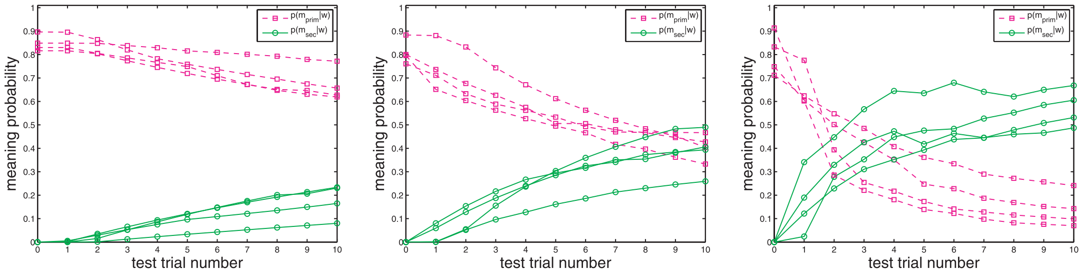
A variety of experiments have been performed on children in order to test their ability in the acquisition of homonyms and synonyms. Children are told stories using picture books, through which they are familiarized with some homonymous or synonymous words. Towards the end of the story, children are then tested for their knowledge of the introduced homonyms or synonyms. In the experiments on learning homonymy, in order to control for a child's familiarity with the primary and secondary meanings of a homonym, and to simulate the child's first encounter with the secondary meaning of the word, pseudo-homonyms are often used in place of real homonyms. A pseudo-homonym is a known word (e.g., *ball*) used to mean something other than its accepted meaning, often a novel meaning not associated with any other word in the child's lexicon (e.g., `dax`). Similarly, pseudo-synonyms are used in experiments on the acquisition of second labels for a familiar object; a pseudo-synonym is a novel word (e.g., *dax*) referring to a familiar object (e.g., `ball`), for which the child already has learned a name (e.g., *ball*). We perform similar experiments to test the ability of our model in learning homonyms and synonyms. We first explain the details of our homonymy learning experiments and then discuss how our model performs in the task of learning synonymous words.

8.1. Learning homonymous words

To simulate homonymy learning in our model, we first train it on 1,000 input pairs (containing noise and referential uncertainty) and then present it with a test trial as follows. We take a typical utterance–scene pair and add to it a pseudo-homonym word–meaning pair—that is, we add a familiar word to the utterance and a novel symbol to the corresponding scene representing the secondary meaning of the pseudo-homonym. The words and meanings in the original utterance–scene pair act as distractors. To understand the effect of exposure on the acquisition of the secondary meaning of a homonymous word w, we present the model with a sequence of 10 test trials and look at the change in the meaning probability of the word's primary and secondary meanings, m_{prim} and m_{sec}, respectively.

Examining a number of these individual simulations reveals an interesting pattern: The acquisition of the secondary meaning of a homonym word in our model is affected by the ''degree of familiarity'' of the pseudo-homonym, as reflected in its frequency of occurrence prior to testing. Instead of averaging over all the simulations, we thus present the results for a random sample of the simulations, each containing a pseudo-homonym from one of three different frequency ranges. In order to examine the patterns in detail, we select four words from each frequency range, showing the results for a total of 12 simulations. Fig. 13 shows the change in the meaning probabilities of the primary and the secondary meanings, $p(m_{prim}|w)$ and $p(m_{sec}|w)$, for sample pseudo-homonyms from the following three frequency ranges: (a) HIGH: the frequency of the pseudo-homonym is higher than twice the number of times the secondary meaning appears with the word, the latter being equal to the number of test pairs (10); (b) MEDIUM: the frequency of co-occurrence of the pseudo-homonym with its primary and secondary meanings is roughly equal; and (c) LOW: the pseudo-homonym has appeared only a few times with its primary meaning.

Fig. 13(a) shows that learning the secondary meaning of a homonym in our model is very difficult if the homonym and its primary meaning are highly familiar. This is consistent with the view that the difficulty children exhibit in the acquisition of a homonym may indicate



(a) HIGH $(freq \geq 20)$ (b) MEDIUM $(5 < freq < 20)$ (c) LOW $(freq \leq 5)$

Fig. 13. Patterns of change in the meaning probabilities of the primary and secondary meanings of homonyms. Results are shown for four representative words randomly selected from each of three frequency ranges, where freq indicates number of appearances of the homonym (with its primary meaning) in 1,000 initial training utterances. The meaning probability of the primary meaning of a homonym, p(mprim|w), is shown as a dashed curve; a corresponding solid curve plots the meaning probability of the word's secondary meaning, p(msec|w). The meaning probabilities are depicted after each test trial, where the trial number 0 shows the values of the probabilities right after training and before testing.

