Joint or Conditional Probability in Statistical Word Learning: Why decide?

Krystal Klein (krklein@indiana.edu) and Chen Yu (chenyu@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E. 10th St
Bloomington, IN 47401 USA

Abstract

Three experiments investigated the ability of human learners to concurrently extract and track both joint and conditional probabilities in statistical word learning. In each experiment, participants were briefly trained on novel word– novel object pairs and asked to learn correct mappings by the end of training. Across a series of learning conditions, we systematically manipulated conditional and joint probabilities individually and in combination to determine whether learners are able to encode multiple statistics in various learning contexts. Our results suggest that participants acquired both joint and conditional probabilities of word-referent co-occurrences. Based on the results from these experiments, we propose that learners are capable of utilizing the most reliable statistics that they acquired in training to make correct judgments in various testing tasks. These results suggest that statistical word learning is not only powerful but also adaptive.

Keywords: Language acquisition; word learning

Introduction

A recent trend in cognitive development is to study how human learners rely on statistical information to gradually reduce the uncertainty in the learning environment and ultimately acquire correct linguistic knowledge. The first study of this sort by Saffran, Aslin, Newport (1997) showed that 8-month-old infants can segment continuous speech into words based on statistical information alone. Subsequently, evidence has been mounting that both infant and adult learners can rely on sequential statistics to extract meaningful units from continuous sequences of stimuli in a variety of different sensory modalities (e.g., visual object sequence, action sequences, and tactile sequences) (Conway & Christiansen, 2005; Kirkham, Slemmer, & Johnson, 2002; Newport & Aslin, 2004; Saffran et al., 1999).

More recently, Yu & Smith (2007) and Smith & Yu (2008) have extended this line of research to word learning by integrating this idea with a well-established proposal – cross-situational learning (e.g. Gleitman, 1990; Siskind, 1996). By doing so, they proposed a new paradigm – cross-situational statistical learning. The basic idea is that a learner who is exposed to multiple words and multiple referents in a single learning experience need not solve the word-referent mapping problem in this moment; if the learner can instead accumulate co-occurrence statistics of words and referents across multiple temporally distinct learning situations, he or she can ultimately figure out the correct pairings from cross-situational statistics.

For example, in one condition reported by Yu and Smith (2007), participants were asked to learn 18 word-referent mappings over a series of learning trials. In each trial, learners viewed four pictures of objects and heard four names in an arbitrary order. Given 16 possible word-referent associations, learners could not have inferred correct pairings from individual trials. However, after being exposed to 27 individually ambiguous trials like this, participants in Yu & Smith (2007) acquired more than 9 out of 18 words. Further, a statistical associative model was developed to explain the underlying learning mechanism that supports cross-situational statistical learning. A set of simulation studies shows that this general associative account can serve as a fundamental framework to incorporate other cues/constraints, such as social cues (Yu & Ballard, 2007), syntactic cues (Yu, 2006), and prior knowledge (Klein, Yu, & Shiffrin, 2009), making statistical learning more efficient and more effective.

Although those recent empirical and computational studies have advanced our understanding of statistical learning in general and cross-situational word learning in particular, important questions regarding the mechanistic nature of cross-situational learning remain unanswered. For instance, what word-referent association information do learners track and compute across multiple learning situations? Do learners represent these statistics in terms of joint probabilities of co-occurring events or conditional probabilities? Do they retrieve statistical information they have accumulated in a probabilistic way or a deterministic way? In particular, if they have represented and accumulated statistical information in more than one form, which do they rely on in different contexts of knowledge retrieval?

There are two general probabilistic representations of co-occurring statistics in the context of statistical word learning: 1) joint probability of co-occurrence of words and objects p(w,o); 2) conditional probabilities of words given an object p(w|o) or conditional probabilities of objects given a word p(o|w). These two representations are not mutually exclusive, but rather, complementary: they reflect different aspects of co-occurring statistics. Joint probabilities represent overall frequencies of two co-occurring events; and conditional probabilities measure the frequency of one event (e.g. a word) given the presence of the other event (e.g. an object), which has predictive power when encoding sequential statistics. Moreover, one can convert between these two representations given knowledge of the statistical base rate of a single event p(w) or p(o):

□□/□=□□,□□□=□□,□□(□);

□(□,□)=□(□/□)□(□)=□(□/□)□(□)

The role of joint and conditional probabilities has been studied in both statistical speech segmentation (Saffran, Aslin & Newport, 1997) and visual statistical learning (Fiser & Aslin, 2001). For example, Fiser and Aslin demonstrated that learners rely on conditional probabilities of co-occurrences in learning statistical structures of visual scenes. In their study, they controlled the frequency (joint probability) of two sets of visual shape pairs and varied the conditional probabilities to distinguish these two groups. They found that learners are sensitive to conditional probabilities when joint probabilities have been equated.

