Integrating Syntactic Knowledge into a Model of Cross-situational Word Learning

Afra Alishahi
Computational Linguistics and Phonetics
Saarland University, Germany
afra@coli.uni-saarland.de

Afsaneh Fazly
Computer Sciences and Engineering
Shiraz University, Iran
fazly@cse.shirazu.ac.ir

Abstract

It has been suggested that children learn the meanings of words by observing the regularities across different situations in which a word is used. However, experimental studies show that children are also sensitive to the syntactic properties of words and their context at a young age, and can use this information to find the correct referent for novel words. We present a unified computational model of word learning which integrates cross-situational evidence with the accumulated semantic properties of the lexical categories of words. Our experimental results show that using lexical categories can improve performance in learning, particularly for novel or low-frequency words in ambiguous situations.

Learning the Meaning of Words

In the course of learning a language, children need to learn mappings between words and their meanings, mostly from noisy and ambiguous contexts. It has been suggested that children learn the meanings of words by observing the regularities across different situations in which a word is used, or the *cross-situational evidence* (Quine, 1960; Pinker, 1989). Experimental studies on children and adult learners have shown that both groups are sensitive to cross-situational evidence, and can efficiently use it to deduce the correct meanings of novel words in ambiguous situations (Smith & Yu, 2007; Monaghan & Mattock, 2009). Moreover, many computational models have demonstrated that cross-situational learning is a powerful and efficient mechanism for learning the correct mappings between words and meanings, and can explain several behavioural patterns observed in children (Siskind, 1996; Yu, 2005; Fazly et al., 2008).

Another valuable source of information for mapping words to meanings is the syntactic structure of the sentence that a word appears in. There is substantial evidence that children are sensitive to the structural regularities of language from a very young age, and that they use these structural cues to find the referent of a novel word (e.g. Naigles & Hoff-Ginsberg, 1995; Gertner et al., 2006), a hypothesis known as syntactic bootstrapping (Gleitman, 1990). The syntactic bootstrapping account is in accordance with children's early sensitivity to distributional properties of language: one-year-old infants can recognize sentences from an artificial grammar after a short period of exposure (Gomez & Gerken, 1999), and 2 to 3-year-olds demonstrate robust knowledge of some of the abstract lexical categories such as nouns and verbs (e.g., Gelman & Taylor, 1984; Kemp et al., 2005). Therefore, it is likely that they draw on their knowledge of the structural regularities of language (and of lexical categories in particular) to facilitate word learning, especially in cases where cross-situational evidence is not reliable. However, a coherent account of word learning that explains the interaction between these two information sources is lacking. Also, despite the extensive body of experimental research on the role of syntactic knowledge in semantics acquisition, few computational models have been developed to explore the usefulness of lexical categories in learning word meanings (but see Yu, 2006).

We present a probabilistic model of word learning which integrates cross-situational evidence and the knowledge of lexical categories into a single learning mechanism. We use an existing computational model of cross-situational learning proposed by Fazly et al. (2008), and augment it with the syntactic categories of words. Our computational simulations show that such information can improve the model's performance in learning words. Especially, the results suggest that the syntactic category of a word and the context the word appears in provide complementary information for the acquisition of word–meaning mappings.

Related Computational Models

A number of computational word learning models have used cross-situational learning as their core mechanism for mapping words to meanings. The rule-based model of Siskind (1996) and the probabilistic models of Yu (2005) and Fazly et al. (2008) all rely on the regularities of the co-occurrences of words and meaning elements, successfully learning word meanings from noisy and ambiguous data. Moreover, these models simulate several behavioural patterns observed in children, such as vocabulary spurt, fast mapping, and learning synonymy and homonymy. However, all these models ignore the syntactic properties of the utterances and treat them as unstructured bags of words.

