To appear in Proceedings of the 30th Annual Conference of the Cognitive Science Society (CogSci 2008), Washington, D.C.

# A Probabilistic Incremental Model of Word Learning in the Presence of Referential Uncertainty

**Afsaneh Fazly, Afra Alishahi and Suzanne Stevenson**
Department of Computer Science
University of Toronto
{afsaneh,afra,suzanne}@cs.toronto.edu

**Abstract**

We present a probabilistic incremental model of early word learning. The model acquires the meaning of words from exposure to word usages in sentences, paired with appropriate semantic representations, in the presence of referential uncertainty. A distinct property of our model is that it continually revises its learned knowledge of a word's meaning, but over time converges on the most likely meaning of the word. Another key feature is that the model bootstraps its own partial knowledge of word–meaning associations to help more quickly learn the meanings of novel words. Results of simulations on naturalistic child-directed data show that our model exhibits behaviours similar to those observed in the early lexical acquisition of children, such as vocabulary spurt and fast mapping.

## Early Word Learning

Acquiring the meaning of words is a challenging task for children: For an utterance that describes a scene, a child must align each word with the right referent in the scene. Over time, such alignments must be used to extract a meaning for each word that is consistent across all of its usages. One well-known problem in word learning is that of *referential uncertainty*, that is, the child may perceive many aspects of the scene that are unrelated to the perceived utterance (Quine, 1960; Gleitman, 1990). For example, a child may hear the sentence *Jo rolled the ball*, but observe that "Jo is happily touching a red ball with her hand and slowly rolling it while her mother is talking to her". However, over time, the child can establish an association between the word "ball" and the round object that the word refers to.

Learning the meaning of words has been suggested to be based on cross-situational observation (Pinker, 1989): The meaning of a word is consistent across multiple usages, and can be learned by detecting the set of meaning elements that are common across all situations in which a word occurs. In its original form, this hypothesis is not precisely specified; moreover, it does not provide the flexibility needed for word learning, especially in handling noisy or ambiguous data. A detailed account of this mechanism is needed in order to explore the possibility of learning word meanings in a naturalistic environment, and to account for many general patterns observed in child experimental data. These patterns include the vocabulary spurt (i.e., a slow stage of learning, followed by a sudden increase in the learning rate), fast mapping (i.e., the ability to map a novel word to a novel object in a familiar context), and the effect of the age of acquisition of words on their processing speed. Computational modeling is a powerful tool for precise investigation of the hypothesized mechanisms of word learning, and for studying the suggested patterns.

Many computational models of word learning have been used to simulate and account for the observed patterns such as fast mapping and the vocabulary spurt. However, most of these models use input data that consists of pairings of a single word and its semantic representation, and ignore the problem of finding the right referent for each word in an utterance (Regier, 2005; Li et al., 2004; Xu & Tenenbaum, 2007). Other models simulate learning the meaning of words in the context of an utterance, but they either use artificial input data with controlled referential uncertainty, which may deviate from children's naturalistic learning environments (Siskind, 1996), or rely on cognitively implausible learning strategies and ignore the problem of referential uncertainty (Yu, 2005). These properties make the existing computational models still inadequate for a careful investigation of the patterns of child word learning in a realistic setting.

We propose a novel incremental model of early word learning in the face of referential uncertainty. Our computational model proposes 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. We evaluate our model on naturalistic child-directed data, and show that the overall behaviour of the model is reminiscent of the general patterns observed in children. Moreover, our experimental results show that learning the meaning of words is a much harder task when the input contains referential uncertainty, illustrating the importance of modeling this aspect of word learning. Our model thus provides an appropriate testbed for investigating the impact of referential uncertainty on the word learning process.

