The Active Role of Partial Knowledge in Cross-Situational Word Learning

Daniel Yurovsky, Damian Fricker, Chen Yu, and Linda B. Smith
{dyurovsk, dfricker, chenyu, smith4} @indiana.edu
Department of Psychological and Brain Science, and Cognitive Science Program
1101 East 10th Street Bloomington, IN 47405 USA

Abstract

A number of modern word learning theories posit statistical processes in which knowledge is accumulated across many exposures to a word and its potential referents. Accordingly, words do not go directly from unknown to known, but rather pass through intermediate stages of partial knowledge. This work presents empirical evidence for the existence of such partial knowledge, and further demonstrates its active driving role in cross-situational word learning. Subsequently, an incremental model which leverages its partial knowledge of word-object mappings from trial to trial is shown to account well for the data. In contrast, models which do not do so cannot explain the data. These results confirm crucial assumptions made by statistical word learning models and shed light on the representations underlying the acquisition of word meanings.

Keywords: word learning; language acquisition; computational modeling; statistical learning

Introduction

We have a tendency to characterize word learning as an all-or-none process: either a child knows a given word, or she has not yet learned it. This is apparent in our methodology (e.g. forced-choice tests, preferential looking), and assessment of vocabulary size via MCDI (Fenson, Dale, Reznick, Bates, Thal, & Pethick, 1994), as well as some theoretical claims. But this implicit all-or-none characterization may stymie our thinking about potential word-meaning representations.

For almost a century we have known that human learning and memory are not binary phenomena (Ebbinghaus, 1913). In learning lists of paired associates, for instance, a failure to recall the correct pair for a prompt does not imply no knowledge of the mapping. Evidence of this knowledge can be recovered using a different test paradigm (e.g. recognition or savings). The knowledge is not absent, but rather partial or sub-threshold. The central idea motivating this work is that such sub-threshold knowledge may play a profound in the course of language acquisition.

Several recent theoretical and computational approaches to word learning have made explicit use of partial knowledge. For instance, McMurray (2007) modeled the learning of a word's meaning as the acquisition of partial meaning tokens. Yu and Smith (2007) argued that early word learning can be thought of as the accumulation of co-occurrence statistics between words and objects across multiple situations. These theories suggest that a word can be learned in bits rather than in a single perfect moment.

Other models make an even stronger claim: not only can one build lexical knowledge by accumulating parts; this partial knowledge is an active driver of the learning system (Blythe, Smith, & Smith, in press, McMurray, Horst, Toscano, & Samuelson, in press, Fazly, Alishahi, & Stevenson, in press, Yu, 2008). These models have been tested predominantly on large corpora, reproducing qualitative patterns found in children's word learning. If they are correct about the presence and role of partial knowledge, however, then we should be able to find empirical evidence for the role of partial knowledge in human word learners.

Yurovsky and Yu (2008) presented indirect evidence of the active role of partial knowledge in cross-situational learning. They exposed participants to a series of individually ambiguous learning trials consisting of multiple words and multiple objects. At the end of each trial, participants were asked to indicate how sure they were (1-10) that they knew the correct label for each object. Yurovsky and Yu showed that a given object's rating could be predicted from the ratings given to the other objects on the same trial, even after the object's rating on its previous exposure was taken into account. Thus, participants seemed to be using partial knowledge of word-object mappings to reduce the set of candidates for other labels.

This analysis, while promising, was performed on participants' subjective knowledge ratings. In the present work, we propose to offer stronger and more direct evidence that partial knowledge plays an active role in word learning. To this end, we expose participants to two consecutive blocks of cross-situational learning. Crucially, half of the words and objects in the second block are those which participants failed to learn in the first block. Comparing the results of block 2 to those of several control conditions, we can determine the role of partial knowledge in cross-situational learning. First, we can ask whether partial knowledge exists in the system, whether learners are really accumulating bits of sub-threshold knowledge.