difficulties in suppressing the primary meaning of the word (e.g., Doherty, 2004). Interestingly, for a homonym with a low frequency of occurrence, the model revises its lexical knowledge, essentially treating the few times the word and its primary meaning have appeared together as noise (see Fig. 13(c)). Of course, if the model later receives more evidence to the contrary, it can easily revise this assumption and learn the primary meaning as a true meaning and not one due to noise or chance co-occurrence. Fig. 13(b) shows that for words with medium frequency, the model learns both meanings with somewhat equal probabilities, as both meanings co-occur with the word with comparable frequency. These results can be considered as interesting predictions of our model; it remains to be tested whether the acquisition of homonymous words in children is similarly affected by the degree of familiarity of their primary meanings.

Note that our probabilistic formulation avoids having to stipulate a special homonymy-learning strategy, such as that in Siskind (1996). This is desirable since the use of such a strategy in Siskind's model prevents it from explaining the observation of early difficulty with homonyms observed in children. However, even though our model is capable of adjusting the meaning probabilities of a familiar word to accommodate a new (additional) meaning, the model does not ''learn'' either meaning according to our definition of learning as exceeding a threshold θ on the meaning probability (here set to 0.7). In order to accommodate a new meaning, the model must lessen its probabilitistic commitment to the original meaning; because the probability mass is now divided among multiple correct meaning primitives, none of them passes the threshold value.

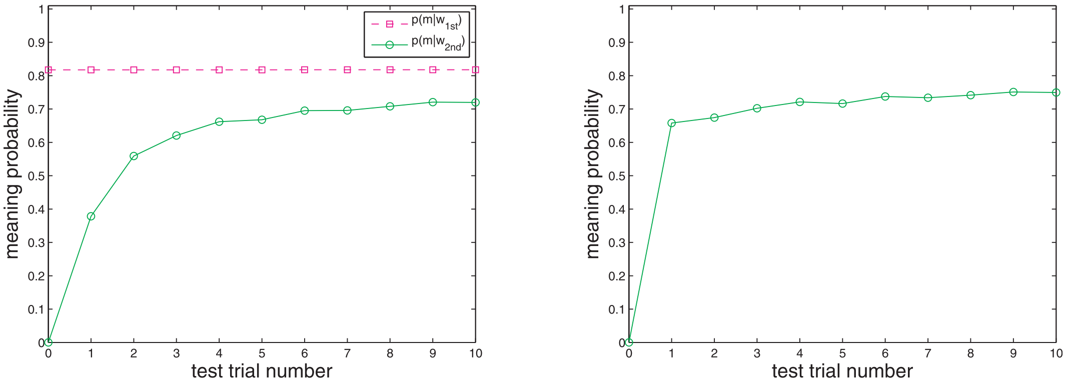
Nonetheless, one of the strengths of probabilistic models is their flexibility. In order to overcome the deficiencies of this current formulation, we can in the future consider alternative mechanisms for determining that a word is learned, without having to change the underlying probability model. For example, we could replace our current threshold-comparing mechanism with one that instead detects significant peaks in the probability distribution.

8.2. Learning synonymous words

Experimental results reveal that children also experience some difficulty in learning a second word for a meaning already associated with a known word; however, given sufficient training, even very young children can acquire the meaning of such new words (Liittschwager & Markman, 1994; Mervis, Golinkoff, & Bertrand, 1994). So while similar to homonyms in some ways, synonyms appear somewhat easier for children. Here we explore how our model behaves in this many (words) to one (meaning) situation, in contrast to the one-to-many situation presented by homonyms.

We test synonymy learning in our model by performing training and test experiments similar to those explained above for homonymy learning. The test trials are different here in that we add a pseudo-synonym word–meaning pair to each test utterance–scene pair. More specifically, we add a familiar meaning to the scene, and a novel word to the utterance representing the second label for the familiar meaning. Here again, the words and meanings in the original utterance–scene pair act as distractors. We perform 20 simulations, each consisting of a training process and a full set of 10 test trials as explained above. Results presented here are averages of the meaning probabilities over the 20 simulations. We have examined the results of the individual simulations, and we have found that they all show very similar patterns.

