Statistical Cross-Situational Learning in Adults and Infants

Chen Yu and Linda B. Smith
Indiana University

1. Introduction

We learn words in ambiguous contexts, with multiple word candidates for any referent and multiple referent candidates for any word. For example, a child may see a boy, a bat, a ball, and a dog and hear "Look at the boy. The dog wants his ball." This is the *word-to-world mapping* problem (e.g. Gleitman, 1990; Bloom, 2000; Smith, 2000). How could a learner who knows no words associate object names with the right referents? Developmentalists have studied a number of solutions to this problem, including ways in which the mature partner limits words and referents and directs attention to the relevant referent (Baldwin, 1993; Tomassalo, 2000), and internal perceptual and conceptual constraints (Genter, 1982). This paper is concerned with an alternative solution, cross-situational statistical learning, a process in which statistics are calculated across different learning instances to determine *across* multiple experiences, the most likely word-referent mappings. We are also interested in how internal constraints, such as whole object assumption or mutual exclusivity, may be realized or embedded in these mechanisms.

Prior research has concentrated on *in-the-moment* solutions to the mapping problem. For example, the mutual exclusivity constraint (Markman, 1990) is hypothesized to direct children to map novel words to unnamed referents. If

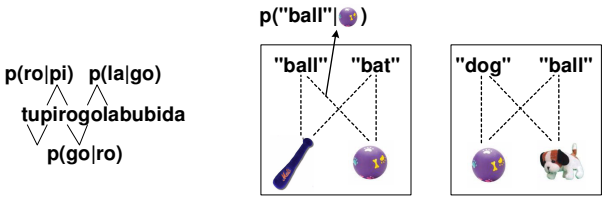

Figure 1: (a) Previous statistical learning studies focus on sequential statistics. Condition Probabilities of adjacent elements from the same stream of repeating elements are most often calculated (e.g. syllabus in speech segmentation studies or visual objects in visual perception studies). (b) Statistical word-to-world mappings involve a different kind of statistical learning. Conditional probabilities include co-occurring elements in a trial from two streams of data, indicating the probability that a word is associated with a referent.

there are two objects present and one has a known name, the child should map a novel name to the second object, solving the word-referent mapping problem in that moment. Does this kind of constraint also contribute, perhaps in a graded way, over multiple encounters with words and potential referents? Children could use broader statistical regularities, keeping track of the associations among many words and referents across trials, using these, and adjusting these, as they encounter potential words and referents. The idea that the learning system may effectively calculate broad cross-situational statistics is suggested by recent findings on statistical learning in infants (Saffran, Aslin, & Newport, 1996). Infants (and children, adults, and nonhuman primates) readily learn transitional probabilities among segments in a temporal stream of syllables, tones, or visual events (Saffran, Johnson, Aslin, & Newport, 1999; Hauser, Newport, & Aslin, 2001; Newport & Aslin, 2004; Conway & Christiansen, 2005). All these studies concerned sequential statistics in streams of repeating segments. Here we examine a different kind of statistical learning – the mapping of units between a word and a referent stream, as illustrated in Figure 1.

As a first step to study this kind of statistical learning, we chose to study adult language learners, asking whether they could compute such statistics over many potential words and referents and asking the nature of the mechanisms that underlie such learning. We present three experiments on both adults and young children examining the capacities and limits of this learning (Yu & Smith, 2007; Smith & Yu, 2008).

2. Statistical Learning in Adults

2.1 Experiment 1

The power of statistical learning to overcome the mapping problem rests on the calculation of cross-situational statistics -- not just tracking, for example, the co-occurrences of ``ball'' with ball or ``cup'' with cup but the co-occurrences of ``ball'' with a scene containing balls and dogs, balls alone, cups, cups and dogs, and so forth. Is this kind of computational mechanism at all feasible for humans? To answer this question, adult subjects were exposed to multiple trials wherein they heard multiple spoken words while looking at multiple pictures of objects. There is a perfect one-to-one mapping of words to referents such that each of the heard words maps to one of the objects. However, each trial consists of multiple words and multiples pictures of objects and there is no information within a trial about the associations between words and referents (including no spatial or temporal cues). We manipulated the degree of ambiguity of each learning trial, presenting in one condition 4 words and 4 possible referents on each trial (16 potential associations), 3 words and 3 possible referents on each trial (9 potential associations), or 2 words and 2 possible referents (4 possible associations).

