Prior Knowledge Bootstraps Cross-Situational Learning

Krystal A. Klein (krklein@indiana.edu)
Chen Yu (chenyu@indiana.edu)
Richard M. Shiffrin (shiffrin@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E. 10th St.
Bloomington, IN 47405 USA

Abstract

Recent research has highlighted the ability of adults as well as infants to learn word-to-world mappings over the course of statistical occurrences containing within-trial ambiguity. Participants in these studies are unfamiliar with all words and referents at the onset of these experiments, which differs from real-life word learning experience, in which learners are unlikely to enter a situation without any prior knowledge. Here we present a variation of the paradigm from Yu and Smith (2007) in which participants are taught a subset of the to-be-learned vocabulary prior to being exposed to the same training and test phases as in the initial experiment. Comparing results in conditions where participants had knowledge of some word-to-referent pairs prior to training with the data without prior knowledge shows the dramatic effects of prior knowledge in statistical word learning. Moreover, the behavioral data in this study provides valuable modeling constraints, as are demonstrated through simulations using several associative models of learning strategy.

Keywords: Language acquisition, word learning, prior knowledge, computational modeling.

Introduction

The problem of language acquisition is not a trivial one. Even considering words that have palpable referents, such as is the case of many of the nouns learned by infants, a child learning language frequently finds himself in a situation in which a stream of continuous speech is heard and a wide variety of potential referents are present, yielding virtually limitless possible pairings. Although to be sure, children acquire some subset of their early vocabulary through social constraints that allow fast mapping of words to referents, a recent hypothesis suggests that an alternative solution is available: rather than learning pairings between words and referents within the course of a single experience, a learner could accumulate co-occurrence statistics of words and referents across a series of temporally distinct situations and eventually use these accumulated statistics to determine the correct mapping between a word and its referent.

We will refer to the theory that people can track statistical co-occurrences of words and images and use this information to resolve ambiguity within individual learning contexts across a series of temporally distinct situations as cross-situational learning (CSL). CSL was proposed by Yu and Smith (2007), and is partially inspired by results obtained over the last decade indicating that both adults and infants can learn transitional probabilities within a stream of events within a single modality (Conway & Christansen, 2005; Newport & Aslin, 2004; Saffran, Aslin, & Newport, 1996); This line of experimentation began with Saffran, Aslin and Newport's (1996) work in solving the speech segmentation problem of language acquisition, which emerges due to the poor diagnosticity of characteristics of the waveforms created by natural speech in designating spoken word boundaries—in particular, the boundary between the end of one word and the beginning of the next. Their study demonstrated that very young infants, previously thought to be incapable of learning at such an age, learn the boundaries between words from a novel vocabulary pool presented in a continuous stream of speech by a mechanical voice containing no prosodic or temporal cues as to the location of boundaries. The authors have taken this as evidence that even 8-month-old infants possess powerful statistical learning capacities that are likely responsible for a great deal of early language learning: Subsequent studies by Saffran (2001) not only replicate the ability of infants to segment speech, but provide convincing evidence that these young subjects consider learned utterances as candidate "words" to be assigned meanings.

Yu and Smith (2007) have extended the studies of sequential statistical learning to cross-modal word-to-referent mappings: They demonstrated that the ambiguity between words and referents can be resolved by adults in a CSL task using cross-situational statistics. In these experiments, learners were able to accumulate statistical evidence of the possible word-referent pairings across multiple individually ambiguous learning trials occurring within a training period lasting less than six minutes (as demonstrated by their surprisingly good performance in a subsequent vocabulary test). Moreover, recent results from CSL studies with infants as young as 12 months have resulted in learning of novel word-to-referent pairs following a 4-minute training period (Smith & Yu, 2008).

In Yu and Smith (2007), speech stimuli were artificial words and visual stimuli were pictures of novel objects. The use of artificial vocabularies in gauging the plausibility of CSL as a robust mechanism for language acquisition is critical due to the control it provides of the knowledge possessed by learners prior to experiment participation. In the original CSL experiments, participants had no way of knowing any of the vocabulary (i.e. the associations

between words and referents) prior to the participation, because the pairings were generated at random for the purpose of the experiments.

