The Pursuit of Word Meanings

Jon Scott Stevens (jonsteve@ling.upenn.edu)
Charles Yang (charles.yang@ling.upenn.edu)
Department of Linguistics, 255 S. 36th Street
Philadelphia, PA 19104 USA

John Trueswell (trueswel@psych.upenn.edu)
Lila Gleitman (gleitman@psych.upenn.edu)
Department of Psychology, 3720 Walnut Street
Philadelphia, PA 19104 USA

Abstract

Experimental studies have drawn conflicting conclusions about the mechanisms of word learning. On the one hand, cross situational learning research (e.g. Yu & Smith, 2007) suggests the capability of tracking word-meaning correlations across learning instances. On the other hand, studies have also shown (Medina, Snedeker, Trueswell, & Gleitman, 2011) that the learner appears to attend to a single meaning hypothesis to confirm or disconfirm it against the data. We present a word learning model that aggressively pursues a single, most favored, meaning hypothesis while maintaining probabilistic associations between words and their meaning hypotheses. The model is congruent with experimental results and outperforms two recent, more computationally complex, cross-situational models on a sample of child-directed English. We propose that this difference in performance is due to a "dilution" effect, where ignoring much of the cross-situational information prevents the learner from diluting her probability space with erroneous hypotheses.

Introduction

The pairing between words and their meanings in the environment is messy and unreliable, which formed the earliest arguments against the associationist approach to language learning (Chomsky, 1959). Major research efforts have been devoted to identifying constraints on word learning and their interactions, and to investigate whether they are domain specific or derive from more general principles (see Bloom 2000 for review).

The recent interest in the cross-situational learning approach marks a new direction in the study of word learning. Surely not all words, or every instance of them, will be neatly aligned with their meanings (Landau & Gleitman, 1988). But if the target meaning is associated with a word sufficiently frequently—and more reliably than its competitors—then the learner may be able to detect and learn from such statistical correlations. A number of recent studies have investigated word learning across situations in the laboratory (Medina et al., 2011; K. Smith, Smith, & Blythe, 2011; L. Smith & Yu, 2008; Trueswell, Medina, Hafri, & Gleitman, 2013; Yu & Smith, 2007), and there are several computational models that explore various aspects of this approach (e.g. Fazly, Alishahi, & Stevenson, 2010; Frank, Goodman, & Tenenbaum, 2009; Siskind, 2000; Yu & Smith, 2007). These models suggest that cross-situational statistics may be exploited to acquire the meanings of words.

In this paper, we develop a new model of word learning that draws insights from recent experimental work pointing to the limits of cross-situational learning (Medina et al., 2011; Trueswell et al., 2013). Based on their experimental results, these researchers offered a "Propose-but-Verify" account of word learning across situations: the learner entertains a single meaning hypothesis, retaining it upon confirmation but replacing it with an alternative upon disconfirmation. Our model draws upon this account but departs from it in incorporating probabilistic learning mechanisms which are commonly associated with cross-situational learning, and which have been used in other domains of language acquisition to integrate domain-specific and domain-general learning processes (Yang, 2002, 2004). Algorithmically, the model (dubbed *Pursuit*) employs a "greedy" variant of Reinforcement Learning (Barto & Sutton, 1998) that rewards a single word-meaning hypothesis per instance of a word if the hypothesis is consistent with the input, and punishes it if it is not, while ignoring all other available meanings (like Propose-but-Verify but unlike cross-situational learning). While making incomplete use of word-meaning correlations would seem to handicap the model, simulations on child directed English data show that the Pursuit model outperforms previous models that keep track of the full range of cross-situational meanings.

We first outline the algorithm and compare its computational properties with other models of word learning. We then discuss the congruence of Pursuit with experimental findings before reporting simulation results that show Pursuit to be a better word learning algorithm for a particular set of CHILDES data.

The Pursuit Model of Word Learning

The Algorithm

Like many other word learning models (e.g. Fazly et al., 2010; Frank et al., 2009; Yu & Ballard, 2007) and in line with current formal models of language acquisition (Yang, 2002, 2004), the Pursuit model represents linguistic hypotheses in probabilistic terms. The learner stores a matrix of association values between words and meanings. An association $A(w,m)$ is best understood as the learner's confidence in the meaning m for word w. Associations are strengthened or weakened incrementally based on observation.