Method

Participants. 38 undergraduate and graduate students at Indiana University were tested in the experiment. Subjects received course credits or $7 for their participation.

Stimuli. Subjects were exposed to three learning conditions, each of which included 18 novel word-object pairs. In total, stimuli consisted of 54 visual-audio pairs in three conditions. The potential words were generated from a computer program to sample broadly from the space of phonotactically probable English. These artificial words were then produced by a synthetic female voice, presented in a monotone. Fifty-four pictures of uncommon objects served as the visual input. The training trials were generated by pairing each word with a single picture. For each training trial, some number (depending on condition) of word-referent pairs were selected. Specifically, on each trial the referents were simultaneously presented on the screen. The names were then presented; however, the temporal order of the spoken names was not related in any systematic way to the spatial location of the referents. This is illustrated for a condition with 2 word-referent pairs and for a condition with 4 word-referent pairs in Figure 2. A 1000 ms silence was inserted between spoken words.

In total, there were three conditions determined by the number of words and referents presented on each trial: 2-2 (2 words and their corresponding referents), 3-3 (3 words and their corresponding referents), and 4-4 (4 words and

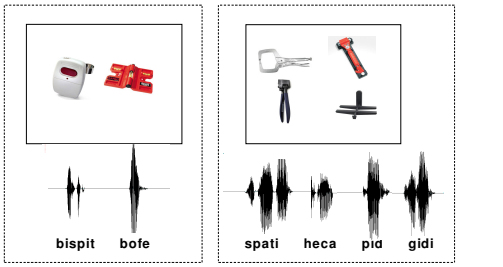

Figure 2: Subjects saw multiple pictures while hearing multiple words in each trial, and were asked to find which spoken word is paired with which

their corresponding referents). In each condition, there were 18 unique word-picture pairs, and each unique word and corresponding unique referent were presented on a total of 6 training trials. This means, as shown in Table 1, that the total number of trials (of two pairs, 3 pairs or 4 pairs) is different over the three conditions. In order to keep the total training time (summed over all trials) constant, we also varied, as shown in Table 1, the length of time of each trial.

Table 1 the statistics of the stimuli in 3 learning conditions.

	# of total pairs	# of occ. per pair	# of trial	time per trial (sec)	total time
2-2-18-6	18	6	54	6	324
3-3-18-6	18	6	36	9	324
4-4-18-6	18	6	27	12	324

Procedure. Visual stimuli were presented by 17 inch LCD flat panel screen and the sound was played by a pair of speakers connected to the same Windows PC. Subjects were instructed to map the pictures of objects showed on the computer screen onto the spoken words in a "nonsense" language. They were told that multiple words and pictures co-occurred on each individual trial and their task was to figure out which word went to which picture across multiple trials. Subjects were asked to participate in three sessions sequentially that corresponded to the three learning conditions. The order of sessions was counterbalanced. After training, subjects received a four-alternative forced-choice test. For each testing question, subjects heard one word and were asked to select the corresponding picture from four options on the computer screen.

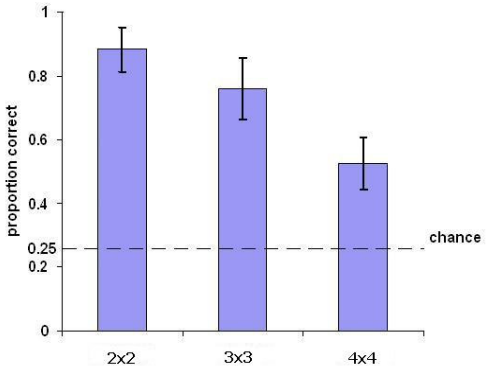

Figure 3: The results of three learning conditions in Experiment 1.

Results and Discussion

The three conditions present learners with different degrees of *in-trial* ambiguity but the same across trial (for a perfect statistical learner) certainty of word-referent pairings. The 4-4 condition –with four labels and four candidate referents for each learning moment (and thus 16 potential associations) – presents the greatest in-trial uncertainty, the 3-3 condition next, and the 2-2 condition least. As shown in Figure 3, in-trial uncertainty appears a relevant factor (ANOVA test, $F(2,74)=76.069$, $p<0.001$): Learners were better able to discover the correct word-referent in the 2-2 condition ($M=15.897$, $SD=2.506$)

and least able in the 4-4 condition (M=9.461, SD=2.907) with performance in the 3-3 condition falling in between (M=13.692, SD=3.507). However, the most important result is that in all conditions, including 4-4, subjects performed reliably above chance ($t(37)=8.785, p<0.001$, one-tailed, for 4-4). Given the in-trial ambiguity, they must be calculating statistics across trials.