## Related Computational Models

The rule-based model proposed by Siskind (1996) is the first to simulate the process of learning word meanings in the presence of referential uncertainty. The model relies on a set of principles to constrain hypotheses about the meaning of words, such as the principle of contrast and the principle of inclusivity. The model is tested on artificially generated input consisting of highly controlled referential uncertainty. It is shown that under these circumstances the meaning of words can be learned, and certain types of noise can be handled by detecting and ruling out the inconsistent input. However, the rule-based nature of the model limits its 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 or ambiguous data.

Another computational model that uses cross-situational inference is proposed by Yu (2005), which is also used to examine the role of various factors, such as syntax (Yu, 2006), in word learning. However, the model uses the original form of the automatic translation learning algorithm of Brown et al. (1993), which lacks cognitive plausibility: It is non-incremental and learns through an intensive batch processing of a whole training data. Moreover, it is tested on limited experimental data containing a very small vocabulary, and with no referential uncertainty.

Most of the existing models rely on a pairing of a semantic representation with a single word form (or its phonological representation)—as opposed to full utterances—as training data. Connectionist models have been proposed for learning such associations, and investigating various patterns in the process of learning. For example, Li et al. (2004) simulate vocabulary spurt and age of acquisition effects, whereas Horst et al. (2006) examine the role of fast mapping. Regier (2005) proposes an associative exemplar-based model that accounts for the changes observed in children's word learning pattern, such as fast mapping and learning synonymy, without a change in the underlying learning mechanism. The Bayesian model of Xu and Tenenbaum (2007), on the other hand, focuses on how humans generalize and learn category meanings from examples of word usages.

## Overview of the Computational Model

Our word learning model is an adaptation of a model of automatic translation between two languages, proposed by Brown et al. (1993). Unlike the original model (as used by Yu, 2005), ours is incremental and does not require a batch process over the entire data. We explain the details of our model in the following subsections.

### Utterance and Meaning Representations

The input to our word learning model consists of a set of utterance–scene pairs that link an observed scene (what the child perceives) to the utterance that describes it (what the child hears). We represent each utterance as a sequence of words, and the corresponding scene as a set of semantic features (including features irrelevant to the utterance), e.g.:

**Utterance:** *Joe rolled the ball*

**Scene:** { ANIMATE, JOE, ACT, MOTION, ROLL, ARTIFACT, OBJECT, GAME EQUIPMENT, MOTHER, HAND, TALK }

In the Experimental Setup section, we explain how the utterances and the corresponding semantic features are selected, and how we add referential uncertainty.

Given a corpus of such utterance–scene pairs, our model learns the meaning of each word $w$ as a probability distribution, $p(.|w)$, over the semantic features appearing in the corpus. In this representation, $p(f|w)$ is the probability of feature $f$ being part of the meaning of word $w$. In the absence of any prior knowledge, all features can potentially be part of the meaning of all words. Hence, prior to receiving any usages of a given word, the model assumes a uniform distribution over semantic features as its meaning.

### The Learning Algorithm

Our model combines probabilistic interpretations of cross-situational learning (Quine, 1960) and a variation of the principle of contrast (Clark, 1990),[1] through an interaction between two types of probabilistic knowledge acquired and refined over time. Given an utterance–scene pair received at time $t$, i.e., $(\mathrm{U}^{(t)}, \mathrm{M}^{(t)})$, the model first calculates an alignment probability $a$ for each $w \in \mathrm{U}^{(t)}$ and each $f \in \mathrm{M}^{(t)}$, using the meaning $p(.|w)$ of all the words in the utterance prior to this time. The model then revises the meaning of the words in $\mathrm{U}^{(t)}$ by incorporating the alignment probabilities for the current input pair. This process is repeated for all the input pairs, one at a time.