The solid line in Fig. 14(a) depicts the pattern of change (over time) in the meaning probability of the second label (the pseudo-synonym) for a familiar meaning/object; the dashed line shows the meaning probability of the first label for the pseudo-synonym for comparison. The learning of a second label for a familiar object in children has often been compared against the acquisition of first labels for novel objects. In Fig. 14(b), we thus show the pattern of change in the meaning probability of a novel word (first label) for a novel meaning. These results are averages over 20 random simulations of the task of learning a novel label for a novel meaning (the test trials in these simulations are formed by adding a novel label to the utterance and a novel meaning to the scene).



(a) second label (synonymous words)

(b) first label (novel words)

Fig. 14. Patterns of change in the meaning probabilities when learning the first label or the second label for a given meaning/object.

Unlike in the case of learning a homonymous word, here the acquisition of the second label does not affect the meaning probability of the first label (the dashed line in 14(a)). This is due to the way meanings are represented in our model's lexicon: Each word has an independent probability distribution over the meaning symbols, representing its lexical entry. The results suggest that, generally, our model can successfully and easily learn second labels for familiar objects. Comparing 14(a) and 14(b), however, indicates that the acquisition of first labels is slightly easier for our model: The meaning probability of a novel word increases substantially (close to 0.7) after only one exposure (Fig. 14(b)), whereas more exposure to the second label is needed before its meaning probability reaches a comparable value (Fig. 14(a)). What is responsible in our model for this small difference in the acquisition of first and second labels is the knowledge accumulated in the meaning probabilities associated with the dummy word (as explained in Section 3). A familiar object has a higher probability of association with the dummy word compared to a novel object, since the latter has not been previously aligned with the dummy word. Thus, compared to a novel object, a familiar object will be more strongly aligned with the dummy word and less strongly aligned with a novel word (its second label). In other words, the more a learner has seen an object, the less likely it is that the learner has not heard a label/word for it. The conflict between aligning the familiar object with the target word or the dummy word results in an alignment between the familiar object and a second label that is not as strong as the alignment between a novel object and a first label. That is, the learner needs more evidence for the existence of a novel (second) label for a familiar object, whereas for a novel object the learner expects to hear a novel label. Overall, our results on synonymy learning are generally consistent with the findings of Mervis et al. (1994) and Liittschwager and Markman (1994), whose experiments show that (2- to 3-year-old) children can learn second labels easily if they receive sufficient training (see also Doherty, 2004, for similar results on 6- to 9-year-old children).

9. General discussion

Much research in psychology and linguistics has focused on identifying the mechanisms that are at play in the course of lexical development in young children. A major source of debate among researchers is the relative contribution of specific biases and constraints on the one hand, and the properties of the input on the other. Whereas some researchers strongly believe in the necessity of special biases for word learning (e.g., Behrend, 1990; Golinkoff et al., 1992; Markman & Wachtel, 1998), others argue that word meanings are learned through general cognitive mechanisms (e.g., Bloom, 2000; Tomasello, 2003). The computational experiments presented in this article provide new insights into this matter. Our word learning model does not include any of the suggested biases or constraints, nor does it incorporate any explicit developmental changes in the underlying learning mechanism. Results of our experiments (presented in Sections 5–8) thus suggest that much about word meanings can be learned from naturally occurring utterances. Our model exhibits a range of behaviors similar to those observed in children, with similar developmental patterns that naturally emerge as a result of processing more data.

Our model adopts a flexible view of word meaning: Instead of establishing a rigid mapping between a word and a meaning element, the model learns a word meaning as a probability distribution over all meaning elements encountered in the input. The model is also probabilistic with respect to the learning algorithm it incorporates: It implements a probabilistic and incremental interpretation of the cross-situational learning mechanism. For each input—an utterance paired with a scene representation—the model uses any knowledge acquired thus far about possible meanings of the words in the utterance to form an alignment probability between them and the meaning elements in the scene. It then updates the meaning probabilities to reflect the current strength of alignment between words and meaning elements in the scene. The probabilistic nature of the model, together with the model's incrementality, makes it robust to noise and uncertainty in the input (as shown in Section 5). The model can easily revise an incorrectly learned word–meaning mapping that has resulted from a noisy input, by the natural adjustment of its probabilities in response to more input. Referential uncertainty—when meaning elements irrelevant to the utterance appear in the scene representation—is handled as a result of the interaction between the two types of probabilistic knowledge (alignment and meaning probabilities) acquired over time. We expect that irrelevant meaning elements do not regularly co-occur with any given word (in contrast to relevant elements). Thus, the association between an irrelevant meaning element and a word (reflected in the meaning probability of the word) is expected to be low. Since the alignments for an utterance–scene pair are calculated from the meaning probabilities, those between a word (in the utterance) and an irrelevant meaning element (in the scene) will be weak. This way, the model implicitly marks the irrelevant meaning elements in the scene by not strongly aligning them with any of the words in the utterance.