In contrast, human development and learning is clearly a lifelong process, and learning in real life seldom occurs in the absence of any useful prior knowledge. Even for young infants—who may start word learning from scratch—surpass this stage quickly upon acquiring even a small vocabulary. Thereafter, they may use prior knowledge as a tool to speed up subsequent learning. In the context of cross-situational word learning, prior knowledge can be seamlessly integrated into the statistical learning mechanism and make the same statistical learning machinery more effective and efficient in at least two different ways.

First, prior knowledge can be directly used to reduce the degree of uncertainty in a learning trial. For example, in an ambiguous situation with three words and three referents {A, B, C, a, b, c}, there are 9 possible associations (A-a, A-b, A-c, B-a, B-b, B-c, C-a, C-b, C-c). However, if the learner somehow has already acquired the mapping A-a, he can use a mutual exclusivity constraint to reduce the search space to be {B, C, b, c}. This space includes only 4 possible associations – a significant reduction from the original situation. In some cases, such a reduction can lead to the elimination of mapping uncertainty altogether. For example, given two words {A, B} and two referents {a, b} in a trial, the learner may consider four possible associations. However, the learner with a prior knowledge A-a can unambiguously associate B with b based on the mutual exclusivity constraint.

Second, based on previous exposures, the learner may retrieve knowledge of having previously encountered a word, but maintain uncertainty regarding its meaning. This kind of partial lexical knowledge may also play an important role in subsequent learning: for instance, given a new situation wherein this previously exposed word (and several other words) co-occurs with several novel referents, the learner may simply exclude this previously exposed word from consideration as mapping onto the new referents on grounds that they have not co-occurred before. In this way, subsequent statistical learning can be simplified.

In brief, there is no doubt that prior knowledge is critical in human learning in general and in language learning in particular. Moreover, the above analyses suggest that prior knowledge can be incorporated easily in the cross-situational learning framework and has a potential to significantly improve learning performance. However, there are relative few studies exploring the role of prior knowledge in language acquisition (but also see the recent work by Lany, Gomez & Gerken, 2007, on syntactic learning). In light of this, the first goal of the current research is to document to what extent prior knowledge may aid, constrain, and bootstrap future learning. To achieve this, we designed and conducted an experiment in which participants learn a subset of vocabulary items (stimuli and their referents) before receiving the same statistical training and testing sequences used in Yu and Smith (2007): i.e. trials containing both the learned pairings and unlearned pairings. We explore the learning mechanisms through which such prior knowledge is utilized by developing and comparing several computational models.

Experiment

As in the CSL paradigm presented by Yu and Smith (2007), we asked participants to learn simultaneously many word-referent pairs from a sequence of highly ambiguous learning trials. On each trial, learners were presented with several words and objects without any information as to which referred to which. Although individual trials were ambiguous, words always co-occurred with their correct referents. The key manipulation here was to pre-train subjects to learn a small number of word-referent pairings before the regular training started. This procedure allows us to assess and explore the way participants utilize prior knowledge in acquiring new knowledge from statistically defined trial pairings.

Participants

213 students at Indiana University participated in the experiment in exchange for course credit.

Stimuli

Computer-generated pronounceable disyllabic nonwords were produced by a computerized voice. Referents were 100 x 100 pixel color images of uncommon objects. The stimuli were the same as those used in Yu & Smith (2007), representing three of the five conditions from that study. 45 unique objects and 45 unique pseudowords were divided into two sets of 18 object-word pairs, and one set of 9 object-word pairs. One 18-pair set was used to form a 3x3 condition wherein each trial consisted of 3 referents and 3 words. The other 18-word set formed a 4x4 condition wherein each trial consisted of 4 objects and 4 words. Finally, the 9-word set was used to form another 4x4 condition with 9 word-object pairs. Each correct word-object pairing occurred six times in the 3x3 and the 4x4 with 18 words condition, while every word-object pair occurred 8 times in the 4x4 with 9 words condition. There were 36 trials in the 3x3 condition, 27 trials in the 4x4 with 18 words condition, and 18 trials in the 4x4 with 9 words condition.

We selected a small number of words to be pre-trained. As shown in Table 1, two pre-training situations were created for each of three learning conditions. In the pre-training with low proportion, there were 3 words in each 18 set and 2 words in the 9-word set selected as pre-training items. In the pre-training with high proportion, there were 5 words in each 18-word set and 3 words in the 9-word set.