The input data is a sequence of utterances $U = (W_U, M_U)$, where W_U and M_U are the sets of words and available meanings in that utterance U. The learner has access to the sets of words, meanings and their associations $(\mathbf{W}, \mathbf{M}, \mathbf{A})$, and adjusts their values after each utterance. For notational convenience, let $\mathbf{A_w}$ for a given word w be the set of associations $\{A(w,x)\}$ for all meanings x that have been hypothesized for w. Similarly, let $\mathbf{A_m}$ for a given meaning m be $\{A(x,m)\}$ for all observed words x. We can speak of the conditional probability of a word meaning, $P(m|w)$, by normalizing $A(w,m)$

INPUT: The learner's words (**W**), meanings (**M**), their associations **A**, and the new utterance $U = (W_U, M_U)$. For every $w \in W_U$:

(a) **Initialization**

If w is a novel word, initiate $\mathbf{A_w} = \{A(w, h_0) = \gamma\}$, where $h_0 = \arg\min_{m \in M_U} \mathbf{A_m}$

(b) **Pursuit**

Select the most probable meaning h for w (i.e., $\arg\max_h A(w,h)$):

i. If h is confirmed ($h \in M_U$), reward $A(w,h)$; go to (c)

ii. If h is disconfirmed ($h \notin M_U$), penalize $A(w,h)$ and reward $A(w,h')$ for a randomly selected $h' \in M_U$

(c) **Lexicon**

If any conditional probability $P(\hat{h}|w)$ exceeds a certain threshold value (θ), then file the $(w, \hat{h})$ into the lexicon.

Box 1

Adjust association $A(w,h)$ against an utterance $U = (W_U, M_U)$ where $w \in W_U$:

If h is confirmed ($h \in M_U$):

$$A(w,m)' = A(w,m) + \gamma(1 - A(w,m))$$

If h is disconfirmed ($h \notin M_U$):

$$A(w,m)' = A(w,m) \times (1 - \gamma)$$

Box 2

with respect to $\mathbf{A_w}$ with smoothing to prevent zero probabilities. This is shown in Equation 1, where N is the number of observed meaning types, and λ is a small smoothing factor.

$$P(m|w) = \frac{A(w,m) + \lambda}{\sum \mathbf{A_w} + N \times \lambda} \quad (1)$$

The term $P(m|w)$ can be viewed as the learner's belief in the word-meaning (w,m) pairing. The smoothing factor λ gives a small amount of probability mass to unseen mappings. If $P(m|w)$ exceeds a certain threshold value, then the learner concludes m to be the meaning of the word w. An outline of the learning algorithm is given in Box 1. Let's first consider the **Pursuit** step, which is the core of our model; we will return to the **Initialization** step, which chooses a meaning candidate for a novel word that the learner encounters for the first time. Our approach could be described as "pursuit with abandon": the privileged status of a single meaning hypothesis comes at the expense of other meanings. This is the fundamental difference between our model and (Medina et al., 2011; Trueswell et al., 2013) from cross-situational learning models, which are defined as the tabulation of multiple, possibly all, word-meaning associations across learning instances.

In **Pursuit**, the learner selects the most favored meaning hypothesis (h) for the word w, i.e., the one with the highest association score. It then adjusts the association score $A(w,h)$ according to its presence or absence in the current utterance. If h is confirmed (i.e., found in M_U, the current set of meanings), then $A(w,h)$ increases, and if h fails to be confirmed, it decreases. In the case of confirmation, the learner ignores all other meanings present in the current utterance and moves on. In the case of disconfirmation, its association score decreases, and the learner randomly chooses a single new meaning from the currently set M_U to boost, again ignoring all other available meanings. After updating $\mathbf{A_w}$ with respect to U, a meaning $\hat{h}$ may emerge as the winner if its conditional probability $P(\hat{h}|w)$, a normalization of $A(\hat{h},m)$ as described in Equation 1, exceeds a certain threshold value (**Lexicon**).

It is instructive to compare the Pursuit model with both cross-situational learning models and the Propose-but-Verify approach (Medina et al., 2011; Trueswell et al., 2013). Like cross-situational learning, the association between words and meanings is probabilistic and dynamically updated in response to the learning data. Like Propose-but-Verify but unlike cross situational learning, the Pursuit model considers only one hypothesis ignores all other meanings upon confirmation. Unlike Propose-but-Verify, however, a disconfirmed meaning is not discarded but only has its association value lowered. Given the Pursuit scheme, a disconfirmed meaning may still remain the most probable hypothesis and will be selected for verification next time the word is presented in the learning data. This crucial features adds considerable robustness to learning behavior as we shall see.