There are only a few existing computational models that explore the role of syntax in word learning. Maurits et al. (2009) has investigated the joint acquisition of word meaning and word order using a batch model. This model is tested on an artificial language with a simple relational structure of word meaning, and limited built-in possibilities for word order. The Bayesian model of Niyogi (2002) simulates the bootstrapping effects of syntactic and semantic knowledge in verb learning, i.e., the

use of syntax to aid in inducing the semantics of a verb, and the use of semantics to narrow down possible syntactic forms in which a verb can be expressed. However, this model relies on extensive prior knowledge about the associations between syntactic and semantic features, and is tested on a toy language with very limited vocabulary and a constrained syntax. Yu (2006) integrates information about syntactic categories of words into his model of cross-situational word learning, showing that this type of information can improve the model's performance. Yu's model also processes input utterances in a batch mode, and its evaluation is limited to situations in which only a coarse distinction between referring words (words that could potentially refer to objects in a scene, e.g., concrete nouns) and non-referring words (words that cannot possibly refer to objects, e.g., function words) is sufficient. It is thus not clear whether information about finer-grained categories (e.g., verbs and nouns) can indeed help word learning in a more naturalistic incremental setting.

An Overview of Our Integrated Model

Consider a young language learner hearing the sentence *the kittie is playing with the yarn*, and trying to find out the meaning of *yarn*. Usually there are many possible interpretations for *yarn* based on the surrounding scene, and the child has to narrow them down using some learning strategy. One such method is to register the potential meanings in the current scene, and compare them to those inferred from the previous usages of the same word (i.e., cross-situational learning). Another way to make an informed guess about the meaning of *yarn* is to pay attention to its syntactic properties. For example, if the child has already heard some familiar words in a similar syntactic context (e.g., *daddy is playing with the ball*, *the kittie is sniffing the slipper*), she can conclude that a group of words which can appear in the context "*is Xing the –*" usually refer to physical objects. Therefore *yarn* must refer to one of the objects present in the scene, and not for example to an action or a property.

We present a computational model that combines these two complementary approaches into a single mechanism of word learning. Our goal is to examine whether using the knowledge of word categories in addition to cross-situational observations can improve the performance in word learning. We use the computational model of Fazly et al. (2008) as the base model of cross-situational learning: the model learns word meanings as probabilistic associations between words and semantic elements, using an incremental and probabilistic learning mechanism, and drawing only on the word–meaning co-occurrence statistics gradually collected from naturally-occurring child-directed input. This model has been shown to accurately learn the meaning of a large set of words from noisy and ambiguous input data, and to exhibit patterns similar to those observed in children in a variety of tasks (see Fazly et al., n.d., for a full set of experiments on this model).

In order to augment the base model with category knowledge, we assume that an independent categorization module can process each sentence and determine the lexical category for each word, e.g., based on its surrounding context. That is, we make the simplifying assumption that prior to the onset of word learning, the categorization module has already formed a relatively robust set of lexical categories from an earlier set of child-directed data. This assumption is on the basis of previous empirical findings that young children gradually form a knowledge of abstract categories, such as verbs and nouns (e.g., Gelman & Taylor, 1984). In addition, several computational models have been proposed for inducing reliable categories of words by drawing on distributional properties of their context (see, e.g. Parisien et al., 2008). However, children's acquisition of categories is most probably interleaved with the acquisition of word meaning, and these two processes must be studied simultaneously. As a first step, we investigate whether the word learning process can benefit from the knowledge of lexical categories, assuming that such knowledge exists.

In the next sections we sketch the base model of cross-situational learning, and explain how we extend it to integrate lexical categories as an alternative source of guidance. During the course of learning in both models, we use the feedback from the categorization model to detect different senses of the same word. That is, the same word types which belong to different categories are represented as separate lexical items. For example, the verb sense and the noun sense of the word *cry* are mapped to two independent meaning representations.

Cross-situational Learning

This section explains the details of the cross-situational word learning model of Fazly et al. (2008), which we use as our base model.

Representation of Input

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, for example:

He hit the rabbit { ANIMATE, MALE PERSON, ACT, MOTION, CONTACT, FORCE, ANIMAL, MAMMAL, RABBIT }

In the Evaluation section, we explain how the utterances and the corresponding semantic features are selected.