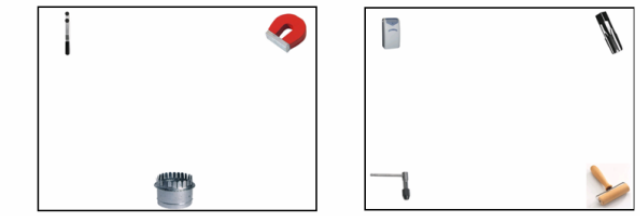

Figure 1: Screenshots from training phase. Left panel is from condition A; right panel is from condition B.

Table 1: Experimental Conditions

Training Condition (Vocab-ulary Size)	Trials	Occur-rences per word	Pre-training condition	Number Pre-trained (Proportion)
A: 3 x 3 (18)	36	6	PRE-HI	5 (0.28)
			PRE-LOW	3 (0.17)
B: 4 x 4 (18)	27	6	PRE-HI	5 (0.28)
			PRE-LOW	3 (0.17)
C: 4 x 4 (9)	18	8	PRE-HI	3 (0.33)
			PRE-LOW	2 (0.22)

Since those pre-trained words were selected from the learning sets, each of them co-occurred six times with some to-be-learned (TBL) words in the regular training. Thus participants could use their prior learning to reduce uncertainty about the pairings present on the regular trials, thereby bootstrapping the learning of the remaining pairings. To control and systematically evaluate the role of pre-trained words were selected in such a way that the possibility that two pre-trained words co-occurred in the same trial was minimized.

Design and Procedure

Proportion of pre-trained vocabulary was manipulated between-subjects; 50 participants learned a higher proportion of items during pre-training (PRE-HI), and 66 learned a lower proportion (PRE-LOW). Due to an error in the experiment administration program, the first block of the PRE-LOW group was unusable, requiring the condition to be repeated with new subjects. Thus, 97 more students were recruited to participate in the repaired condition.

Participants experienced three blocked learning conditions, the details of which are summarized in Table 1. Each block consisted of three consecutive phases. The first phase provided pre-training on a subset of that block's selected words; this phase represents the primary modification to the original paradigm reported by Yu & Smith (2007). During this phase, several objects were displayed on the computer screen, along with a button labeled "Ready for Test". Participants were told that they were to learn the correspondence between the objects displayed and the utterances that correspond with them, and that clicking on an object would cause its corresponding word to be played over the computer speakers. They were instructed to study these pre-training items freely until they were ready to be tested on them. During this test, words for the pre-trained items were played and participants had to select the correct referent from a set composed of the pre-training objects and five novel objects. These novel objects were not used in any other part of the experiment. If performance at test was not perfect, the pre-training screen was redisplayed and participants were instructed to study more before being retested.

The second phase was the same as the training phase in the corresponding condition of Yu & Smith: a series of training trials proceeded, during which three (in the case of learning condition A) or four (in the case of learning condition B and C) referents appeared on spatially separated areas of the screen (see Figure 1), and the corresponding utterances were presented auditorily in a random order (having no relation to the locations of the referents on the screen). Participants did not make any responses during the study phase; they were simply instructed that they would be trying to learn a set of word-referent correspondences that included the pre-trained items as well as other items.

The third phase was a four alternative multiple choice test; this was also modeled after the testing phase employed by Yu & Smith. During each trial, four objects appeared in the four corners of the screen and the sound corresponding to one of those objects was presented auditorily. Participants were instructed to select the object to which the sound corresponded. A response was required to advance to the next trial. Every vocabulary item was tested once; thus, there were 18 test trials in blocks A and B, and 9 in block C.