The function that specifies the magnitude of reward/penalty is found in Box 2. Following Yang's Variational model of language acquisition (2002, 2004), we use a simple reinforcement learning model to adjust the association scores for words and their meanings (Bush & Mosteller, 1951). The amount of adjustment is determined by the learning rate γ, usually a small value between 0 and 1. The Pursuit model falls in the subclass of *greedy* algorithms: instead of sampling over hypotheses thereby giving every hypothesis a chance to be selected, the learner simply chooses the most favored hypothesis and ignores all the rest. Under the Pursuit model, as long as the most favored meaning continues to be confirmed, the learner ignores all other competing meanings. This is a familiar idea sometimes known as *error-driven learning* in the formal studies of language acquisition (Berwick, 1985), and appears to be consistent with the experimental findings (Medina et al., 2011; Trueswell et al., 2013). Traditional use of error-driven learning has frequently been tied to categorical learning models (e.g., Gibson and Wexler 1994) but that is neither necessary nor sufficient. The Pursuit model is more in line with the current understanding (Saffran, Aslin, & Newport, 1996; Xu & Tenenbaum, 2007; Yang, 2004) that human language acquisition has a probabilistic component. Additional meanings are only added to the hypothesis set if the most favored hypothesis fails to be located in the current learning instance. Even here, the single-mindedness of Pursuit is evident: the learner chooses only one additional meaning to boost.

Let's now consider the **Initialization** step which deals with novel words that the learner has not seen before. This step encodes a probabilistic form of the Mutual Exclusivity Constraint (Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; Markman & Wachtel, 1988), which has been implemented in various ways by many computational models of word learning (Fazly et al., 2010; Frank et al., 2009; Yu & Ballard, 2007): when encountering novel words, children favor mappings to novel rather than familiar meanings.[1] In our model, the learner chooses a hypothesis that is least likely to be referred to by another word in the learner's hypothesis space. More specifically, the learner associates a novel word with the meaning about which she is least confident, where the learner's confidence in a meaning is defined as the highest probability with which this meaning is associated with some other word. For example, say the new word is "cat", and both CAT and DOG are available meanings.[2] If DOG is already paired with the word "dog" with probability 0.8, and CAT is paired with the word "whisker" with probability 0.6, then learner is less sure about CAT than DOG, making CAT a better guess for the new word "cat"; DOG is ignored completely.

Properties of Pursuit Learning

We now illustrate the key differences between the Pursuit model and cross-situational learning models. For concreteness, we focus on the model of Fazly et al. (2010). This model is an effective incremental variant of Yu & Ballard's original batch learning model (Yu & Ballard, 2007) and most directly captures cross-situational learning experimental results (e.g., L. Smith and Yu 2008; Yu and Smith 2007). Like the current model, Fazly et al. (2010) use associations between words and meanings, which are updated incrementally as new input is received, to calculate word-meaning probabilities. But unlike the current model, Fazly et al.'s cross-situational learner boosts the associations (and thus the probabilities) for *all* meanings that are present during the utterance of a given word; this is a key design feature of cross-situational learning. By contrast, the Pursuit model can only raise the probability of one hypothesis per instance of a word.

Let us first consider the standard cross-situational learning scenario, adapted from L. Smith and Yu (2008) (Fig 1).

By keeping track of word object co-occurrence frequencies, a cross-situational learner can determine that the most likely meaning for "ball" is in fact BALL, which is present across both learning instances.

The Pursuit model, by contrast, will not fare as well. Suppose on the first scene, the learner guesses BALL (with probability of 0.5), it will be selected and confirmed on the scene,

[1] Fazly et al. (2010) build in this preference by having the learner give larger association boosts to newer meanings and smaller association boosts to meanings that are already associated with other words in the utterance. The Bayesian model of Frank et al. (2009) penalizes many-to-one mappings and places a higher prior probability on lexicons with a smaller number of word-meaning pairs, which is also a probabilistic encoding of the Mutual Exclusivity Constraint.

[2] Throughout this paper, we use quotations to denote the form of the word and uppercase to denote meanings.

Figure 1: A canonical cross-situational learning scenario

and learning succeeds. But if the first guess is BAT, it will be disconfirmed on the second scene and the learner selects another meaning from BALL and DOG randomly. Taken together, the Pursuit model only learns BALL 75% of the time, which is transparently sub-optimal compared to the cross situational learning model.