The present study has three aims. First, we seek to extend previous studies on joint and conditional probabilities to statistical word learning. Learning co-occurrence statistics between words and referents differs from statistical sequential learning; it is thus unclear whether the results from previous studies will generalize to word learning. Second, having observed that most previous studies have focused on conditional probabilities (for example, Fiser and Aslin's 2001 study showed that given the same frequency, learners were sensitive to conditional probabilities), we aim to provide a more complete picture of the role of joint and conditional probabilities – individually or in combination – in statistical word learning. Thus, in addition to equating frequency and varying conditional probability, we equate conditional probabilities while varying joint probabilities, and even attempt to vary both such that conditional probabilities and joint probabilities may potentially compete with each other. Third, we will probe participants' statistical knowledge in different ways, to determine whether learners retrieve their acquired statistical knowledge differently according to retrieval situation.

Experiment 1

The aim of Experiment 1 was to determine whether the result obtained by Fiser and Aslin (2001) in their within-modality visual statistical learning paradigm would extend to the cross-modality mapping task inherent in cross-situational statistical word learning paradigms. We accomplished this by creating a single-factor experiment wherein a set of to-be-learned words was divided into two groups: each word in the first group occurred almost twice as often as each word in the second group, but some of these occurrences were paired with one referent, and other occurrences were paired with another referent. Each word in the second group always appeared with its correct referent; thus, the joint frequencies were equated across two groups, while conditional probabilities differed dramatically between the two conditions. Because we wanted to be sure of what co-occurrences were being stored by participants, we modified the standard cross-situational paradigm (wherein multiple words and multiple referents occur on a trial) to include only one word and one referent on each trial (see Vouloumanos, 2007).

Method

Participants 74 Indiana University undergraduate students participated in this study for course credit.

Design A single factor design was employed. Conditional probability of images given words was manipulated within subjects, such that half of the experimental words occurred three times with a single referent image (*object* heretoforth), and half of the words occurred three times with each of two objects, only one of which was a To-Be-Tested (TBT) object. In an attempt to disguise the design of the study from participants, each word also occurred one time with another random TBT object. A "correct" answer for a word at test was operationally defined as the object from the set of 12 TBT objects that most often occurred with that test word. Thus, during training, the conditional probability that an object from the high-conditional probability condition that appeared on a given training trial was accompanied by its correct sound was .857; similarly, conditional probability was .461 in the low-conditional probability condition.

Stimuli Word stimuli were 12 computer-generated pseudowords pronounced by a computerized voice. Referents were 18 169 x 169 pixel color images of uncommon objects. These images were also resized to 100 x 100 px for use in test, in order to comfortably fit all test items on the computer screen.

Procedure. The experiment consisted of a training phase, followed immediately by a test phase. In the training phase, a series of trials were displayed to participants, during each of which one object appeared in the center of the screen, and one word was presented auditorily. Each object appeared for 3 seconds. Participants made no responses during the training phase; they were instructed that they would be trying to learn a set of word-to-object correspondences over a series of trials. The order of the 66 training trials was randomized for each participant.

The test phase commenced immediately following training. There were 12 test trials, during each of which one of the trained words was presented auditorily and the twelve TBT objects appeared on the computer screen. Participants were instructed to select the object that corresponded to the test word using a computer mouse (chance = 0.08). A response was required to advance to the next trial. Each word was only tested one time, and the order of test items was randomized across participants.

Results

Percent of correct responses was tabulated for each subject in each condition, and then a within-subjects t-test was computed. Mean performance was .462 (SD=.265) for words with high conditional probability and .432 (SD=.232) for words with low conditional probability. The difference between these means was not significant, $t(73) = .907$, $p > .05$, 2-tailed.