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

The model proposes a probabilistic interpretation of cross-situational learning (Quine, 1960) through an interaction between two types of probabilistic knowledge acquired and refined over time. Given an utterance–scene pair ($\mathrm{U}^{(t)}$, $\mathrm{S}^{(t)}$) received at time t, the model first calculates an alignment probability a for each $w \in \mathrm{U}^{(t)}$ and each $f \in \mathrm{S}^{(t)}$, using the meaning $p(.|w)$ of all the words in the utterance prior to this time. It 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 $f \in \mathrm{S}^{(t)}$ 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, $a(w|f, \mathrm{U}^{(t)}, \mathrm{S}^{(t)})$ 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. Combining these two requirements:

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

Note that a feature can have a non-zero alignment with more than one word in an utterance. For example, if two concrete nouns occur in a sentence, they both need to be aligned with the single feature ARTIFACT.

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 $f \in \mathrm{S}^{(t)}$ 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:

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

where $\mathrm{assoc}^{(t-1)}(w, f)$ is zero if w and f have not co-occurred before. The model then uses these association scores to update the meaning of the words in the current input, as in:

$$p^{(t)}(f|w) = \frac{\mathrm{assoc}^{(t)}(f, w)}{\sum_{f_j \in \mathcal{F}} \mathrm{assoc}^{(t)}(f_j, w)} \quad (2)$$

where $\mathcal{F}$ is the set of all features seen so far. We use a smoothed version of this formula to accommodate noisy or rare input, as explained in Fazly et al. (n.d.).

Word Acquisition Score

To evaluate our model, we need to verify how accurately the model learns the meaning of words. We thus define the *acquisition score* of a word w at time t as an estimation of how closely the meaning probability $p^{(t)}(.|w)$ resembles the *correct* meaning of w, or $\mathcal{T}_w$. The correct meaning of a word is a set of semantic features according to an input-generation lexicon.[1] Ideally, a word is accurately learned when its relevant semantic features (those in $\mathcal{T}_w$) are ranked at the very top of the distribution $p^{(t)}(.|w)$. We use average precision[2] to measure how well $p^{(t)}(.|w)$ separates the relevant features of w from irrelevant ones.

Adding Lexical Categories to the Model

As mentioned before, we assume that prior to the onset of word learning, the child has formed a number of lexical categories, each containing a set of word forms. More formally, we assume that the word learning model has access to a categorization function $\mathrm{cat}(w, \mathrm{U}^{(t)})$ which at any time t during the course of learning can determine the category of a word w in utterance $\mathrm{U}^{(t)}$. We do not make any assumptions about the details of the categorization process, except that it does not rely on the meaning of words in order to find their appropriate category.

As the model learns meanings for words, the categories that these words belong to are implicitly assigned a meaning as well. Once the word learning process begins, we assign a meaning distribution to each category on the basis of the meanings learned for its members. Formally, we define the meaning of a category C as the average of the meaning distributions of its members, as in:

$$p^{(t)}(f|C) = \frac{1}{|C|} \sum_{w \in C} p^{(t)}(f|w) \quad (3)$$

where $|C|$ is the number of word forms in category C, and $p^{(t)}(f|w)$ is the meaning probability of word w for feature f at time t. Prior to observing any instances of the members of a category in the cross-situational input, we assume a uniform distribution over all the possible meaning elements for each category.

[1] The model does not have access to this lexicon for learning; it is used only for input generation and evaluation.

[2] Precision is calculated as the proportion of the number of features from $\mathcal{T}_w$ to the total number of features at each cut-off point in the ranked list $p^{(t)}(.|w)$. The acquisition score is the average over the precisions for all the cut-off points up to the point where all the features in $\mathcal{T}_w$ are included in the ranked list. Note that this score is 1 when the probabilities assigned to all of the relevant features of w are higher than those assigned to the irrelevant features.