Because our framework handles noisy data, the behavior of our model can be evaluated using actual child-directed sentences. Our results in Section 5.2 show that using realistic input data with noise and referential uncertainty significantly increases the difficulty of the task of learning the mapping between words and their meanings. This raises the question of whether existing computational models that assume a much simpler format for their input (e.g., a pairing of a phonological form and a symbolic meaning) are scalable to more complex settings and would show the same patterns of behavior given more realistic data. In contrast, our model provides a testbed that can be used to examine various aspects of word learning in a more naturalistic setting and to make predictions about the behavior of young word learners in novel situations.

Our probabilistic incremental learning algorithm also enables the model to explicitly use its partially acquired knowledge of word meanings to accelerate the acquisition of novel words. This is possible through the interaction between the two types of probabilities acquired and updated over time. This bootstrapping mechanism is responsible for some of the observed developmental patterns in the behavior of our model, such as the pattern of vocabulary growth presented in Section 6. Our model exhibits the same developmental pattern observed in children, without having to posit different learning mechanisms over time.

This is in contrast to theories which suggest that children become more efficient word learners later in their life due to a change in the underlying learning mechanisms, for example, from associative to referential (Kamhi, 1986; see also Li, Xiaowei, & MacWhinney, 2007 for a computational model that explicitly incorporates two modes of learning to account for the vocabulary growth in children).

A probabilistic model is also flexible in that it can support different views of the developing representation (the acquired word meanings) without requiring different learning mechanisms or biases. For example, our experiments in Section 7 show that a difference in the behavior of a learner in two conditions of the same task (referent selection with familiar or novel target words) may be attributed to the learner's use of two different interpretations of the probabilistic knowledge of word meanings. There, we argued that when the meaning probabilities are uninformative in guiding referent selection for a new word, the model can perform induction over them to yield a different ''view'' of the learned information, which we call the referent probability. The particular use of the referent probabilities exhibits surface behavior that appears the same as the result of using the mutual exclusivity (ME) bias. That children are good at ruling out a familiar object as the referent of a novel word is sometimes taken as evidence for the use of the ME bias as a word learning mechanism (e.g., Markman & Wachtel, 1988). According to the results of our fast mapping experiments, however, the ME bias is not a necessary mechanism for word learning, but rather may appear as a consequence of using a certain problem-solving method in a particular task, such as referent selection.

It has been suggested that the ME bias—a bias towards a one-to-one mapping between words and meanings—is responsible for the observed difficulty in children's acquisition of homonymous and synonymous words: Children must override the one-to-one mapping bias in order to learn homonyms and synonyms (e.g., Liittschwager & Markman, 1994; see also Yurovsky & Yu, 2008 for related experiments on adults). Our model does not incorporate the ME bias as part of its learning mechanism, but still exhibits similar behaviors to those observed in children with respect to learning synonyms and homonyms. Our results (presented in Section 8) suggest that the initial reluctance of children for learning the secondary meaning of a homonymous word might be due to a conflict with the primary meaning of the word (as also suggested by, e.g., Doherty, 2004), as opposed to a need for overriding the ME bias. In fact, our results show that the more familiar the primary meaning of a homonymous word, the harder it is for the model to acquire the word's secondary meaning. A similar conflict is observed when learning a second label for a familiar meaning (a synonymous word): The model can readily learn a synonym (second label); however, it is still easier for it to learn a novel word (first label).

To summarize, our model successfully accounts for many of the observed patterns in the course of early vocabulary acquisition in children, while learning word–meaning mappings from large-scale, naturalistic data that resemble what children receive as input. Importantly, the model accomplishes this by incorporating only general mechanisms of learning, and without a need for the explicit use of specific word-learning biases, suggesting that the input children receive can to a large extent shape the developmental patterns they exhibit. Nonetheless, the model has a number of limitations that need to be addressed in future work, which we briefly mention here.