At a deeper level, we pursue a more interesting question: does partial knowledge of *individual* word-referent pairs – interacting in a *system* with partial knowledge of other word-referent pairs – facilitate the acquisition of new words. To answer this question, a set of computational models are fit to the data to understand the underlying learning mechanisms which give rise to the empirical results. We compare a *simple associative model*, a *biased associative model* which increments associations in proportion to their current strength, and a *competitive associative model* which adds within-trial competition. In the simple associative model, partial knowledge is not used in learning. In the biased associative model, partial knowledge of a word-

referent pair drives learning of that individual pair. Finally, in the competitive associative model, partial-knowledge of multiple word-referent pairs interacts and, by so doing, facilitates the learning of other pairs and thus the whole system of words and referents.

Experiment 1

To demonstrate the role of partial knowledge in word learning, we used the cross-situational word learning paradigm (Yu & Smith, 2007). In this task, participants are exposed to a series of individually ambiguous learning trials, each of which contains multiple co-occurring words and potential referents. While each trial is individually unambiguous, words always co-occur with their correct referent, and thus participants who correctly track co-occurrence between words and objects across trials can learn the correct pairings.

In Experiment 1, participants were exposed to two consecutive blocks of cross-situational word learning. At the end of block 1, participants were asked to select the correct referent for each of the trained words. For participants in the *unlearned* condition, half of the stimuli in block 2 were word-object pairs from block 1 for which they selected incorrect referents. For participants in the *new* condition, all stimuli in the second block were new.

If participants encoded nothing about words for which they selected incorrect referents in block 1, participants in the *unlearned* and *new* conditions should learn equally well in block 2. Alternatively, since no feedback is provided at test, if participants who selected incorrect referents did so as the result of binary hypotheses, and carried these wrong hypotheses to block 2, we might expect participants in the *unlearned* condition to underperform those in the *new* condition. However, if participants who selected incorrectly possess sub-threshold knowledge of the correct referent, we would expect participants in the *unlearned* condition to perform better than *new* participants in block 2. Most interesting would be if sub-threshold knowledge of one pair interacted with sub-threshold knowledge of other word-referent pairs to facilitate learning new pairs in block 2.

Method

Participants. Ninety-two Indiana University undergraduates participated in exchange for course credit; 50 in the *unlearned* condition and 42 in the *new* condition. However, to ensure a fair comparison across conditions, data from only a subset were analyzed (criteria explained in procedure). The final analysis was conducted on 23 participants in the *unlearned* condition, and 10 participants in the *new* condition.

Stimuli. Referents were represented by pictures of unusual objects which were easy to distinguish from each other, but difficult to name. Words were 1-2 syllable synthesized nonsense words constructed to be phonotactically probable in English. All words and objects have been used in previous cross-situational learning experiments (Yu & Smith, 2007, Yurovsky & Yu, 2008). Forty-two unique words and objects were used in total – 24 in block 1 and 18 in block 2.

Training slides for block 1 presented two pictures – one on each side of the screen – and played two labels, following Yu and Smith's (2007) 2x2 condition. Training slides for block 2 presented four objects – one in each corner of the screen – and played four labels, following Yu and Smith's (2007) 4x4 condition. Test slides for each block displayed all of the objects from that block (24 for block 1, 18 for block 2) in random positions and played one label.

Procedure. Each participant was exposed to two blocks of cross-situational learning – first a 2x2 block and then a 4x4 block. Each block consisted of a training phase followed by a test phase. The training phases consisted of a series of trials each displaying a set of objects and playing an equal number of words. Screen position and word order were randomized, such that they provided no information about which word labeled which object.

Following training, participants were given a series of alternative forced choice tests in which they were asked to select the correct referent for each label. Each word was tested once, and all objects from a block were presented on each test trial, so the content of test trials was uninformative as to correct mappings.

Block 1 contained 24 novel words, each of which occurred 5 times with its correct referent and less often with other objects. This resulted in 60 2x2 trials in total. Block 2 contained 18 words, each of which occurred with its correct referent 4 times and less often with other objects. This resulted in 18 4x4 trials. Word-object pairings and trial orders were selected randomly for each participant.