Table 2: Empirical Results

Training Condition (Vocabulary Size)	Pretraining condition	*N*	Performance (All Items)			Performance (Untrained Items)			Performance (Pretrained Items)		
			Median	Mean	SE	Median	Mean	SE	Median	Mean	SE
A: 3 x 3 (18)	PRE-HI	66	.944	.883	.018	.923	.851	.023	1.00	.967	.001
	PRE-LOW	97	.944	.900	.014	.933	.885	.017	1.00	.976	<.001
B: 4 x 4 (18)	PRE-HI	66	.833	.825	.017	.769	.776	.023	1.00	.952	.001
	PRE-LOW	50	.833	.759	.029	.800	.721	.033	1.00	.947	.001
C: 4 x 4 (9)	PRE-HI	66	1.00	.929	.016	1.00	.909	.023	1.00	.970	.001
	PRE-LOW	50	1.00	.900	.022	1.00	.877	.028	1.00	.980	.001

Results and Discussion

Results from all experimental conditions are displayed in Table 2. To provide a thorough characterization of the data, we have included performance over all test items, as well broken down results between pre-trained items and those that were to-be-learned during training (TBL). We have included means and standard errors, as well as medians for the data; median is included due to the significant skewing of the data toward perfect scores. There are several noteworthy observations in the results.

Pre-trained words. As can be seen in the rightmost portion of the table, participants performed very well on pre-trained items; in fact, the median performance was 100% for all six conditions. We have taken this to mean that very little forgetting of these items occurs over the course of training. This is critical to ensure prior knowledge is available through the whole training.

To-Be-Learned Words. Learning performance is significantly improved across all the learning conditions. In particular, we are interested in how well subjects acquired TBL words. This is illustrated in Figure 2 alongside the results from the original conditions (i.e. having no pre-trained items) reported by Yu & Smith (2007). On average, with two or three pre-trained word-referent pairs, subjects improved the learning of TBL words from 75% to 85% in the 3x3 condition, from 52% to 72% in the 4x4 with 18 words condition, and from 54% to 87% in the 4x4 with 9 words condition. There is no condition effect, suggesting both that subjects can robustly make use of prior knowledge in various learning conditions, and that the underlying learning mechanism of using prior knowledge may be similar in those training conditions.

PRE-LOW Vs. PRE-HI. The overall improvement from PRE-LOW (2 or 3 trained words) to PRE-HI (3 or 5 trained words) is limited at best. The potential improvement is of course limited by ceiling effects, Nonetheless, the lack of improvement with number of pre-trained items is a surprise. It is possible that the memorization of more words in pre-training requires additional computational and memory resources in the subsequent regular training, thereby hindering new learning.

Why is the role of prior knowledge in subsequent statistical learning so significant? This has to do with the nature of cross-situational learning. Compared with one-trial learning, cross-situational learning requires compiling and relating information extracted from multiple trials. There is high ambiguity about the true pairings that is gradually reduced as information accumulates across multiple trials.

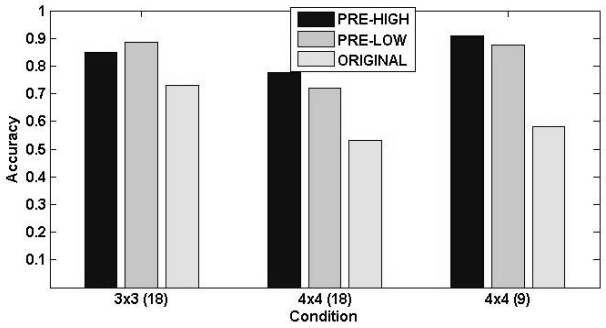

Figure 2: Results of experiment plotted alongside results from Yu & Smith (2007).

Prior knowledge can reduce much of the ambiguity, perhaps in the same way that using smaller numbers of pairs on each trial reduces ambiguity. More generally, whatever the causal mechanisms at play, these results point to the role of prior lexical knowledge in shaping, constraining, and improving subsequent word learning.

Modeling

To provide a formal account of the effects of prior knowledge in statistical learning, this section lays out and applies several computational models. The general principle shared between those models is conceptually simple – a Hebbian-like associative learning process continues accumulating statistical evidence trial by trial. The link of a word-referent pairing is strengthened when they co-occur in a trial but is weakened with disconfirming evidence (the word occurring without the referent). Based on this general principle, all the models also share the same representation of word-referent mappings. In all current model attempts, associations between sounds and pictures are represented by a V [vocalization] x V [referents] matrix (M) of weights, where V is equal to the vocabulary size. It is generally assumed that a small, variable residual association exists between any two given items prior to any previous exposure; this could be described as inherent associability of the two items at hand. However, beyond these two points, the models differ greatly in their implementation, as will be described.