Discussion

Although words in the low-conditional probability condition occurred as many times with a second, untested object as they did with their correct object, the degree of learning (as measured by percent correct) demonstrated by participants was not significantly less than in the high-conditional probability condition. The results of this experiment are inconsistent with the hypothesis that people are more likely to form a mapping between a picture and a sound when conditional probability of a referent is high than when it is low, controlling for number of joint exposures.

That participants did not perform better in the high conditional probability condition than the low conditional probability condition is especially surprising due to the presence of an equally strong (i.e. frequent) competitor object to the TBT object during training. If some learners had been using a mutual exclusivity assumption, for example, it would have been just as probable that they would have learned the untested object as the correct referent for the word as opposed to the TBT object. The fact that no difference emerged begs the hypothesis that participants do not "decide" upon a correct pairing until they are faced with a situation in which they have to decide.

Experiment 2

A second experiment was developed with two aims. First, we aimed to expand the question into a complete cross of joint frequency and conditional probability factors, so that it could be determined whether a main effect of joint frequency or an interaction effect between joint frequency and conditional probability would be evident. Second, we sought to compare high- and low- conditional probability conditions in a context where the correct answer had no strong competitor. Thus, rather than assigning words in the low conditional probability condition to two equally-frequently-occurring referents, we included multiple spurious correlations that only occurred one time each.

Method

Participants 45 Indiana University undergraduate students participated in this study for course credit.

Design A 2 (joint frequency) x 2 (conditional probability) within-subjects design was used. Each participant studied 4 words in each of the four permutations of these factors (LL, LH, HL, and HH). Three "noise" sounds and three "noise" pictures were also included during training in order to form the conditions of interest; however, these words and pictures were excluded from testing.

The high and low joint probability conditions included 6 correct co-occurrences and 3 correct co-occurrences, respectively. In the "high" conditional probability conditions, neither the word nor its correct picture occurred in the absence of the other; thus, conditional probability was 1.0 in both directions (from word to picture, and from picture to word). Each word in the "low" conditional probability condition occurred as many times with other pictures as it did with its correct referent. Likewise, its referent occurred as many times with other sounds as it did with the correct sound. Thus, the conditional probabilities were 0.5 in both directions. Both the words and the pictures in the low conditional probability condition were constrained such that they were never presented with any other item more than one time. Thus, there were no strong competitors in terms of joint probability.

Table 1: Design summary and results from Experiment 2

Cond	fr (w,o)	p (w | o)	fr (w)	M (SD)
LL	3	0.50	6	0.2444 (.216)
LH	3	1.0	3	0.4278 (.280)
HL	6	0.50	12	0.4556 (.234)
HH	6	1.0	6	0.4222 (.281)

Procedure The procedure was the same as in Experiment 1, except that there were 120 training trials and 16 test trials.

Results

Percent correct was tabulated for each participant in each condition. The results are plotted in Figure 1; means and standard deviations are displayed in Table 1. Performance in all conditions was significantly above chance (LL: t(44)=5.64, $p < .05$; LH: t(44)=8.75, $p < .05$; HL: t(44)=11.26, $p < .05$; HH: t(44)=9.05, $p < .05$). A 2x2 repeated measures ANOVA was calculated using SPSS 16.0. The results indicate significant main effects of both joint frequency [$F(1,44) = 9.522$, $p = .004$], conditional probability [$F(1,44) = 5.450$, $p = .024$], as well as a significant interaction of the two factors [$F(1,44) = 14.137$, $p < .001$].

Discussion

Unlike in Experiment 1, higher conditional probability led to more correct answers in Experiment 2, as indicated by a significant main effect of conditional probability. A significant main effect of frequency and an interaction effect between frequency and probability were also observed. Although performance in all conditions was above chance, performance was dramatically lower in the LL condition with respect to all the others. This suggests that as long as either joint frequency or joint probability was strong, learners could adaptively rely on that cue, learning nearly