**Step 1: Calculating the alignment probabilities:** For a feature $f \in \mathrm{M}^{(t)}$ and a word $w \in \mathrm{U}^{(t)}$, the higher the probability of $f$ being part of the meaning of $w$ (according to $p(f|w)$), the more likely it is that $f$ is aligned with $w$ in the current input. In other words, $a(w|f, \mathrm{U}^{(t)}, \mathrm{M}^{(t)})$ is proportional to $p^{(t-1)}(f|w)$. In addition, if there is strong evidence that $f$ is part of the meaning of another word in $\mathrm{U}^{(t)}$—i.e., if $p^{(t-1)}(f|w_k)$ is high for some $w_k \in \mathrm{U}^{(t)}$ other than $w$—the likelihood of aligning $f$ to $w$ should decrease (principle of contrast). Combining these two requirements:

$$a(w|f, \mathrm{U}^{(t)}, \mathrm{M}^{(t)}) = \frac{p^{(t-1)}(f|w)}{\sum_{w_k \in \mathrm{U}^{(t)}} p^{(t-1)}(f|w_k)} \qquad (1)$$

General features such as ARTIFACT or ENTITY are part of the meaning of, and thus co-occur with, many words in language. Therefore, in an input pair, they are usually aligned with more than one word in the utterance. Over time, the model correctly learns a relatively strong association between such features and the appropriate words, although their association is less strong than those of more specific features.

**Step 2: Updating the word meanings:** We need to update the probabilities $p(.|w)$ for all words $w \in \mathrm{U}^{(t)}$, based on the evidence from the current input pair reflected in the alignment probabilities. We thus add the current alignment probabilities for $w$ and the features $f \in \mathrm{M}^{(t)}$ to the accumulated evidence from prior co-occurrences of $w$ and $f$. We summarize this cross-situational evidence in the form of an association score, which is updated incrementally:

$$\mathrm{assoc}^{(t)}(w, f) = \mathrm{assoc}^{(t-1)}(w, f) + a(w|f, \mathrm{U}^{(t)}, \mathrm{M}^{(t)}) \qquad (2)$$

where $\mathrm{assoc}^{(t-1)}(w, f)$ is zero if $w$ and $f$ have not co-occurred before. The association score of a word and a feature is basically a weighted sum of their co-occurrence counts: Instead of adding one each time the two have appeared in a pair together, we add a probability (a value between zero and

[1] We assume that a feature in a scene is highly associated with only one of the words *in the corresponding utterance*. This differs from what is widely known as principle of contrast, in that the latter assumes contrast across the entire vocabulary.

one) that reflects the confidence of the model that their co-occurrence is indeed because $f$ is part of the meaning of $w$.

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

$$p^{(t)}(f|w) = \frac{\text{assoc}^{(t)}(f,w) + \lambda}{\sum_{f_j \in \mathcal{F}} \text{assoc}^{(t)}(f_j,w) + \beta \times \lambda} \quad (3)$$

where $\mathcal{F}$ is the set of all features seen so far, $\lambda$ is a smoothing factor for allowing a small probability for unseen features, and $\beta$ is the expected number of feature types. The denominator is a normalization factor to get proper probabilities.

### Handling Referential Uncertainty

Our model updates the meaning of a word every time it is heard in an utterance. This flexibility, in addition to the probabilistic nature of the learning, allows the model to handle referential uncertainty. Recall that we simulate referential uncertainty in the form of additional semantic features that are irrelevant to the meanings of the words in an utterance. We expect that the irrelevant features do not regularly co-occur with a given word (in contrast to the relevant features). Thus the overall association score between an irrelevant feature and the word is expected to be lower than that of a relevant feature. This in turn will lower the probability of the irrelevant features in $p(.|w)$.

Another strategy we adopt for handling referential uncertainty is the addition of a dummy word to every utterance when updating the alignment probabilities. This is to allow the possibility of aligning the irrelevant semantic features to the dummy word, hence lowering their alignment probability with the words in the utterance. Note, however, that nothing indicates to the model a priori which features are relevant and which are irrelevant.