Using Categories in Alignment

Knowledge of word categories is integrated into the base model in the alignment phase (i.e. **Step 1** of the learning algorithm), where we decide which semantic feature in an observed scene must be aligned with which word(s) in the accompanying utterance. Given a new utterance–scene pair, we can align words in the utterance with the semantic features in the observed scene based on the cross-situational evidence that we have accumulated so far. Alternatively, we can find the category for each word and use the meaning associated with the word category as a guide to align it with the best matching semantic features from the scene. We can merge these two pieces of information into an extended version of Eqn. (1):

$$\begin{aligned} a(w|f,\, \mathrm{U}^{(t)},\, \mathrm{S}^{(t)}) \quad &= \mathit{weight}(w) \cdot a_w(w|f,\, \mathrm{U}^{(t)},\, \mathrm{S}^{(t)}) \\ + \quad &(1 - \mathit{weight}(w)) \cdot a_c(w|f,\, \mathrm{U}^{(t)},\, \mathrm{S}^{(t)}) \end{aligned}$$

The word-based alignment score $a_w(w|f,\, \mathrm{U}^{(t)},\, \mathrm{S}^{(t)})$ is calculated as in Eqn. (1). The category-based alignment score $a_c(w|f,\, \mathrm{U}^{(t)},\, \mathrm{S}^{(t)})$ is calculated in a similar fashion, except it relies on the meaning of the word category:

$$a_c(w|f,\, \mathrm{U}^{(t)},\, \mathrm{S}^{(t)}) = \frac{p^{(t-1)}(f|\mathrm{cat}(w, \mathrm{U}^{(t)}))}{\sum_{w_k \in \mathrm{U}^{(t)}} p^{(t-1)}(f|\mathrm{cat}(w_k, \mathrm{U}^{(t)}))}$$

where the meaning probability $p^{(t-1)}(f|\mathrm{cat}(w_k, \mathrm{U}^{(t)}))$ is calculated as in Eqn. (3).

The relative contribution of the word-based versus the category-based alignment is determined by the function $\mathit{weight}(w)$. It has been shown that cross-situational evidence is a reliable cue for frequent words: Fazly et al. (n.d.) show that once their model receives a few instances of a word form, it can reliably align it with proper semantic features. On the other hand, the category-based score is most informative when the model encounters a low-frequency word. Therefore, we define $\mathit{weight}(w)$ as a function of the frequency of w:

$$\mathit{weight}(w) = \frac{\mathit{freq}(w)}{\mathit{freq}(w) + 1}$$

Once the overall alignment score is calculated for the new input pair, the meaning probabilities of words are updated through **Step 2** of the original learning algorithm, and the meaning of their corresponding categories are updated accordingly.[3]

Evaluation

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

[3]For each word w in $U^{(t)}$, the meaning distribution of the corresponding category C is incrementally updated as $p^{(t)}(f|C) = p^{(t-1)}(f|C) + \frac{1}{|C|}(p^{(t)}(f|w) - p^{(t-1)}(f|w))$.

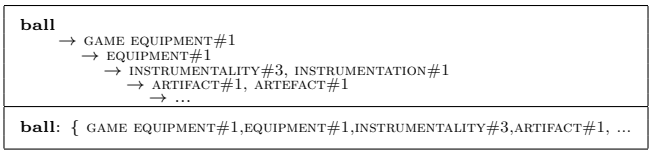

Figure 1: Semantic features for *ball* from WordNet.

(MacWhinney, 1995), which 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. We automatically construct a scene representation for each utterance based on the semantic features of the words in that utterance. For nouns, we extract the semantic features from WordNet[4] as follows: We take all the hypernyms (ancestors) for the first sense of the word, and add the first word in the synset of each hypernym to the set of the semantic features of the target word (see Figure 1 for an example). For verbs, we extract features from WordNet as well as from a verb-specific resource, VerbNet.[5] For adverbs, adjectives and closed class words we use the features of Harm (2002). Words not found in these three resources are removed from the utterance.

To form the initial lexical categories, we use a non-overlapping portion of the part-of-speech tagged version of the Manchester corpus. The original corpus has 60 fine-grained tags, which we map to 11 coarser-grained categories, such as Noun, Verb, and Preposition.[6]

Learning Curves

To understand whether category information improves learning of word–meaning mappings, we compare the pattern of word learning over time for two models: the base model which only uses cross-situational evidence, and the extended model which incorporates lexical categories into learning. For each model we measure the average acquisition score (defined on page 3) of all words that the model has encountered up to each point in time.