There are a number of potential explanations of the differences among the three conditions, including the central variability of degree of in-trial uncertainty but also the additional necessary confounding of numbers of trials and length of trial. We investigate these factors in Experiment 2.

2.2 Experiment 2

Experiment 2 was designed to investigate under what circumstances, subjects would be able to achieve significantly better performance in the most ambiguous condition of Experiment 1, the 4-4 condition in which each trial offered 16 possible word-referent associations. We maintained across conditions, constant within-trial ambiguity but manipulated: (1) the number of occurrences of each pair to add more exposures to the whole training; and (2) the total number of word-referent pairs to be learned to make the learning task easier.

Method

Participants. 28 undergraduate students at Indiana University were tested in this experiment. None of them participated in Experiment 1. They received course credits for their participation.

Stimuli. The stimuli were selected from the same 54 word-object pairs used in Experiment 1. The three learning conditions are shown in Table 2, described by number of words-referent pairs presented on each trial (4-4), the number of total pairs to be learned (9 or 18), and the total number of occurrences of each unique word-referent pair across trials (8, 12, or 6). The time per trial was held constant for all conditions.

Table 2 the statistics of the stimuli in 3 learning conditions.

	# of total pairs	# of occ. per pair	# of trial	time per trial (sec)	total time
4-4-9-8	9	8	18	12	216
4-4-9-12	9	12	27	12	324
4-4-18-6	18	6	27	12	324

Procedure. The procedure is the same with that of Exp. 1.

Results and Discussion

There were 4 words and 4 pictures in a single trial in the three conditions, which contained a high degree of ambiguity at each individual moment (trial). However, the results in the three learning conditions consistently demonstrate that adult learners can still achieve a significant amount of correct pairings. As illustrated in Figure 4, they definitely acquire some lexical knowledge from

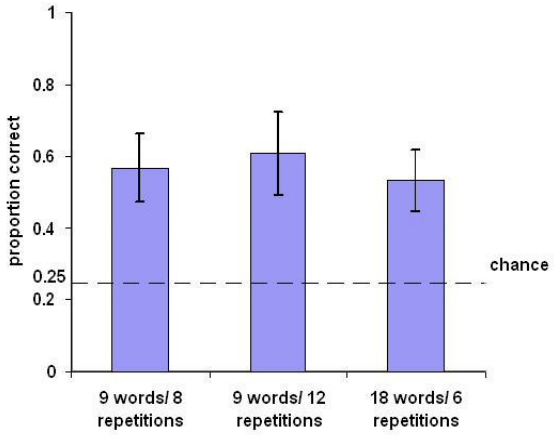

Figure 4: the results in three conditions in Experiment 2.The Y-axis shows the percentage of correct answers.

exposures. In addition, the results in 4-4-18-6 condition of this experiment (M=9.629,SD=3.076) are very similar to the same condition in Experiment 1, suggesting that our results are reliable and duplicable. The above two observations are quite in line with what we expected. However, as shown in Figure 4, there was no significant difference in these three conditions. A direct comparison between 4-4-18-6 (M=0.534,SD=0.232) and 4-4-9-12 (M=0.609, SD=0.176) conditions shows that they have the same number of trials, the same training time and the same within-trial ambiguity. The only two different factors are: (1) the number of unique pairs and (2) the number of occurrences per pair, which was expected to make one condition easier than the other. However, the results in these two conditions are quite similar. An intuitive explanation is that the number of co-occurring pairs plays a dominant role in statistical word learning and other factors are not so important conditioned on that factor. But why is that? Herein lies the power of cross-situational statistical learning: Even when the referent of a word cannot be unambiguously determined on any single learning trial, across multiple trials involving many different words and many different potential referents, the word and referent will occur most systematically than any other. The more words and referents there are to learn and that may co-occur together on any learning trial, the more discernible the systematicity –across trials – of the underlying correct mappings. Could bigger lexicons (more pairs) really be easier to learn than smaller ones? Learning requires multiple processes, some of which (for example, memory for particular items and attention) will be negatively impacted by the size of the learning set. However, within these constraints, statistical learning of a *system* of word-referent pairs may well benefit from larger as opposed to smaller data sets.