Integrating word learning with other aspects of language acquisition. Assuming that word learning is an isolated process in language acquisition, we represent each utterance in the input as an unordered set of words, thus ignoring the syntactic structure of the sentences. The two processes of syntax acquisition and word–meaning mapping are more likely to be intertwined, and in fact there is evidence that children use syntactic knowledge to learn the meaning of words (e.g., Gertner et al., 2006; Gleitman & Gillette, 1994; Hoff & Naigles, 2002; Naigles, 1990). By extending the model to also incorporate syntactic information, we can examine the role of syntax in facilitating the acquisition of word meanings (i.e., syntactic bootstrapping). One possible way of integrating this information into the word learning process is to use the distributional properties of words (i.e., the context words appear in) for categorizing them. The lexical category of a word can provide useful cues about the semantic properties of the word, and this information can be used along with other information sources to boost the learning of word meanings. We can also examine the role of knowledge of word meanings in the acquisition of word (syntactic and/or semantic) categories. Specifically, we can plug our word learning model into a computational model of word category learning. Existing computational models that focus on the acquisition of such categories either assume perfect word–meaning mappings (as in Alishahi & Stevenson, 2008) or ignore such mappings altogether (as in Parisien, Fazly, & Stevenson, 2008). A computational model that combines these different (though related) aspects of language acquisition could shed further light on the interactions among the various components of early language acquisition in children. Some recent work has investigated the joint acquisition of word meaning and word order (Maurits, Perfors, & Navarro, 2009), assuming certain built-in knowledge of grammatical functions and the relational structure of meaning. Given our framework of basic word–meaning mapping, the joint learning of syntactic or semantic categories (as opposed to word order) seems more relevant for our model.

Using richer semantic representations. Although the input sentences to our model are selected from recorded conversations between children and their parents, the semantic representations paired with these sentences are rather naive. One important shortcoming of our approach is the model's inability to represent the relation between super/subordinate terms. Currently, we represent the correct meaning of the related terms *cow* and *animal* as two individual and unrelated symbols `cow` and `animal`. In Fazly et al. (2008) we represented the meaning of each word as a set of semantic features (instead of a single semantic symbol), where the meaning sets of related terms have a significant overlap (or, in the case of superordinate terms, one meaning representation is the subset of the other). We plan to explore this alternative representation in the future and specifically study the acquisition of superordinate terms by our model (see Xu & Tenenbaum, 2007, for one such study).

Even with this type of enhancement, the semantic representation in our model would still be impoverished. Although it is not well understood how humans represent the meaning of a sentence, it is clear that these representations must have a much more complex (and highly relational) structure than a mere collection of symbols (see, e.g., the logical representation used by Siskind, 1996, following Jackendoff, 1990). It must be assumed that the meanings children learn support some conception of relational semantics. This poses two open research questions to address. First, we must consider how our probabilistic formulation can be extended to deal with acquiring relational structures. Second, we must determine how to automatically generate such semantic structures to pair with child-directed speech (CDS) data, since there are no large CDS corpora annotated with such meaning representations.

Incorporating other information sources. Using a more appropriate semantic representation would further enable us to investigate a range of phenomena that are not possible to study in the current setting. For instance, children are shown to be sensitive to social–pragmatic cues in the input and to use them to map words to their meanings, especially at earlier stages of word learning (see, e.g., Hoff & Naigles, 2002, the references therein). By incorporating a more realistic semantic representation in our model, we would be able to pay attention to such cues and to use them in the word learning process (see Frank et al., 2007; Yu & Ballard, 2008, for two such approaches). Another important source of information that a more elaborate semantic representation can provide is the distinction between different categories of words. Currently we do not distinguish between different semantic categories, such as words that refer to perceivable entities versus words that describe a state or action. Nor do we pay attention to the distinction between different syntactic categories, such as verbs and nouns. It has been shown that children tend to learn nouns (more specifically, object names) before verbs, and concrete words before abstract ones (Gentner, 2006). It has also been shown that, when learning the meaning of a new word, humans have a tendency to assign the new word to a ‘‘basic’’ semantic category (i.e., a category at a particular level of the category inclusion hierarchy) (Rosch et al., 1976). A richer representation that reflects the differences between (syntactic or semantic) categories would make it possible to study such learning preferences in our model.

Acknowledgments

This article is an extended and updated version of two papers that appeared, respectively, in the *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (CogSci 2008) and the *Proceedings of the 12th Conference on Natural Language Learning* (CoNLL 2008); we thank the anonymous reviewers of those papers. We also thank Arthur Markman and the anonymous reviewers of this paper for their insightful recommendations that helped improve the quality of the article. We are grateful to our colleagues in the Computational Linguistics group at the University of Toronto. We especially thank Aida Nematzadeh and Chris Parisien for their feedback during the preparation of this article. The research was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Faculty of Arts and Science at the University of Toronto.