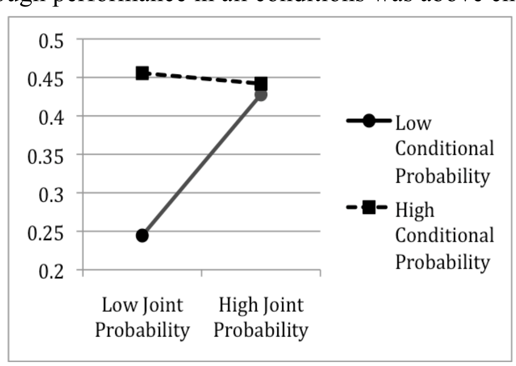

Figure 1: Experiment 2 results in percent correct as a function of condition.

half of the items in all those conditions within a single training session. In the absence of a strong cue (i.e., the LL condition), participants were only able to learn about 1 of 4 items. Conversely, having two strong cues (the HH condition) did not lead to better performance than in the other conditions; the observed mean was only the second highest—indeed, differences between the HL, LH and HH conditions did not approach significance. The reason for this cannot be determined from these data, but one may speculate that ceiling effects of computation came into play.

Experiment 3

The previous experiments compared the effects of joint frequency and conditional probability through independent observations, allowing for standard statistical analysis. In Experiment 3, we were interested in pitting different combinations of joint probability and conditional probability levels directly against one another. Thus, we gave each word two possible referents, and created a number of conditions in which these two candidate referents had various combinations of levels on the factors of interest. Then, we implemented 2-alternative forced choice testing between the two candidate referents during the testing phase, so that participants would be forced to select the referent that they most thought was designated by the word.

Method

Participants 14 Indiana University students and postdoctoral candidates participated in this study for course credit or $6 payment.

Design The experimental conditions were formulated from an incomplete cross of two factors of interest. A condition's level on the first factor refers to the co-occurrence frequency of a word and each of two possible referent pictures. The three possible levels are low-low frequency (LLF), low-high frequency (LHF), and high-high frequency (HHF). A condition's level on the second factor refers to the probability of an experimental word conditional upon each of its possible referents. The three possible levels of this factor are low-low probability (LLP), high-low probability (HLP), and high-high probability (HHP). Table 2 summarizes the seven conditions employed in this study.

Cond	fr(w,o1)	fr(w,o2)	p(w|o1)	p(w|o2)	Trials
HH-HL	6	6	1	.5	108
LH-HH	3	6	1	1	54
LH-HL	3	6	1	.5	90
LH-LL	3	6	.5	.5	108
LL-HL	3	3	1	.5	54

Table 2: Design summary for Experiment 3

Stimuli. Word stimuli were computer-generated pseudowords pronounced by a computerized voice. Referents were 169 x 169 pixel color images of uncommon objects.

Each of the seven blocks included six experimental words, with each word mapping to two distinct images. Additionally, some blocks contained 3-6 "noise" words, which occurred on at most one occasion with images in the block, according to the particular block's design (see Table 2). In total, there were 51 distinct pseudowords and 60 distinct pictures employed in the experiment; these were assigned to the blocks randomly within the numerical constraints inherent to the conditions.

Procedure. Participants experienced five blocked learning conditions; the order of blocks was randomized across subjects. Each block consisted of a training phase, followed immediately by a test phase. In the training phase, a series of training trials were displayed to participants, during each of which one image appeared in the center of the screen, and one name was presented auditorily. Each image appeared for 3 seconds. Participants made no responses during this phase; they were instructed that they would be trying to learn a set of name-referent correspondences over a series of trials, although sometimes the wrong name for a picture would be said. The number of trials in a given block depended on the composition of the corresponding experimental condition (see Table 2). The training trials were presented in a random order for each participant.

After each training phase, the 6 experimental words for the just-occurring block were tested using a two-alternative forced-choice method. During each trial, one of the experimental words was presented auditorily, and subsequently, the two exposed referents appeared on the left and right of the computer screen. Participants were instructed to select the referent to which the name corresponded. A response was required to advance to the next trial. Every word was tested once, but the order of test items was randomized across participants.

Results

Figure 2 plots the mean proportion of correct responses in each condition of Experiment 3, and means and standard deviations are reported in Table 3. 95% confidence intervals on the standard error of the mean were calculated in order to determine whether each condition differed from chance performance (0.50).