Block 1 was identical for participants in both the *new* and *unlearned* groups. The stimuli for block 2 differed across conditions. In the *new* condition, block 2 consisted entirely of novel stimuli – 18 words and their associated objects. For participants in the *unlearned* condition, however, 9 of the words and objects in the second block were those for which they had selected the incorrect response at test in block 1 (see Figure 1). Thus, for participants in the *unlearned* group, half of the stimuli in block 2 were words and objects for which they had not successfully learned correct associations. We will refer to the words and referents carried over from block 1 as old and those which are seen for the first time in block 2 as new.

Since participants could complete block 2 of the *unlearned* condition only if they had selected incorrect referents for at least 9 words in block 1, we could analyze participants who learned at most 15 of the 24 possible mappings. However, this could produce a skewed measure of average learning performance in block 2 since we would be rejecting data from those who learned "too much" in block 1. To help compensate, we also excluded participants who learned less than 9 correct pairings. Thus, only participants who learned between 9 and 15 correct pairings in block 1 continued on to block 2 of either condition.

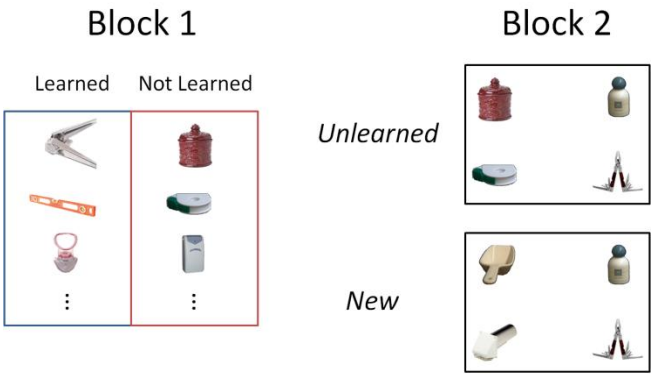

Figure 1: Selection of stimuli for block 2. In the *unlearned* condition, half of the items on each trial of block 2 were those for which the participant had given the incorrect response in block 1. The other half were new. In the *new* condition, all stimuli were new.

Results and Discussion

As described above, only a subset of the participants run in block 1 of either condition proceeded on to block 2.

Importantly, the proportion of words learned in block 1 did not differ between the selected subset and the set of all participants in the *unlearned* condition (M_s = .51, M_a = .50, t = .234, *n.s.*) nor in the *new* condition (M_s = .47, M_a= .43, t = .470, *n.s.*). Neither was there a significant difference between the proportion of words learned in block 1 by the selected participants in the *unlearned* vs. the *new* condition (M_u = .51, M_n= .47, t = 1.62, *n.s.*). This is to be expected given that block 1 was identical across conditions. Thus, all further analysis will be performed on selected participants.

In the second block, participants in both the *unlearned* and *new* conditions learned a significant proportion of word-object pairings (M_u = .56, t_u = 9.74, p < .001, M_n= .27, t_n= 6.62, p < .001, *chance* = .056). However, as shown in Figure 2, participants in the *unlearned* condition successfully mapped more than twice as many words to their correct referents as those in the *new* condition (t = 3.63, p = .01). Further, this benefit was not only for the 9 old pairings carried over from block 1 (M_u = .6, M_n = .26, t = 3.69, p < .001), but for the 9 new pairings as well (M_u = .51, M_n = .27, t = 2.48, p < .05).

Thus, partial knowledge of word-object pairings in block 1 allowed participants in the *unlearned* condition to learn significantly more mappings in block 2 than participants in the *new* condition. Further, the benefit was not just for the pairings for which participants had partial knowledge, but for novel pairings as well. This suggests that partial knowledge plays an active role in organizing cross-situational learning. Even though knowledge of word-object pairings was below threshold in block 1, it was sufficient to drive learning of novel pairings in block 2.