Notes

1. A study by Brent and Siskind (2001) shows that isolated words form a considerable portion of infant-directed speech (around 9%). However, receiving such input is not considered to be necessary for word learning since all normal children eventually learn all words in their language, even though the vast majority are not presented in isolation.
2. Specifically, if an inconsistency in the meaning symbol sets is detected when processing a word, the model defines a new ''sense'' for the word and starts to build its meaning from scratch. This mechanism allows the model to handle some degree of noise and to learn multiple meanings for the same word form, at the expense of creating many incorrect word senses that clutter the lexicon and affect the model's efficiency.
3. Experimental data in these studies show that infants as young as 8 months old are sensitive to speech cues and have substantial segmentation capabilities. Several computational models have also demonstrated the usefulness of speech cues for word segmentation (see, e.g., Brent & Cartwright, 1996; Goldwater, Griffiths, & Johnson, 2007).
4. Note that this approach contrasts strongly with that of Siskind (1996), in which words are associated with sets of ''possible'' and ''necessary'' meaning elements. For our model, *all* meaning elements are possible for a word, and none is necessary—instead, the probabilities are refined over time to reflect stronger or weaker associations, enabling the model to not be permanently misled by noisy data.
5. Note that due to the probabilistic nature of the alignments, one meaning symbol in the scene may be *weakly* aligned with many words in the corresponding utterance. This is indeed the case when the model does not have any prior knowledge of the meaning of words in an utterance. Also note that this assumption differs from the principles of contrast (Clark, 1990) and mutual exclusivity (Markman & Wachtel, 1988), both assuming contrast across the entire vocabulary. Instead, we assume a probabilistic contrast among the meanings of the words within an utterance.
6. As is standard in statistical estimates of probabilities, we use a smoothing factor added to all counts of the assoc score to avoid zero counts. To better understand the above denominator in this context, it can be rewritten as $\sum_{m' \in \mathscr{M}} (\text{assoc}^{(t)}(m', w) + \lambda) + (\beta - |\mathscr{M}|) \times \lambda$, where the first term adds the assoc score plus the smoothing factor λ for all observed meaning symbols (of which there are $|\mathscr{M}|$), and the second term adds 0+λ (or just λ) for all unseen meaning symbols (of which there are $\beta - |\mathscr{M}|$) which have an assoc score of 0. Thus, adding $\beta \times \lambda$ in the denominator above ensures that all meaning elements have been taken into account in apportioning the probability mass.
7. We do not expect substantial changes to the qualitative pattern of our results under different settings of these parameters, as long as β is set to a reasonably large value, and λ is very small, reflecting the probability of an unseen word–meaning pair.

8. It is interesting to note that the model of Siskind (1996) is much less sensitive than ours to referential uncertainty. In his approach, referential uncertainty in the input consists of having a large number of complete conceptual representations for each utterance; however, most of these can easily be ruled out due to the strong compositionality constraint built into the model. Thus, the practical consequences of a higher degree of referential uncertainty is less severe in that model compared to ours. Note that ''noise'' in the sense that we refer to it in this paper (i.e., having words in an utterance whose meaning is missing from the corresponding scene) is not simulated in the input to the model of Siskind (1996).
9. A different conclusion was made by Pan et al. (2005) who did not find an independent (significant) effect of mothers' tokens on vocabulary growth in children. Nonetheless, their results revealed a positive effect of the produced types on word learning and a positive correlation between types and tokens.
10. We calculate MLU over the first 100 utterances to better understand its effect on word learning at early stages, since a boost in the number of learned words at early stages of learning results in an overall faster acquisition of words in our model.
11. Note that even though the model/learner has heard the words *dax* and *cheem* once, the words are still not considered as *familiar*, and hence we look at the referent probabilities here.

12 Our focus is on the acquisition of homonyms by young children; hence, we are not concerned with the distinction between homonyms such as *bank* (''financial institution'')/*bank* (''river bank'') and homophones such as *bear*/*bare*.

References

Alishahi, A., Fazly, A., & Stevenson, S. (2008). Fast mapping in word learning: What probabilities tell us. In A. Clark & K. Toutanova (Eds.), *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL)* (pp. 57–64). Manchester, UK: Association for Computational Linguistics .

Alishahi, A., & Stevenson, S. (2008). A computational model for early argument structure acquisition. *Cognitive Science*, *32*(5), 789–834.

Aslin, R., Saffran, J., & Newport, E. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324.

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

Barrett, M. (1994). *Early lexical development*. Malden, MA: Wiley-Blackwell.

Bates, E., & Carnevale, G. F. (1993). New directions in research on language acquisition. *Developmental Review*, *13*, 436–470.

Behrend, D. A. (1990). Constraints and development: A reply to Nelson (1998). *Cognitive Development*, *5*, 313–330.