Figure 2 shows the learning curve for each model over 5000 time units (or processed input pairs). The curves show that the extended model consistently outperforms the base model. The improvement is more pronounced as the model receives more input, since by learning more about the meanings of words the model also forms a more reliable knowledge about the meanings of categories and can use them more efficiently in aligning the novel words with their referents.

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

[5]`http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`

[6]We thank Chris Parisien for providing us with the coarse-grained tagging of the corpus.

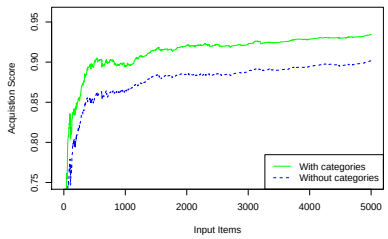

Figure 2: Avg. acquisition score for all words over time, with and without using lexical categories.

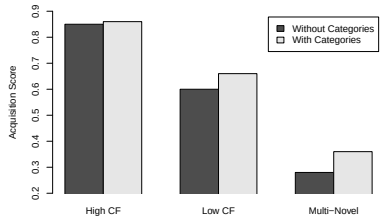

Figure 3: Avg. acquisition score for words in contexts with different degrees of familiarity.

Categories and Context Familiarity

The learning curves presented above show an overall improvement when lexical categories are incorporated in word learning. However, we expect the gain from including categories to vary across different situations. For example, experimental and computational studies have shown that cross-situational learning can account for accurate mapping of a novel word to a novel object in a familiar context (see Fazly et al. (n.d.) for a discussion on this phenomenon). The same pattern is expected in our base model, where the alignment between a word and a semantic feature is in part determined by what the model has learned about the possible meanings of the co-occurring context words (see Eqn.1). Therefore, it can learn a lot about a novel word from a single exposure if that word appears in a familiar context.

We hypothesize that categories can be particularly helpful in cases where a novel word first appears in an unfamiliar context (where not all words in the utterance are accurately learned), or when an utterance contains more than one novel word. To investigate this hypothesis, we introduce a context familiarity measure CF as the mean *familiarity* of all words that co-occur with a target word, where the familiarity of a word is determined by its frequency range. The mappings between familiarity values and frequency ranges are as follows: 0 (0), 1 (1), 2 (2–4), 3 (5–9), 4 (10–29), and 5 ($\geq$ 30), where the numbers in parentheses specify the frequency range.

Figure 3 shows the average acquisition score of words with high and low context familiarity (CF $\leq$ 3 vs. CF $>$ 3), and for novel words which appear in the company of other novel words (this last condition is marked as Multi-Novel in Figure 3). The average scores are calculated by both models after the first occurrence of each word. As can be seen, the inclusion of categories leads to a statistically significant improvement for words in all three conditions ($p < 0.05$).[7] However, the improvement is much more pronounced for words with low context familiarity, and particularly when an utterance includes more than one novel word (i.e. a highly unfamiliar context). These results support our hypothesis, and suggest that in learning the meaning of words, the context of a word and its lexical category can be seen as complementary sources of knowledge.

[7]The p-values are measured according to a two-sided sign test for a confidence interval of 95%.

Comparing Different Categories

To better understand the impact of lexical categories on word learning, we examine the pattern of improvement for words with different parts of speech. Lexical categories differ in their frequency of occurrence and in their semantic properties. For example, open-class categories such as Verb and Noun tend to have lower token frequency, higher type frequency, and more within-class meaning variability compared to closed-class categories such as Determiner and Preposition.

Recall that words in our test corpus are tagged with one of 11 coarse-grained parts of speech. Three of these categories (Auxiliary, Infinitive and Negation) each contain only a single word type, and one (Other) is not a coherent and meaningful category. The average acquisition score in both models for the remaining categories are shown in Figure 4. Out of these seven categories, four are open-class: Noun (599 word types), Verb (261), Adjective (60), and Adverb (25), and three are closed-class: Determiner (23), Preposition (17), and Conjunction (8).