These findings provide initial support for the idea that sub-threshold knowledge of word-object mappings drives cross-situational learning. Partial knowledge of some pairs may influence the learning of other pairs on a trial-to-trial basis by constraining the pairs that are associated within a trial. An alternative explanation, however, is that participants in this experiment are benefitting from knowledge of which of the stimuli in block 2 had been seen previously in block 1. This could allow participants in block 2 of the *unlearned* condition to actively reduce the ambiguity of each training trial by mapping old words to old objects and new words to new objects. Some evidence for this second hypothesis comes from the errors made by participants in block 2 of the *unlearned* condition. When participants made errors in selecting referents for new words, they selected new referents at a probability significantly different from chance (M = .70, t = 3.73, p < .01, *chance* = .44). To provide further insights into the nature of the partial knowledge and its role in learning novel items, we constructed a new condition that was designed to assess the influence of sub-threshold mappings over and above possible knowledge of old/new.

Experiment 2

In Experiment 1 we tested the role of partial knowledge in word-referent mapping by exposing participants to two consecutive trials of cross-situational learning. Crucially, half of the pairings in block 2 were pairings for which participants failed to learn correct mappings in block 1. Learning results in block 2 showed that partial knowledge of these word-object pairings allowed participants to perform more than twice as well as participants exposed to a second block consisting of all new pairings. One possibility is that this benefit is entirely due to participants preferentially mapping old words to old objects and new words to new objects because they categorized them into two groups by mere exposure.

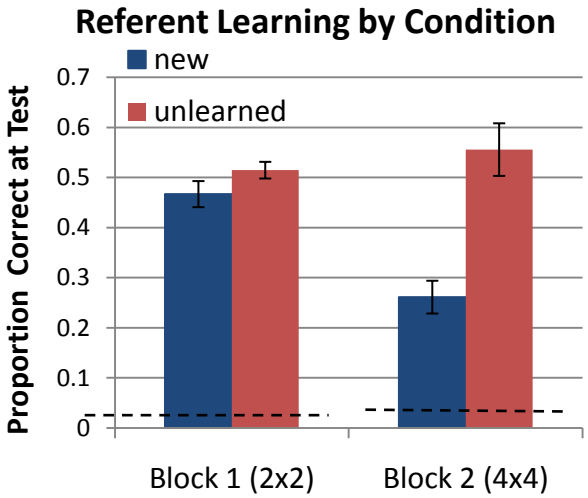

Figure 2: Proportion of word-referent pairings learned by participants in each condition across Blocks 1 and 2. Dotted lines indicate chance levels.

To establish a more stringent baseline for comparison, in Experiment 2 we constructed a *control* condition in which participants were exposed to the same word and object stimuli as participants in the *unlearned* condition, but without any opportunity to learn their associations. These same stimuli then appeared again in block 2. The *control* condition allows us to determine a second baseline – the effect of mere exposure to the stimuli of block 1.

Method

Participants. Ten Indiana University undergraduates participated in exchange for course credit. None had previously participated in Experiment 1.

Stimuli. Stimuli for Experiment 2 were identical to those for Experiment 1.

Procedure. The procedure for the *control* condition was similar to that used in the *unlearned* condition of Experiment 1. The crucial difference, however, was in the co-occurrence statistics of the words and objects of block 1. Whereas all words co-occurred with their correct referents 5 times in Experiment 1, in Experiment 2 half of the words occurred at most one time with each possible referent. Thus, there was essentially no correct referent for these 12 words. These unlearnable words and objects were matched for frequency of occurrence with those in block 1 – only co-occurrence statistics changed.

After the test phase of block 1, participants were exposed to a second cross-situational learning task as before. This time, however, 9 of the words and objects in block 2 were drawn randomly from the set of 12 unlearnable words and objects of block 1. In the second block these words each occurred 4 times with a single correct referent just like the 9 novel words. Thus, participants could distinguish the old words from the new words by their appearance in block 1, but they could not use potential partial knowledge of word-referent mappings to bootstrap their learning in block 2.