However, recent experimental work suggests that learners do not necessarily keep track of multiple hypotheses as suggested by cross situational learning models (Medina et al., 2011; Trueswell et al., 2013). Subjects were presented with nonsense words accompanied by visual scenes similar to those in Fig 1. After each instance of a word, subjects were asked to guess the word's meaning. It is found that when instances of a word were separated in time by instances of other words, learners appeared to be capable of remembering only their previous guess. If that previous guess was disconfirmed by new data, they appeared to be incapable of remembering which alternative meanings had been present before. To illustrate, imagine the scenes in Fig 1 accompanying utterances of the nonsense word "dax". Subjects who guess the meaning BALL for "dax" after instance 1 are likely to guess BALL after instance 2 as well. But subjects who guess BAT are at chance when choosing between DOG and BALL after instance 2. In other words, the fact that BALL was present in the first instance does not help the learner who made the wrong initial guess. The Pursuit model captures these results, where the favored hypothesis is determined by the highest association score, where other possibilities–both other meanings in the scene and other, but lower ranked, meanings from past experience–are ignored.

This sub-optimal way of making use of information across situations ought to handicap the Pursuit model. However, simulation results show that it actually performs better than cross situational learning. As we discuss below, simulation results and the analysis of word learning data in naturalistic settings suggests that scenarios like the one in Fig 2 are more common than those like in Fig 1, in which case the Pursuit model has an advantage.

After instance 1, the learner has only one choice of BALL as the hypothesized meaning for "ball". This initial guess receives an initial association score of γ (**Initialization**); as the only and thus the most probable hypothesis for "ball", it is

Figure 2: A learning scenario where Pursuit obviates probability dilution

"ball"	BALL	ELEPH.	DOG	BEAR
# of rewards (Pursuit)	3	0	0	0
Probability ≈ (Pursuit)	1	0	0	0
# of rewards (Cross-sit.)	3	2	1	1
Probability ≈ (Cross-sit.)	0.43	0.29	0.14	0.14

Table 1: Pursuit vs. cross-situational on Fig 2

rewarded during the **Pursuit** step. BALL is therefore chosen again after instance 2, confirmed and further rewarded. Instance 3 is similar to instance 2; BALL is chosen again, and the association between "ball" and BALL increases once more. After these three instances, the conditional probability $P(BALL|``ball'')$ will be very high, since the learner completely ignored the presence of ELEPHANT, DOG and BEAR during the second and third instances. This guarantees that the correct meaning BALL will be learned as the correct meaning. Space limitations prevent us from illustrating the behavior of the Pursuit model when the unambiguous scene is presented at the second and third instance. As the reader can readily verify, the probability of learning BALL will be lower; interestingly, that is also the findings of Medina et al. (2011), where high informative cues presented earlier in the learning sequence are more effective than presented later.

Contrast that performance with how the cross-situational model fares on the same instances. After instance 1, the probability of BALL will be high because that is the only meaning that has co-occurred with the target word. After instance 2, BALL receives a further boost, but crucially, DOG and ELEPHANT receive a boost as well. After instance 3, BALL and ELEPHANT receive a further boost, while the new meaning BEAR receives some share of the probability for "ball". The result is that ELEPHANT has been rewarded twice, DOG once, BEAR once, and the correct meaning BALL three times—the majority of the association boosts for "ball" have been incorrect meanings. In this simple one-word example, the Fazly et al. (2010) cross situational learning model will have meaning probabilities directly proportional to the number of association boosts for each meaning. Thus, after seven meaning tokens, the cross-situational model assigns approximately $\frac{4}{7}$ (more than half) of the total probability for "ball" to incorrect meanings. In order to succeed, Fazly et al.'s learner must have a very low threshold for words to be learned. However, to set the threshold this low has the unwanted effect of allowing many other weak hypotheses into the lexicon, as we shall see in our simulation results. Table 1 summarizes the differences between cross-situational and Pursuit learners on Fig 2. In the case of the cross-situational learner, the presence of ELEPHANT, DOG and BEAR, ignored by our learner, *dilutes* the probability space for "ball", making the correct meaning less likely to be added to the lexicon. The advantage of the Pursuit model over cross-situational models derives from its apparent sub-optimal design. The pursuit of the most favored hypothesis limits the range of competing meanings. But at the same time, it obviates the dilution of cues, especially the highly salient first scene in Fig 2, which is weakened by averaging with more ambiguous learning instances—but these are precisely the types of highly salient instances that the learner takes advantage of (Medina et al., 2011).