Simple Association (SA) Models

General Description. The SA models are intended to represent a simple and quite probably automatic associative process that might occur if associations are learned through passive observation of the training trials. These model is based on one reported by Yu et al. (2007), with modifications made in order to represent prior knowledge accumulated during the pre-training phase. During each simulated training trial, α randomly-chosen associations are made. An association is exactly a strengthening of connection between a single referent and a single vocalization. We refer to this model as "simple" because

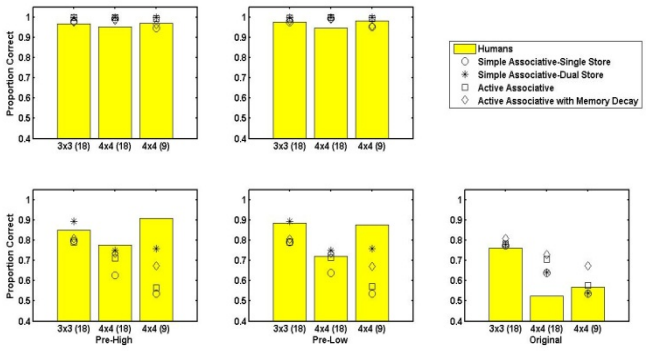

Figure 2: Fits of four models to empirical data. *First row:* Pre-trained words. *Second row:* To-be-learned items.

selection of mappings between pictures and sounds to be stored is made randomly from those presented during that particular trial; thus each association is independent of other associations, either occurring within the same trial or during other trials of the training phase. The act of making an association between picture i and sound j results in an increment of 1 associative unit in cell $M_{(i,j)}$ of the associative matrix.

We report here simulation results with two versions of the SA model, having two different methods of encoding and using prior knowledge.

SA-Single Store (SASS). The first simple associative model uses a single associative matrix for both pre-trained and to be learned (TBL) items. Associations between TBL pictures and sounds are initialized uniformly with a residual strength of 1 unit, while associations between pre-trained pictures and their associated pre-trained sounds are given an additional p units of strength, where p is a small integer free parameter of the model chosen to make it somewhat unlikely for pre-trained items to be forgotten, but so that it happens occasionally, as it does with human participants.

During test, the four cells corresponding to the test choices from the matrix row that designates the test word are "activated". The strongest association among four choices is selected a response; in the case of a tie, the answer is chosen at random from the tied items.

SA-Dual Store (SADS). The second SA model uses an associative matrix for TBL items and stores pre-trained items separately in a table that allows unambiguous lookup. Simulated learners are assumed to recall the pairings of pre-trained items confidently, and then implement a mutual exclusivity principle, thus reducing the ambiguity within trials containing pre-trained items.

During test, the presentation of a sound triggers first access to the lookup table. If the sound is contained in the lookup table, the picture is looked up and then selected, resulting in the correct response 100% of the time.[1] If the sound is not contained in the lookup table, a decision is made in the same was as in the SASS model. It should be noted, however, that there may be fewer than four activated cells, as the picture choices may include one or more pre-trained items, which are excluded from the associative matrix.

Simulations. The α parameter for both SA models was fixed at 3 because it was most successful in fitting the ORIGINAL conditions reported in Yu & Smith (2007). A was the only parameter to be chosen for the SADS model, but the SASS model required a t parameter to be chosen. Small-scale simulations indicated that a t value of 2 allowed the best approximation of the amount of forgetting of pre-trained items that was seen in the empirical study.
Figure 3 shows the mean performance of 1,000 simulated subjects in each of the four models described in this paper as different symbols superimposed over bar graphs of the empirical results. The SASS and SADS models are represented by the circles and asterisks, respectively. The two models perform about equally in the ORIGINAL condition (far right column), but their similarities end there. The SASS model predicts that the proportion correct of TBL items in the 4x4 9-words condition will be lower than in the 4x4 18-words
It is clear from these simulations that a "smarter" model is needed in order to account for the high performance in the pre-training conditions. For this reason, we built a new version of the associative model that uses previous-trial information to choose which associations to make.
Another characteristic of this model is to ignore the real-time nature of the audio stream (as opposed to the visual stream) within a single trial. In the case of the SA model, this can lead to the theoretically dubious situation of associating multiple referents with the last utterance heard, suggesting that no storage was occurring earlier in the trial. Although this is certainly possible, we suggest that participants are thinking about the utterances as they hear them and trying to associate them with references; thus, in subsequent versions of the model we assume that each utterance is sequentially heard and an attempt to store it with a referent is made momentarily at that time, with some probability of success. This has the additional feature of leading to variable number of associations stored per trial, which seems more likely than a fixed number, as participants are likely to find some combinations easier to store than others.