Results and Discussion

Because half of the words in block 1 of the *control* condition were unlearnable, it is unsurprising that these participants learned less words in block 1 than those in Experiment 1 (M_1 = .5, M_2 = .28, t_u = 7.21, p < .001). However, when only those words for which there was a correct answer in both Experiments are considered, participants performed equally well in both Experiment 1 and 2 (M_1 = .49, M_2 = .48, t = 0.19, *n.s.*). It is thus reasonable to compare block 2 performance across conditions.

In Experiment 2, we test the hypothesis that the benefit experienced due to participants in the *unlearned* condition of Experiment 1 was due not to partial knowledge, but to the ability to partition stimuli into two sets: old and new. If this is the case, mere exposure to the stimuli of block 1 – without the underlying co-occurrence statistics – should have been sufficient to reproduce this benefit. This is precisely the condition experienced by participants in the *control* condition. However, counter to this hypothesis, participants in the control condition did not outperform those in the *new* condition (M_c = .27, M_n = .26, t = .16, *n.s.*). They did, however, significantly underperform those in the unlearned condition (M_c = .27, M_u = .56, t = -3.22, p < .01). This difference was separately significant for old (M_c = .2, M_u = .6, t = -4.06, p < .001) and trending in the right direction for new (M_c = .34, M_u = .51, t = -1.55, p = .13) words. Figure 3 shows these results. This weighs against the hypothesis of mere exposure and lends credence to the hypothesis that partial knowledge is an active driver of cross-situational learning.

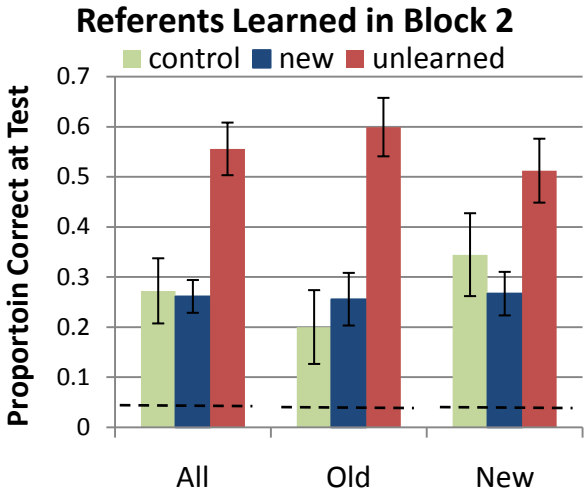

Figure 3: Proportion of word-referent pairings learned by participants in each condition. Old words are those which have been carried over to block 2 from block 1. In the *new* condition there are no old words, so the old words are those which fill the same slots in the training trials as the old words in the *unlearned* and *control* conditions. Dotted lines indicate chance.

Computational Models

To more fully analyze the role of partial knowledge of word-referent mappings in driving cross-situational learning, we implemented three incremental associative models that were exposed to simulated trials identical to those seen by experimental participants. The three models allow us to explicitly test hypotheses about how partial knowledge is used.

The first model – the *simple associative model* – maintains a word x object co-occurrence matrix and simply increments the cell corresponding to a word-object association each time the pair appears on a trial. This model thus learns the pure frequency of each of the possible word-object pairs.

The second model – the *biased associative model* – similarly maintains a word x object co-occurrence matrix. However, instead of incrementing the association strength

between a word and object by one whenever they co-occur, it increments their association by the strength of the current association. Whereas the *simple* model produces linear growth, the *biased* model produces geometric growth. This rich-get-richer scheme capitalizes on partial knowledge of a pairing in order to learn that pairing.

The final model – the *normalized associative model* – adds a competitive process to the *biased* model. On each trial, the increase in association between words and objects are computed as in the *biased* model, but the increment for a given word-object pair is normalized by the sum of all increments made for that object on that trial. This implements competition between all of the words in one trial. Intuitively, as one word accounts better for the presence of an object, the association between other words and that object are depressed. This mechanism is similar to the alignment mechanism used by Fazly et al.'s (in press) iterative version of the IBM Machine Translation Model (Brown, Pietra, Pietra, & Mercer, 1994).