Three of the five conditions were significantly different from chance. In Condition 1, both referent objects for a sound had high frequency, but Choice 1 had $p(o \mid w) = 1$, and Choice 2 had $p(o| w) = 0.5$; participants reliably chose the high-conditional probability object more often than would be expected by chance, $p < .05$. In Condition 2, both referent objects for a sound had high conditional probability of $p(o \mid w) = 1$, but Choice 1 had low joint frequency and Choice 2 had high joint frequency; participants reliably chose the high joint-frequency object, $p < .05$. In Condition 5, both referent objects had low frequency, but Choice 1 had high conditional probability, and Choice 2 had low conditional probability; participants more frequently chose the high conditional-probability object, $p < .05$.

Two of the conditions were not reliably different from chance. Condition 3, in which Choice 1 had low joint frequency and high conditional probability, and Choice 2 had high joint frequency and low conditional probability, did not approach significance. This indicates that participants did not find either candidate object more viable than the other as the correct referent for the word. In Condition 4, both objects had low conditional probability, but Choice 1 had low joint frequency and Choice 2 had high joint frequency. Participants selected Choice 1 less frequently than half the time, and this approached significance, $t(13) = -1.975$, $p < .07$.

Table 3: Results from Experiment 3.

Condition	Joint Frequency; Conditional Probability	Mean (SD)
HH-HL	Choice 1 = Choice 2 (H); Choice 1 > Choice 2	0.6548 (.201)*
LH-HH	Choice 1 < Choice 2; Choice 1 = Choice 2 (H)	0.1548 (.190)*
LH-HL	Choice 1 < Choice 2 Choice 1 > Choice 2	0.5595 (.274)
LH-LL	Choice 1 < Choice 2 Choice 1 = Choice 2 (L)	0.3929 (.203)
LH-HL	Choice 1 = Choice 2 (L); Choice 1 > Choice 2	0.6429 (.205)*

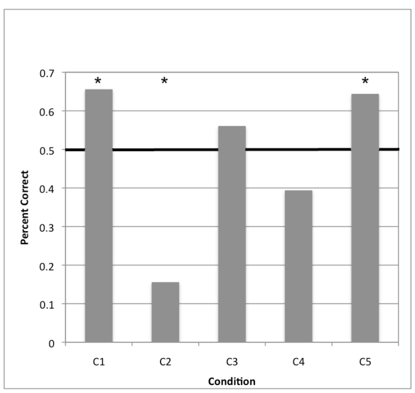

Figure 2: Percent correct in the five experimental conditions. Columns marked with an asterisk were significantly different from chance, $p < .05$.

Two paired samples t-tests were computed in order to compare independent effects of conditional and joint probability. In the first test, Condition 2 and Condition 4 were compared, because in both conditions, Choice 1 had low joint frequency and Choice 2 had high joint frequency, but both choices had high conditional probability in Condition 2, and both had low conditional probability in Condition 4. This difference was not significant, $t(13) = 3.552$, $p < .05$. In the second test, Condition 1 and Condition 5 were compared; in both of these conditions, Choice 1 had high conditional probability and Choice 2 had low conditional probability, but in Condition 1, both items had high frequency, and in Condition 5, both items had low frequency. This difference was not significant, $t(13) = .221$, $p > .05$.

Discussion

The three conditions that differed significantly from chance (with one more approaching significance) indicated that both joint frequency and conditional probability came into play: when conditional probability was equated, participants reliably chose the high-frequency object, and when frequencies were equated, participants chose the high-conditional probability object.

Experiment 3 differed from the previous two in that it directly pitted multiple referent objects for the same word against one another during a 2AFC test. This different methodology allowed us to observe which cue (conditional or joint probability) participants preferred to rely upon when forced to choose. The result is that conditional probability is preferred over joint probability.