### Word Comprehension Score

To evaluate our model, we need to verify how accurately the model learns the meaning of words. We thus define a comprehension score, $c^{(t)}(w)$, for each word $w$ at time $t$, which compares the learned meaning of $w$, or $p(.|w)$, to the word's *correct* meaning, $\mathcal{T}_w$. The correct meaning of a word is the set of semantic features for that word in our input-generation lexicon.[2] The comprehension score is calculated as in:

$$c^{(t)}(w) \quad = \quad \sum_{f_j \in \mathcal{T}_w} p^{(t)}(f_j|w) \quad (4)$$

where $0 \leq c^{(t)}(w) \leq 1$. Ideally, a word is accurately learned when most of its probability mass $p(.|w)$ is concentrated around its true features (those in $\mathcal{T}_w$). We thus consider a word as learned when its comprehension score exceeds a predefined threshold, $\theta_c$. In our experiments reported in the following sections, we examine the behaviour of our model by looking into the comprehension scores.

[2]Note that the model does not have access to this lexicon for learning; it is used only for input generation and evaluation.

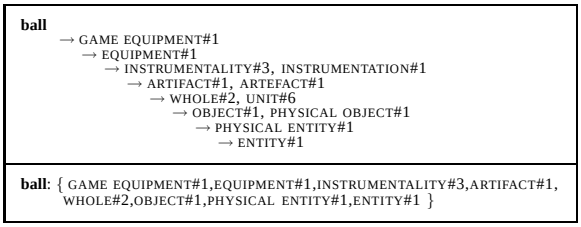



Figure 1: Semantic features for *ball* from WordNet.

## Experimental Set-UP

### The Input Corpora

The training data for our model consists of a sequence of utterances, each paired with a set of semantic features. We extract utterances from the Manchester corpus (Theakston et al., 2001) in the CHILDES database (MacWhinney, 1995). The Manchester corpus contains transcripts of conversations with children between the ages of 1;8 and 3;0. We use the mother's speech from transcripts of 6 children, remove punctuation and lemmatize the words, and concatenate the corresponding sessions as our test data.

There is no semantic representation of the corresponding scenes available from CHILDES. Therefore, we automatically construct a scene representation for each utterance based on the semantic features of the words in that utterance. For nouns and verbs, we extract the semantic features from WordNet[3] as follows: We take all the hypernyms (ancestors) for the first sense of the word, where each hypernym is a set of synonym words (or synsets), tagged with their sense number. For each hypernym, we add the first word in its synset to the set of the semantic features of the target word (see Figure 1 for an example). For adjectives and closed class words (e.g., pronouns), we extract the semantic features using the system of Harm (2002). Other words not found in either of the two resources (e.g., adverbs) are removed from the utterances.

We need to evaluate our model on input that includes referential uncertainty. That is, the representation of the scene must contain semantic features that do not come from the perceived utterance. To simulate such data, we use every other sentence from the original corpus (preserving their chronological order), paired with its own scene representation as well as that of the following sentence. The extra semantic features that are added to each utterance thus correspond to meaningful semantic representations, as opposed to randomly selected features. The resulting corpus has a high rate of referential uncertainty, where on average an utterance is paired with twice as many semantic features as there are in its original meaning set.

### Parameters

We set the parameters of our learning algorithm using a development data set which is similar to our test data, but is selected from a non-overlapping portion of the Manchester corpus. The expected number of semantic features, $\beta$ in Eqn. (3), is set to 7000 based on the total number of distinct features

[3]`http://wordnet.princeton.edu/`

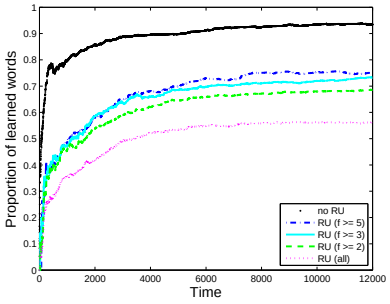



Figure 2: Proportion of learned words over time.

extracted for the development data. The smoothing parameter $\lambda$ in Eqn. (3) is set to a very small value, $10^{-7}$. We set the comprehension threshold, $\theta_c$ in Eqn. (4), to a reasonably high value, 0.7 (recall that comprehension scores are between 0 and 1). This was a value with which our model showed reasonable performance on development corpus. Moreover, 0.7 is a reasonably large portion of the probability mass, given that only a small fraction of the semantic features appear as part of the correct meaning of a word.