The models are each tested for their knowledge of word-object associations in the same way as experimental participants. At the end of training, they are exposed to a series of alternative-forced choice tests and make their selections using the Shepard-Luce Choice Rule (Luce, 1959, Shepard, 1957). The simulated participant selects each alternative with a probability proportional to the exponential function of the strength of its association with the tested word.

Each model has only one parameter: a sensitivity parameter (λ) which weights each of the exponentiated probabilities in the Shepard-Luce Choice rule. Higher values of λ indicate that participants are more sensitive to differences in associative strengths between alternatives. To simulate Experiments 1 and 2, we exposed simulated participants to exactly the same stimuli as real participants. For instance, simulated participants in the *unlearned* condition were exposed to all of the training trials of the first block one at a time. Then, each simulated participant made selections at test using the Shepard-Luce Choice Rule. Nine of the items for which the model gave the wrong answer were then carried over to block 2, which were once again presented to the participant one trial at a time. Finally, the same decision rule was used to select a referent for each tested word. One thousand simulated participants were run in each of the three conditions using each model.

As can be seen in Figure 4, all of the models make essentially the same predictions for block 1. However, they make differing predictions for block 2 – the block during which partial knowledge may play a role. Figure 5 shows that the *simple associative* model is unable to produce the trend found in the data at even a qualitative level. It predicts that participants in the *unlearned* condition should underperform those in the *control* and *new* conditions. The other two models produce qualitatively similar trends. The *competitive* model, however, performs quantitatively better than the *biased* model (SSE_c = .0071, SSE_b = .0191, Bayes Factor = 2.69). As both have an equal number of

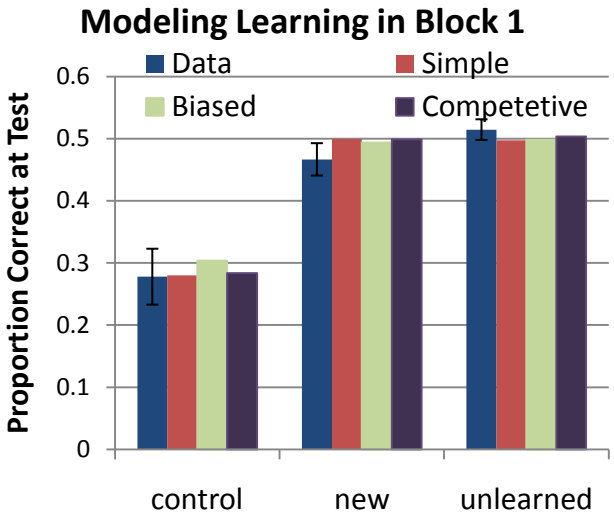

Figure 4: Proportion of word-referent pairs learned in block 1 by experimental participants and each of the three models across all experimental conditions.

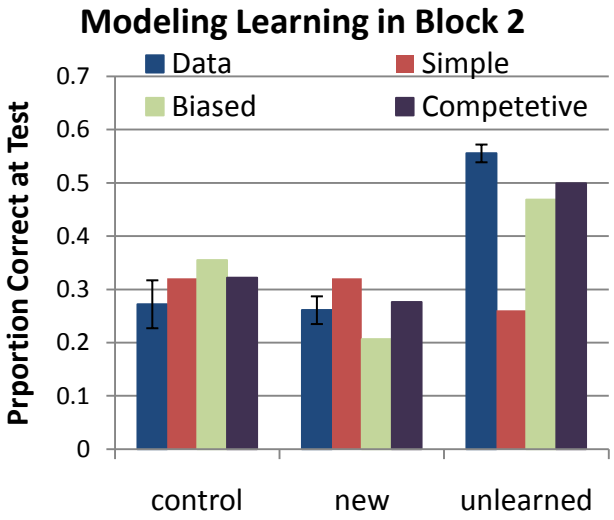

Figure 5: Proportion of word-referent pairs learned in Block 2 by experimental participants and each of the three models across all experimental conditions.

parameters, we can conclude that the *competitive model* is the better model for this empirical data. This supports the hypothesis that partial knowledge plays an active role in cross-situational learning, with partial-knowledge of multiple word-object associations interacting to support the acquisition of new word-object associations.