Computational Simulations and Results

Learning data

We manually coded two videos of mother-child interaction from the Rollins corpus on the CHILDES database, about 15 minutes' worth of data in total. For each of the 496 distinct utterances, we coded which concrete noun meanings were available to the learner (e.g., visible on the video and judged not to be outside the baby's visual field). The coding reflects children's prior assumptions, amply documented in the literature (see Markman, 1992, and references therein), that words map to basic categories of discrete and whole objects. Thus, for example, if we included the meaning BIRD in the interpretation of a scene, we did not also include BEAK or ANIMAL, in line with previous computational simulations of word learning (Frank et al., 2009; Yu & Ballard, 2007).[3]

The performance of all models was evaluated based on comparison to a single gold standard lexicon consisting of the set of word-meaning pairs that, in the judgment of the experimenters, could have reasonably been learned. The two criteria for inclusion in the gold standard were: (1) the word must refer to a concrete object (since only concrete object meanings were coded), and (2) the word must appear more than once in the data, and must refer to a meaning that is at some point visible to the child as judged from the video.

[3]The simulations of (Fazly et al., 2010) were carried on a much larger dataset but with artificially constructed sets of meanings; for the sake of comparison, we ran that model on the present set of Rollins data.

Simulations

We tested four models on the derived data set from CHILDES. First, we ran the Bayesian cross-situational model of Frank et al. (2009) on our dataset, with the code obtained from the first author of that paper. Second, we ran an implementation of Fazly et al. (2010), exactly following the description in their paper. Third, we implemented the the strict version of 'Propose-but-Verify', in which only a single hypothesized meaning is ever maintained in memory for a given word (Trueswell et al., 2013), along with a variant that probabilistically retrieves the conjectured hypothesis. And finally, we tested the Pursuit model. We tried a wide range of parameter values for all models to optimize performance. For the Fazly et al. (2010) and Pursuit models, the parameter values were optimized to two decimal places. For the Bayesian model, we used the best of a manageable number of simulations using a range of values for the learning rate parameter, which determines lexicon size. Further optimization was impractical, as each simulation of the Bayesian model can take several hours to run. The results reported below were obtained by running the models with best-case parameter values.

Models were evaluated based on precision, recall and the combined F-score (the harmonic mean of precision and recall) of the learned lexicon against the gold standard. Precision refers to the percentage of accurate word-meaning pairs, compared to the gold standard lexicon, that the model has learned, and recall refers to the percentage of all word-meaning pairs in the gold standard that have been learned by the model. Due to instances of random choice in the algorithm, Pursuit yields slightly different lexicons each time the model is run. Therefore, the results reported for the current model were obtained by averaging precision and recall over 500 simulations and using those averages to calculate F-score.

Results and Analysis

Table 2 compares the output of the three models. First, the low precision of the propose-but-verify model (Trueswell et al., 2013) underscores the robustness of probabilistic learning on realistic data. Second, our results replicate the basic pattern reported in Frank et al. (2009) that the Bayesian model outperforms the simplest non-Bayesian cross-situational model (Fazly et al., 2010; Yu & Ballard, 2007).[4] The Pursuit model outperforms all other models. Out of 500 simulations, only 8 produced a lexicon whose F-score was below the best result obtained by the Bayesian model.

	Precision	Recall	F1
Bayesian	**0.50**	0.29	0.37
FAS '10	0.28	0.21	0.24
Propose/Verify	0.04	0.31	0.08
P/V (prob. retrieval)	0.05	0.29	0.09
Current	0.44	**0.38**	**0.41**

TABLE 2: MODEL COMPARISON

We believe the advantage of the Pursuit model is revealed through the nature of the learning data the child learner faces during acquisition, which is more similar to the experimental conditions in Medina et al. (2011) and Trueswell et al. (2013), rather than the standard illustration of cross-situational learning in Fig 1. An analysis of the words successfully learned (and not learned) by the word learning models clearly illustrates this. Consider Fig 3, which shows the number of correctly learned words in the cross-situational model's lexicon vs. an average lexicon output by the Pursuit model, grouped by average ambiguity. We define the average ambiguity of a word as the average number of meanings in all utterances that contain that word; for instance, if a scene that accompanies a word contains only one noticeable object as judged by the human coder, then that occurrence of the word has ambiguity score of 1. Fig 3 shows that the Pursuit model does considerably better on words with lower ambiguity scores. For words with an ambiguity score greater than or equal to 3, both models learn rather conservatively, with the cross-situational model making only 3 correct mappings and the Pursuit model making 4 correct mappings. But when we look at those words that have an average ambiguity score less than three, we see that while the cross-situational model doesn't do much better on these words (4 correct mappings), our model learns more than twice the number of words (9 correct mappings). For each of the 6 words that were learned by the Pursuit model but not by the cross-situational model, there was at least one object which co-occurred with that word multiple times, but which was incorrect. This is exactly analogous to Fig 2, where the probability space for a word becomes diluted by competing hypotheses. While the averaging effect of cross-situational learning dilutes the few but relatively salient learning instances, the Pursuit model greatly benefits from their presence.