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

Bornstein, M. H., Haynes, M. O., & Painter, K. M. (1998). Sources of child vocabulary competence: A multivariate model. *Journal of Child Language*, *25*, 367–393.

Bowerman, M., & Choi, S. (2003). Space under construction: Language specific spatial categorization in first language acquisition. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and cognition* (pp. 387–428). Cambridge, MA: The MIT Press.

Brent, M. (1996). Advances in the computational study of language acquisition. *Cognition*, *61*, 1–38.

Brent, M., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125.

Brent, M., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), 1333–1344.

Brown, P. F., Della Pietra S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263–311.

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

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

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*(4).

Casenhiser, D. M. (2005). Children's resistance to homonymy: An experimental study of pseudohomonyms. *Journal of Child Language*, *32*, 319–343.

Choi, S., & McDonough, L. (2007). Adapting spatial concepts for different languages: From preverbal event schemas to semantic categories. In J. M. Plumert & J. P. Spencer (Eds.), *The emerging spatial mind*. New York: Oxford University Press.

Clark, E. V. (1990). On the pragmatics of contrast. *Journal of Child Language*, *17*, 417–431.

Coady, J. A., & Aslin, R. (2004). Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology*, *89*(3), 183–213.

Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. *Developmental Psychology*, *37*(5), 630–641.

Doherty, M. J. (2000). Children's understanding of homonymy: Metalinguistic awareness and false belief. *Journal of Child Language*, *27*, 367–392.

Doherty, M. J. (2004). Children's difficulty in learning homonyms. *Journal of Child Language*, *31*, 204–214.

Fazly, A., Alishahi, A., & Stevenson, S. (2008). A probabilistic incremental model of word learning in the presence of referential uncertainty. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 703–708). Austin, TX: Cognitive Science Society.

Fisher, C. (1996). Structural limits on verb mapping: The role of analogy in children's interpretations of sentences. *Cognitive Psychology*, *31*(1), 41–81.

Forbes, J. N., & Farrar, M. J. (1995). Learning to represent word meaning: What initial training events reveal about children's developing action verb concepts. *Cognitive Development*, *10*(1), 1–20.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2007). A Bayesian framework for cross-situational word-learning. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems, volume 20* (pp. 1212–1222). Cambridge, MA: MIT Press.

Ganger, J., & Brent, M. R. (2004). Reexamining the vocabulary spurt. *Developmental Psychology*, *40*(4), 621–632.

Gentner, D. (1978). On relational meaning: The acquisition of verb meaning. *Child Development*, *49*(4), 988–998.

Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek & R. Golinkoff, (Eds.), *Action meets word: How children learn verbs* (pp. 544–564). New York: Oxford University Press.

Gershkoff-Stowe, L., & Hahn, E. R. (2007). Fast mapping skills in the developing lexicon. *Journal of Speech, Language, and Hearing Research*, *50*, 682–697.

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science*, *17*(8), 684–691.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 135–176.

Gleitman, L., & Gillette, J. (1994). *The role of syntax in verb learning. The handbook of child language*. Oxford, England: Blackwell.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2007). Distributional cues to word segmentation: Context is important. In H. Caunt-Nulton, S. Kalatilake, & I. Woo (Eds.), *Proceedings of the 31st Boston University Conference on Language Development* (pp. 239–250). Somerville, MA: Cascadilla press.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wegner, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology*, *28*(1), 99–108.

Golinkoff, R. M., Hirsh-Pasek, K., Mervis, C. B., Frawley, W. B., & Parillo, M. (1995). Lexical principles can be extended to the acquisition of verbs. In M. Tomasello & W. E. Merriman (Eds.), *Beyond names for things: Young children's acquisition of verbs* (pp. 185–216). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gopnik, A., & Meltzoff, A. (1987). The development of categorization in the second year and its relation to other cognitive and linguistic developments. *Child Development*, *58*(6), 1523–1531.

Halberda, J. (2006). Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cognitive Psychology*, *53*, 310–344.

Halberda, J., & Goldman, J. (2008). One-trial learning in 2-year-olds: Children learn new nouns in 3 seconds flat. (in submission).

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child Development*, *73*(2), 418–433.

Horst, J. S., McMurray, B., & Samuelson, L. K. (2006). Online processing is essential for learning: Understanding fast mapping and word learning in a dynamic connectionist architecture. In R. Sun (Ed.), *Proceedings of CogSci'06* (pp. 339–344). Austin, TX: Cognitive Science Society.

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128–157.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*(2), 236–248.

Jackendoff, R. (1990). *Semantic structures*. Cambridge, MA: MIT Press.