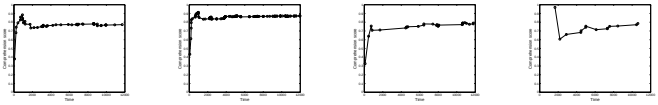
## Experimental Results

In the following sections, we provide a qualitative analysis of our model through examination of its learning patterns. We train the model on the input corpus with referential uncertainty, as explained in the previous section. In order to examine how adding referential uncertainty affects word learning, we repeat most of the reported experiments on the input with no referential uncertainty, and compare the results.

### Effects of Referential Uncertainty

As noted before, one of the main challenges of word learning is the uncertainty inherent in the children's learning environment. To better understand the effect of referential uncertainty (RU) in learning, here we compare the behaviour of our model in two conditions, without RU and with RU. Figure 2 shows the change in the proportion of learned words (those whose comprehension scores exceed the specified threshold) over time, where time is measured as the number of input utterance–scene pairs processed. The bottom curve shows the learning pattern for input with RU, and the top one shows the results for data without RU. As expected, 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. In addition, Figure 2 shows that the task of word learning is much easier in the absence of RU, reflected in the sharp vocabulary growth, as well as in the high proportion of learned words (above 90%) in this condition.

To further elucidate the notable drop in the proportion of learned words when there is RU, we look into the relation between a word's frequency and how easily the model learns it. We examine the learning curves when low frequency words are removed. (Note that low frequency words are only removed from the evaluations, and not from the input data.)



*book* (f=54) *car* (f=148) *fish* (f=31) *kiss* (f=14)

Figure 3: Comprehension scores of four words over time.

Figure 2 displays three learning curves including only words which are heard at least twice, three times or five times.

A comparison of the curves shows that the more frequent a word is, the more likely it is to be learned. These results confirm our hypotheses that in the presence of RU, the model needs more instances of a word usage to learn it with high confidence. Nonetheless, even for words with a minimum frequency of 5, the learning is still more difficult when there is referential uncertainty.

### Convergence and Learning Stability

Our learning algorithm revises the meaning of a word every time it is heard in an utterance (in contrast, e.g., to Siskind, 1996's model). This is a key property that makes our model flexible so it 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 the 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.

Figure 3 shows the change in the comprehension scores of four sample words over time. As expected, the comprehension scores show some fluctuation at the beginning, but they converge on a high value as more examples are observed. We also examine 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 $\theta_c$). The average comprehension score of all words increases rapidly and becomes stable around 0.7 after processing almost $4,000$ input pairs, reflecting the stability in learning. As expected, the average comprehension score of the learned words increases more quickly and reaches a higher value (around 0.8). With no RU, the average comprehension scores show similar increasing patterns, but are generally higher, reflecting easier learning.

### Vocabulary Growth and Fast Mapping

Longitudinal studies of early vocabulary growth in children have shown that vocabulary learning is slow at earlier stages of learning, then proceeds to a rapid pace, and finally becomes less active (Carey, 1978)—a phenomenon often referred to as "vocabulary spurt". Here, we look at the change in the learning rate over time to see whether the pattern of vocabulary growth in our model matches this observation.