Interestingly, we observe that category information helps more with the acquisition of open-class words, in particular Noun ($p < 10^{-16}$) and Verb ($p < 0.0001$). We believe this difference is due to the high token frequency of closed-class words which makes them very easy to learn, even for the base model that does not take into account the information about their categories. Moreover, using categories does not significantly improve the acquisition of adjectives and adverbs. We suspect that this is a result of the small number and the inconsistent meaning representations of these categories in the resource of Harm (2002). In general, we predict that using better resources for extracting semantic features will boost the contribution of lexical categories in word learning.

Conclusions and Future Directions

Our computational model of word learning demonstrates the advantage of integrating lexical categories into a cross-situational model of word learning. Drawing on

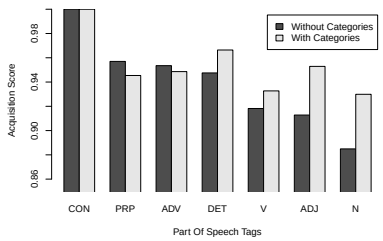

Figure 4: Avg. acquisition score for different categories.

the meaning probabilities of individual words, the model gradually associates each lexical category with a meaning representation, which in turn can boost learning of novel words. Our simulations of the model over the course of acquisition show that using lexical categories consistently improves learning over a base model which only relies on cross-situational evidence. Moreover, our analyses of the results suggest that lexical categories can have a significant impact on the acquisition of open-class words which appear in less familiar context.

The model in its current form makes simplifying assumptions that must be addressed in future work. It is assumed that lexical categories are formed prior to the onset of the word learning process, and that the category of each word can be precisely determined upon its first appearance in the input data. In the future, we intend to use an incremental model of category induction to simultanously learn lexical categories and word meanings. In fact, using a finer-grained set of categories induced by such a model might be more suitable for our purpose, since they can represent more specialized meanings (e.g., fruits and animals instead of nouns). Moreover, the categorization process can benefit from the integration of word meanings in addition to the distributional context. This extension will allow us to study how the early stages of word learning and category formation interact.

Acknowledgment

We would like to thank Grzegorz Chrupała for his invaluable help, and the anonymous reviewers for their insightful comments on our paper. Afra Alishahi was funded by IRTG 715 Language Technology and Cognitive Systems provided by the German Research Foundation (DFG).

References

Fazly, A., Alishahi, A., & Stevenson, S. (n.d.). A probabilistic computational model of cross-situational word learning. *Cognitive Science*. (To appear)

Fazly, A., Alishahi, A., & Stevenson, S. (2008). A probabilistic incremental model of word learning in the presence of referential uncertainty. In *Proc. of CogSci'08.*

Gelman, S., & Taylor, M. (1984). How two-year-old children interpret proper and common names for unfamiliar objects. *Child Development*, 1535–1540.

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*(1), 3–55.

Gomez, R., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135.

Harm, M. W. (2002). *Building large scale distributed semantic feature sets with WordNet* (Tech. Rep. No. PDP.CNS.02.1). Carnegie Mellon University.

Kemp, N., Lieven, E., & Tomasello, M. (2005). Young Children's knowledge of the "determiner" and " adjective" categories. *Journal of Speech, Language and Hearing Research*, *48*(3), 592–609.

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

Maurits, L., Perfors, A. F., & Navarro, D. J. (2009). Joint acquisition of word order and word reference. In *Proc. of CogSci'09.*

Monaghan, P., & Mattock, K. (2009). Cross-situational language learning: The effects of grammatical categories as constraints on referential labeling. In *Proc. of CogSci'09.*

Naigles, L., & Hoff-Ginsberg, E. (1995). Input to verb learning: Evidence for the plausibility of syntactic bootstrapping. *Dev. Psychology*, *31*(5), 827–37.

Niyogi, S. (2002). Bayesian learning at the syntax-semantics interface. In *Proc. of CogSci'02.*

Parisien, C., Fazly, A., & Stevenson, S. (2008). An incremental Bayesian model for learning syntactic categories. In *Proc. of CoNLL'08.*

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

Quine, W. (1960). *Word and object.* Cambridge, MA: MIT Press.

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

Smith, L., & Yu, C. (2007). Infants rapidly learn words from noisy data via cross-situational statistics. In *Proc. of CogSci'07.*

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. *J. of Child Language*, *28*, 127–152.

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.*