Johnson, E., & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*, 548–567.

Jusczyk, P., & Aslin, R. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *28*, 1–23.

Kalagher, H., & Yu, C. (2006). The effects of deictic pointing in word learning. In, *Proceedings of the 5th International Conference of Development and Learning*. Bloomington, IN.

Kamhi, A. G. (1986). The elusive first word: The importance of the naming insight for the development of referential speech. *Journal of Child Language*, *13*, 155–161.

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

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

Liittschwager, J. C., & Markman, E. M. (1994). Sixteen- and 24-month-olds' use of mutual exclusivity as a default assumption in second-label learning. *Developmental Psychology*, *30*(6), 955–968.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk, volume 2: The database (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mandler, J. M. (1992). How to build a baby: II. conceptual primitives. *Psychological Review*, *99*, 587–604.

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

Mattys, S., Jusczyk, P., Luce, P., & Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*, 465–494.

Maurits, L., Perfors, A. F., & Navarro, D. J. (2009). Joint acquisition of word order and word reference. In N. Taatgen & H. Van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1728–1733), Austin, Tx: Cognitive Science Society.

Mazzocco, M. M. M. (1997). Children's interpretations of homonyms: A developmental study. *Journal of Child Language*, *24*, 441–467.

McMurray, B. (2007). Difusing the childhood vocabulary explosion. *Science (Brevia)*, *317*, 631.

Merriman, W. E., & Bowman, L. L. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the Society for Research in Child Development*, *54*, (Serial No. 220).

Mervis, C. B., Golinkoff, R. M., & Bertrand, J. (1994). Two-year-olds readily learn multiple labels for the same basic-level category. *Child Development*, *65*, 1163–1177.

Naigles, L.. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, *17*, 357–374.

Naigles, L.. & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? Effects of input frequency and structure on children's early verb use. *Journal of Child Language*, *25*, 95–120.

Naigles, L., & Kako, E. T. (1993). First contact in verb acquisition: Defining a role for syntax. *Child Development*, *64*, 1665–1687.

Nazzi, T., & Bertoncini, J. (2003). Before and after the vocabulary spurt: Two modes of word acquisition? *Developmental Science, 6*(2), 136–142.

Ninio, A. (1980). Picture-book reading in mother–infant dyads belonging to two subgroups in Israel. *Child Development*, *51*, 587–590.

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development*, *76*(4), 763–782.

Parisien, C., Fazly, A., & Stevenson, S. (2008). An incremental Bayesian model for learning syntactic categories. In A. Clark & K. Toutanova (Eds.), *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL)* (pp. 89–96), Manchester, UK: Association for Computational Linguistics.

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

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*, 819–865.

Reznick, S. J., & Goldfield, B. A. (1992). Rapid change in lexical development in comprehension and production. *Developmental Psychology*, *28*(3), 406–413.

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

Schachter, F. F. (1979). *Everyday mother talk to toddlers: Early intervention*. London: Academic Press.

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

Smith, L. B. (2000). Learning how to learn words: An associative crane. In R. M. Golinkoff & K. Hirsh-Pasek, (Eds), *Becoming a word learner, a debate on lexical acquisition* (pp. 51–80). New York: Oxford University Press.

Smith, L. B., & Yu, C. (2007). Infants rapidly learn words from noisy data via cross-situational statistics. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. (pp. 635–658). Austin, TX: Connitive Science Society.

Smith, L. B., Yu, C., & Pereira, A. (2007). From the outside-in: Embodied attention in toddlers. In, *Proceedings of 9th European Conference of Artificial Life*. Lisbon, Portugal.

Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb–argument structure: An alternative account. *Journal of Child Language*, *28*, 127–152.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Veneziano, E. (2001). Displacement and informativeness in child-directed talk. *First Language*, *21*(63), 323.

Woodward, A. M., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18-month-olds. *Developmental Psychology*, *30*(4), 553–566.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Yu, C. (2005). The emergence of links between lexical acquisition and object categorization: A computational study. *Connection Science*, *17*(3–4), 381–397.

Yu, C. (2006). Learning syntax–semantics mappings to bootstrap word learning. In R. Sun (Ed.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 924–929). Austin, Tx: Cognitive Science Society.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development*, *4*(1), 32–62.

Yu, C., & Ballard, D. H. (2008). A unified model of early word learning: Integrating statistical and social cues. *Journal of Neurocomputing*, *70*(13–15), 2149–2165.

Yurovsky, D., & Yu, C. (2008). Mutual exclusivity in cross-situational learning. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 703–708). Austin, Tx: Cognitive Science Society.