Figure 4 depicts the proportion of learned words against the number of word types heard at each time, both without

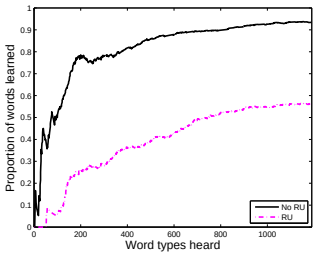


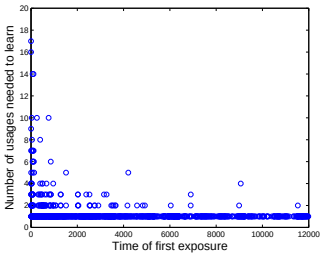

Figure 4: Rate of vocabulary growth as new words are heard.

and with referential uncertainty. Without RU, the learning rate is immediately high, rather than a period of slow growth followed by a spurt as in children. The expected vocabulary growth pattern is more pronounced with RU: There is little learning prior to hearing about 150 words. This can be attributed to the property of our model that uses its own learned knowledge of word meaning to facilitate the learning of new words. After this sudden increase in the number of learned words, the learning proceeds with a nearly constant rate. It is important to note that since we test our model on realistic data, we do not have a fixed vocabulary, and therefore, new words are heard continually. The learning thus does not stop, but it gradually becomes slower, perhaps mainly due to a corresponding decrease in the rate of hearing new words.

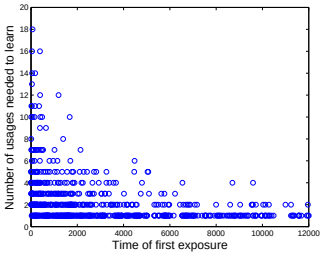
The observed shift from slow to fast word learning is sometimes tied with a phenomenon referred to as fast mapping (Carey, 1978). Fast mapping states that once children have learned a repository of words, they can easily link novel words to novel objects in a familiar context based only on a single (or few) exposures. Many researchers believe that the delay in the onset of fast mapping in children is not due to a change in the underlying learning mechanisms, but is a result of processing more input data (Regier, 2005; Horst et al., 2006). To examine this hypothesis in our model, we look at the interaction between the number of usages that the model needs to learn a word, and the word's age of exposure, defined as the first time the word is heard. Figure 5 depicts the plots (for the words that are learned at some point in time), both without and with RU. In both cases, the model shows clear fast mapping behaviour: Words received later in time, on average, require fewer usages to be learned. With referential uncertainty, fast mapping occurs much more gradually. These results show that our model exhibits fast mapping patterns once it has been exposed to enough word usages, and that no change in the learning mechanism is needed.

## Age of Acquisition Effect

Recent studies have suggested that age of acquisition (AoA), independently of word frequency, affects the speed of processing a word. For example, AoA is shown to be a good predictor of the adult's speed in word naming (Tamminen & Gaskell, 2006) or lexical decision (Nazir et al., 2003; Tamminen & Gaskell, 2006). One problem with these studies is that they cannot accurately estimate the age of acquisition of a word, and mostly rely on the subjective adult AoA ratings.




(a) No referential uncertainty.




(b) With referential uncertainty.

Figure 5: Number of usages needed to learn a word vs. the time of first exposure for that word.

Studying the AoA effect in a computational model such as ours offers the advantage of having direct access to the exact age at which the model has acquired each word. More importantly, different interpretations of the age of acquisition can be investigated: Whether AoA refers to the time of the first *exposure* to a word, or the age at which the model/child can correctly *comprehend* or *produce* the word. Moreover, most of the reported studies on AoA effects in humans, as well as the computational modeling of these effects, have been focused either on the association between the phonological form of a word and the corresponding written form (as in the word naming task), or on the familiarity of a written form (as in the lexical decision task). Few studies have been performed on whether similar AoA effects can be observed in tasks that rely on the association between a word form and its *meaning* (but see Li et al., 2004). Our model provides an appropriate testbed for investigating whether AoA effects can be observed in the context of learning word meaning, especially in the presence of referential uncertainty.

To simulate AoA effects in our model, we need to estimate two factors. We estimate the processing speed of a word $w$ at a time $t$ as its comprehension score $c^{(t)}(w)$, as given in Eqn. (4) (assuming that words that have a higher comprehension score can be accessed and processed faster). We consider two different estimations for the age of acquisition of a word: First, the onset of the word in the training data, or its age of 'Exposure'; and second, the first time the model correctly learns the word, or its age of 'Learning'.