General Discussion

A series of experiments in the present study attempted to understand the role of conditional and joint probabilities in the context of statistical word learning. Taking together the results across these experiments, several consistent and intriguing observations emerge. First, statistical word learners seem to be able to keep track of both joint and conditional probabilities. Direct evidence comes from Experiment 1 and 2, in which participants were able to learn several word-referent pairs with different conditional and joint probabilities in a single learning session. More specifically, in Experiment 2, they relied on joint probabilities in the testing of those pairs with high joint probabilities and meanwhile also learned the pairs with low joint probabilities but high conditional probabilities in the same training session. Further evidence is from Experiment 3, in which participants demonstrated the sensitivity to both joint probabilities (when condition probabilities were equated) and conditional probabilities (when joint probabilities were equated). Joint and conditional probabilities are complementary (e.g. one form cannot be derived from the other) in that they reflect different aspects of co-occurring statistics. Therefore, learning both joint and conditional probabilities concurrently enables human learners to extract and acquire more statistical regularities from the same learning input. More generally, the information contained in different probabilistic forms (e.g. conditional and joint probabilities in our case) may allow human learners to infer more complex knowledge by integrating this information to perform more complex inferences in the future. Indeed, recent work in a variety of fields suggests that considerable latent structure is derivable from the statistical analyses of large data sets (Landauer & Dumais, 1997; Steyvers & Tenenbaum, 2005). Our experiments provide direct support to those simulation studies in that we demonstrated that different kinds of statistics can be extracted from the learning environment and remain in memory. Moreover, this learning mechanism seems to be rapid, robust and effective.

Second, our experiments show not only that statistical word learning is robust and powerful, but also that statistical learning mechanisms may also be quite flexible and adaptive. In Experiment 2, participants adaptively relied on either conditional or joint probability knowledge, depending on which information is more reliable. In Experiment 3, when we equated one form of probability, they made their judgment based on the other. In addition, the fact that participants from two counterbalanced conditions in Experiment 2 – high joint probabilities with low conditional probabilities and high conditional probabilities with low joint probabilities – performed equally well also suggests that they may adaptively and automatically switch between these probabilistic representations depending on which is more reliable. Thus, in addition to extracting and storing different kinds of statistical information from the same training stimuli, statistical learners also know how to efficiently retrieve the information.

Third, despite the major findings from our research indicating both probabilistic representations are extracted in training for adaptive use at test, conditional probabilities seem to be more influential to the learning system than joint probabilities. In Experiment 3, when participants were forced to make a decision between items with high conditional probabilities but low joint probabilities, and items with low conditional probabilities but high joint probabilities, they relied more on conditional probabilities. However, participants were more sensitive to the changes in joint probabilities than the changes in conditional probabilities. We note that these observations may be unreliable partially due to the special parameters we used in our studies (e.g. conditional probabilities 1 or 0.5, and with 3 or 6 repetitions) and the limited number of subjects recruited. Therefore, we intend to vary these parameters in a future study will and test whether the same conclusion can be generalized to other situations.

In summary, we have shown that human observers can extract and keep track of both conditional and joint probabilities in various learning situations. They perform this statistical learning within a short, unsupervised training session. Moreover, statistical learners can apparently take advantage of acquired statistical information in a way that they always count on the more reliable information in knowledge retrieval. This work represents our first efforts on this topic and there are intriguing questions that are unanswered. For instance, are there any difference between two forms of conditional probabilities p(w|o) and p(o|w) in the context of word learning? Are human learners able to integrate different kinds of information to make a joint decision? Will the results reported here be generalized to other statistical learning tasks, such as statistical sequential learning and visual statistical learning? More studies will be needed to further document the role of both conditional and joint probabilities in word learning.

Acknowledgments

This research was supported by National Institute of Health Grant R01HD056029 and National Science Foundation Grant BCS 0544995.

References

Aslin, R.N., Saffran, J.R., & Newport, E.L. (1997). Computation of conditional probability statistics by 8-month old infants. *Psychological Science, 9,* 321-324.

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.

Fiser, J., & Aslin, R.N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12,* 499-504.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition, 1,* 3-55.

Kirkham, N.Z., Slemmer, J.A., & Johnson, S.P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition, 83,* B35-B42.

Klein, Yu, & Shiffrin (2009). Prior knowledge bootstraps cross-situational learning. Manuscript in preparation.

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.

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

Saffran, J.R., Johnson, E.K., Aslin, R.N., & Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*, 27-52.

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

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

Steyvers, M., & Tenenbaum, J.B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29,* 41-78.

Vouloumanos, A. (2007). Fine-grained sensitivity to statistical information in adult word learning. *Cognition, 107,* 729-742.

Yu, C. and Ballard, D. H. (2007) A Unified Model of Early Word Learning: Integrating Statistical and Social Cues. Neurocomputing, 70(13-15), 2149-2165.

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