Conclusion

We implemented a model of word learning that combines insights from general considerations of probabilistic learning as well as experimental demonstrations of the word learning process. Cast in Marr's familiar level of analysis (1982), we have aimed to develop models at the algorithmic/representational level that could be related closely to behavioral studies of language learning. Future work will explore the predictions of the Pursuit model in an experimental setting, which may further refine the details of the model. The model also provides a general framework in which other cues

[4]It should be noted that, although we used the same videos as Frank et al. (2009), we report lower numbers for the Bayesian model even though we used the authors' original code. We believe this is due to the difference in data coding. Our coding is more inclusive as we encoded more meaning candidates in the video: more specifically, the average number of possible meanings per utterance in our coding is just under 3, but is 2 in Frank et al. (2009) (meaning our coding is 50% more ambiguous on average). Given the severity of the ambiguity in word-meaning associations (Gillette, Gleitman, Gleitman, & Lederer, 1999), a more inclusive coding procedure gives a better approximation of word learning in realistic situations.

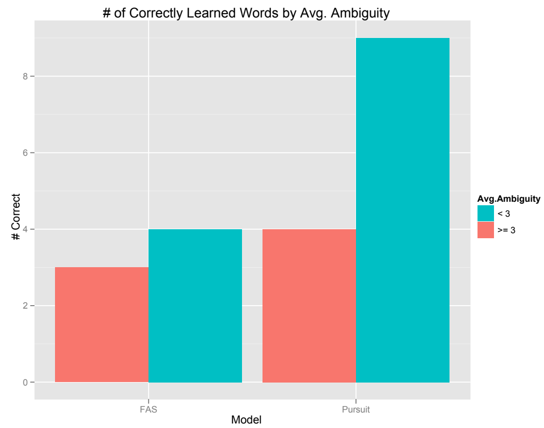

Figure 3: Comparison of correctly learned words

for word meaning can be incorporated; for instance, gestural information may further constrain the space of possible meanings, which can be straightforwardly implemented in our framework. Our model pursues the highly valued, and thus probabilistically defined, word meaning at the expense of other meaning candidates. By contrast, cross-situational models do not favor any one particular meaning, but rather tabulate statistics across learning instances to look for consistent co-occurrences. While the cross-situational approach seems optimally designed, simulation results show that its advantage is outweighed by dilution effects that distract the learner away from clear, unambiguous learning instances. It is notable that the apparently sub-optimal Pursuit model produces superior results over more powerful models with richer statistical information about words and their associated meanings: word learning is hard, but trying too hard may not help.

References

Barto, A., & Sutton, R. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.

Berwick, R. (1985). *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press.

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

Bush, R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, *68*, 313-323.

Chomsky, N. (1959). Review of verbal behavior. *Language*, *35*(1), 26-58.

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

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5).

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*(3), 407-454.

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

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

Landau, B., & Gleitman, L. R. (1988). *Language and experience: Evidence from the blind child*. Cambridge, MA: Harvard University Press.

Markman, E. (1992). Constraints on word learning: Speculations about their nature, origin, and domain specificity. In *Modularity and constraints on language and cognition: The minnesota symposium.* Mahwah, NJ: Erlbaum.

Markman, E., & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, *20*(2), 121-157.

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

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

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Siskind, J. (2000). Learning word-to-meaning mappings. In P. Broeder & J. Murre (Eds.), *Models of language acquisition.* Oxford: Oxford University Press.

Smith, K., Smith, A., & Blythe, R. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480-498.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558-1568.

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

Xu, F., & Tenenbaum, J. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2).

Yang, C. (2002). *Knowledge and learning in natural language*. Oxford: Oxford University Press.

Yang, C. (2004). Universal grammar, statistics, or both? *TRENDS in Cognitive Sciences*, *8*(10), 451-456.

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

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414-420.