For each AoA condition, we compile two sets of words from the training data, an 'Early' set, which contains words acquired at an earlier stage of learning, and a 'Late' set, containing words acquired at a later stage. We consider the time span 5,000–10,000 as the earlier stage, and the time span 10,000–15,000 as the later stage (we skip over the time span

| | Low Frequency | | High Frequency | |
|---|---|---|---|---|
| Condition | Early | Late | Early | Late |
| Exposure | 0.65 (70) | **0.70** (35) | **0.84** (8) | 0.82 (1) |
| Learning | 0.82 (38) | **0.85** (32) | **0.85** (18) | 0.79 (4) |

Table 1: Average comprehension score for Early and Late word sets for the Exposure and Learning conditions, controlled for frequency. Size of the sets is shown in parentheses.

1–5,000 in order for the model to stabilize). To control for frequency, we only include words whose frequency after processing all 15,000 input pairs falls into a certain range.

Table 1 shows the average comprehension scores of the words in the Early and Late sets after processing 15,000 pairs. The scores are calculated for two conditions, Exposure and Learning, and for two different frequency ranges, Low Frequency (between 2 and 4) and High Frequency (between 6 and 10). The results show an interesting pattern: for more frequent words, an AoA effect can be observed, i.e., words acquired earlier are, on average, easier to comprehend (and therefore easier to process). In contrast, for low frequency words, an opposite effect can be observed for both conditions, i.e., words acquired later are easier to comprehend. This suggests that age of acquisition mainly affects the processing speed of words that are well-entrenched, and that the infrequent words can be remembered only if they have been acquired later. However, this prediction may not be reliable due to the small number of the high frequency words in each set, and further research is needed to confirm it.

## Conclusion and Future Directions

We have presented a computational model of word learning that draws on cognitively plausible mechanisms, such as cross-situational observation and the principle of contrast. The model employs a probabilistic learning algorithm that incrementally updates word meanings based on the observed pairings of utterances and scene representations. Our experimental results show that the model can successfully handle referential uncertainty, and many general patterns in child experimental data can be observed and accounted for in our model. A key property of the proposed model is that its previously acquired knowledge is not corrupted by processing more input. This makes the model suitable for handling synonymy and homonymy, which we plan to explore in the future. Also, we have shown that the model is robust against noisy data. In the future, we need to add new types of noise to the training data (e.g., having words in the utterance whose meanings do not appear in the scene representation), and evaluate the the model under these conditions.

Our model processes words in the context of a sentence, in contrast to the majority of the existing computational models that study words in isolation. Having access to the context of words would enable us to embed additional cues in our learning algorithm, such as the word order of the utterance that a word appears in. Moreover, we intend to explore syntactic and semantic categorization of words, using the word meanings acquired by our model. Such word categories can in turn be used as feedback to our word learning model through a bi-directional bootstrapping process. In future work, we will explore these options, and examine the impact of new factors on the learning pattern of the model.

## References


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.* The MIT Press.

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

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

Harm, M. W. (2002). *Building large scale distributed semantic feature sets with WordNet*. Carnegie Mellon University.

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 *Proc. of CogSci'06.*

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

MacWhinney, B. (1995). *The CHILDES project: Tools for analyzing talk* (2nd ed.). Lawrence Erlbaum Associates.

Nazir, T. A., Decoppet, N., & Aghababian, V. (2003). On the origins of age-of-acquisition effects in the perception of printed words. *Developmental Science*, *6*(2), 143–150.

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

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

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

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

Tamminen, J., & Gaskell, M. G. (2006). Learning new words: Effects of lexical competition and age of acquisition. In *Proc. of CogSci'06.*

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.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psych. 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 *Proc. of CogSci'06.*