Active Association Models (AA)

Original AA model Initialization of memory in the AA model occurs in the same manner as in the SASS model.

[1] This version thus has no mechanism by which to "forget" pre-trained items; however, since forgetting occurred so rarely in the empirical results, it seems worthwhile to simulate the SADS model in spite of this drawback.

However, the selection of which associations are made during the training phase is more strategic. As each utterance is presented sequentially, a random variable determines whether an association to a picture will be made. With probability *c*, the current sound will be successfully associated to *some* picture—although not necessarily the correct one—during the trial (and thus, with probability *1-c*, it will not be associated with any picture during that trial). If an association is to be made, the corresponding picture for that word is then chosen from all available pictures on the screen according to a choice rule weighted by the current strengths of association of the various pictures to the target word. Selection of a referent R after hearing a test utterance U is stochastic and is made according to the choice rule

$$P(R = R_x \mid U_j) \sim \frac{M_{(x,j)}}{\sum_i M_{(i,j)}}$$

As in the SA models, making an association leads to an increment of 1 strength unit in the memory matrix *M*.

Memory Decay Variation We have also run a simulation using a slight variation of the AA model, which incorporates a decay of associative strengths over time. This "Active Association Memory Decay" (AAMD) is identical to the AA model, except that memory experiences some uniform decay after each training trial. The purpose of this is to cause more recent associations to contribute more significantly to the choice rule, making it more likely to associate the same referent with a sound that you associated with it in a very recent experience than if it was in a more distant trial.

Simulations The AA model provides a decent fit to human performance on pre-trained items; however, performance on TBL items underperforms human participants somewhat in the 4x4 (18) condition and dramatically in the 4x4 (9) conditions. Thus, by taking temporal proximity into account, the AAMD provides a qualitatively better fit. However, both models incorrectly predicted performance would be worse in the 4x4 9-word condition than in the 4x4 18-word condition, which requires further investigation.

Conclusion

The study reported here indicates that prior knowledge helps to bootstrap subsequent learning in cross-situational word learning. Participants' percentage performance on TBL items significantly exceeds performance levels in the ORIGINAL conditions reported by Yu and Smith.

Several initial modeling attempts have been made to provide a mechanistic explanation of the superior performance of adult learners. Simple associative (SA) models using an increase in strength for prior trained items greatly underperform in all PRE-HI and PRE-LOW conditions, suggesting a more strategic use of prior information in learning new pairs. The more active associative (AA) model can reach higher performance levels, but still fails to capture the very high performance in 9-item condition. The second-generation, modified versions of these models do slightly better: adding a temporal decay to the active association model reduced the error in the difficult-to-fit 9 item condition, and interestingly, adding the simple assumption that pretrained items are being kept separate from TBL items while random associations are being made (i.e. the SADS model) best captured the TBL performance of human participants. In summary, the present simulation efforts not only highlighted the intriguing observations in our behavioral results but also provided useful insights to future work.

Acknowledgments

This research was supported by National Institutes of Health Grant R01HD056029 and National Science Foundation Grant BCS 0544995. The authors extend thanks to Juliette McNamara and Anna Bern for help administering this study.

References

Conway, C.M., & Christiansen, M.H. (2005). Modality-constrained learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory & Cognition, 31,* 24-39.

Lany, J., Gomez, R.L., & Gerken, L. (2007). The role of prior experience in language acquisition.

Newport, E.L., & Aslin, R.N. (2004). Learning at a distance I. Statistical learning of nonadjacent dependencies. *Cognitive Psychology, 48,* 127-162.

Saffran, J.R. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition, 81,* 149-169.

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

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50,* 86-132.

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

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

Yu, C., Smith, L.B., Klein, K.A., & Shiffrin, R.M. (2007). Hypothesis Testing and Associative Learning in Cross-Situational Word Learning: Are They One and the Same? In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 64-70). Austin, TX: Cognitive Science Society.