General Discussion

Whereas many methods for measuring word learning treat it as if it were binary – either the correct referent of a word is known or it is not – recent theoretical and computational models have argued that it is a gradual, accumulative process. Rather than learning a word's referent from a single perfect moment, learners may hone in on the correct referent through exposure to environmental statistics.

Empirical work has demonstrated that co-occurrence statistics alone are sufficient for learning word-object pairings (Yu & Smith, 2007). Furthermore, Vouloumanos (2008) showed evidence that learners are not only sensitive to the most frequently associated object for a given word, but also show deep knowledge of the statistical structure. Still, these results probed statistically acquired word-object knowledge only in its final state – producing a binary learned/unlearned data point for each potential pairing. Empirical evidence of graded states of partial knowledge has been indirect at best (Yurovsky & Yu, 2008).

The present work provides direct empirical evidence of not only the presence of such partial knowledge, but also its active role in driving word learning from exposure to exposure. The compared incremental models of statistical word learning show that partial knowledge may be leveraged on a trial-to-trial basis to bootstrap learning. Crucially, the better quantitative fits of the *competitive* model suggest that partial knowledge of a word-object association does not merely facilitate learning of that one association, but also combines with partial knowledge of other word-referent pairs to bootstrap learning of the whole system of words and referents. When words are learned as an interacting system, partial knowledge of one component gives a learner a leg up on acquiring others (Landuaer & Dumais, 1997).

While there is no denying the importance of word learning models at the computational level (Frank, Goodman, & Tenenbaum, 2009, Xu & Tenenbaum, 2007, Yu, 2008), this work again underscores our need to understand the continuous interaction of knowledge and learning on a moment-to-moment basis. Word learning is a constructive process, with initial successes cascading on themselves to empower even more successful learning (Smith, 1999). By digging deeper into word learning – understanding the latent representations that drive the system – we can hope to come to terms with its incredible complexity.

Acknowledgments

This work was supported by a National Science Foundation Graduate Research Fellowship to the first author and National Institute of Health Grant R01HD056029.

References

Brown, P. F., Pietra, S., Pietra, V., & Mercer, R. L. (1994). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics, 19*, 263–311.

Blythe, R. A., Smith, K., & Smith, A. D. M. (in press). Learning times for large lexicons through cross-situational learning. To appear in *Cognitive Science*.

Ebbinghaus, H. (1913). Memory. A Contribution to Experimental Psychology. New York: Teachers College, Columbia University.

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

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S.J.(1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59. Chicago: University of Chicago Press.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20*, 579-585.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Luce, R. D. (1959). *Individual choice behavior.* New York: Wiley.

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

McMurray, B., Horst, J., Toscano, J., & Samuelson, L. (in press). Towards an integration of connectionist learning and dynamical systems processing: case studies in speech and lexical development. In J. Spencer, M. Thomas, & J. McClelland (Eds.), *Toward a new grand theory of development? Connectionism and Dynamic Systems Theory reconsidered*.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22*, 325-345.

Smith, L. B. (2000). Learning how to learn words: An associative crane. In R.M. Golinkoff, K. Hirsh-Pasek, L. Bloom, L. Smith, A. Woodward, N. Akhtar, M. Tomasello, & G. Hollich (Eds.), *Becoming a word learner: A debate on lexical acquisition* (pp. 51-80). New York, NY: Oxford Press.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition, 107*, 729-742.

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

Yu, C. (2008). A Statistical Associative Account of Vocabulary Growth in Early Word Learning. *Language Learning and Acquisition, 4*, 32-62.

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

Yurovsky, D. & Yu, C. (2008). Mutual Exclusivity in Cross-Situational Statistical Learning. In B. C. Love, K. McRae, & V.M. Sloutsky (Eds.). *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 715-720). Austin, TX: Cognitive Science Society.