3. Statistical Learning in Infants

The above two adult experiment clearly showed that statistical cross-situational learning is within human repertoire, at least for adult learners. Our next question is to ask whether young children possess similar cognitive capabilities. In the following experiment, 12- and 14- month old infants were taught 6 word-referent pairs via a series of individually ambiguous trials. On each trial, two word forms and two potential referents were presented with no information about which word went with which referent. Although word-referent pairings were ambiguous within individual trials, they were certain across trials. For example, for a particular infant, whenever the form *tobi* occurred its assigned referent always occurred. After training, infants were presented with a single word and two potential referents, the cross-trial correct referent and a foil. Past research (e.g., Golinkoff, et al, 1997; Swingley & Aslin, 2000) shows that within this kind of preferential looking task, infants look longer at the labeled test object. Thus if infants have calculated the statistics appropriately, despite the uncertainty on individual learning trials, they should look longer at the correct referent of the word form.

Method

Participants. The participants, drawn from a working and middle-class population of a midwestern college town, were 28 12-month old infants (range -- 11 mo 17 days to 13 mo - 0 days; mean -- 12 mo 7 days; 13 males, 15 females) and 27 14-month old infants (range --14 mo 2 days to 15 mo 14 days; mean --14 mo 12 days; 14 males, 13 females). Two additional children began but did not finish the experiment.

Stimuli. The 6 "words" -- *bosa, gasser, manu, colat, kaki and regli* --followed the phonotactic probabilities of English and were recorded by a female speaker in isolation and were presented to infants over loudspeakers. The 6 "objects" were drawings of novel shapes; each was a unique bright color. On each trial, two objects (12 by 14 inches in projected size and separated on the screen by 30 inches) were simultaneously presented on a 47 by 60 inch white screen.

Procedure. Infants sat (on their mother's lap) 3.5 feet in front of screen with the mother's chair set at the center of the screen. Infants' direction of eye gaze was recorded from a camera centered at the base of the screen and pointed directly at the child's eyes. Parents were instructed to keep their own eyes shut through the entire procedure so as to not to influence their infant's behaviors. A camera directed on the parent through out the procedure confirmed their adherence.

There were 30 training slides. Each presented two objects on the screen for 4 sec; the onset of the slide was followed 500 msec later by the two words –each said once with a 500 msec pause between. Across trials, the temporal order of the words and spatial order of the objects were varied such that there was no

relation between temporal order of the words and the spatial position of the referents. Each correct word-object pair occurred 10 times. The two words and two objects appearing together on a slide (and creating the within trial ambiguities and possible spurious correlations) were randomly determined such that each object and each word co-occurred with every other word and every other object at least once across the 30 training trials. The first four training trials each began with the centered presentation of a Sesame street character (3 sec) to orient attention to the screen. After these first four trials, this attention grabbing slide was interspersed every 2 to 4 trials to maintain attention. The entire training – an effort to teach six word-referent pairs – lasted less than 4 minutes (30 training slides and 19 interspersed Sesame Street character slides).

There were 12 test trials, each 8 seconds. This duration was chosen from pilot studies to optimize the number of participants able to complete all 12 test comparisons (2 per target word). Each test trial presented one word, repeated 4 times with 2 objects – the target and a distracter – in view. The distracter was drawn from the training set. Each of the 6 words was tested twice. The distracter for each trial was randomly determined such that each object occurred twice as a distracter over the 12 test trials.

There were 2 unique sets of training slides with different orderings of objects, different mappings of words to the objects, and different combinations of word-referent pairs on the slides. For each set, the left-right locations of objects on the slides and the order with which the names were presented were randomly generated with the constraint that the object on the left was the target referent for the first presented word on half the trials and the target referent for the word presented second on the other half. There were also two unique test orders with unique randomly generated pairings of target and distracter, with the target appearing on the left on half the slides on the right on the other half. Half the infants at each age level were randomly assigned to each slide set.

Two coders naïve to condition and trial type coded direction of eye gaze from the video recorded from the camera directed at the infant's eyes. They coded, frame-by-frame, all frames from the start to the end (indicated by light on the video) of each training and test trial. The coder's task for each frame was to categorize the direction of look as right, left or away from the screen (hands, ceiling, mother's face, floor, etc). For reliability, the two coders each coded the same random sample of 25% of the frames. Agreement on these frames was 90.8%.

Results

Training trials. Infants were highly attentive to the training slides, looking (sum of right and left looks) at each 4 sec slide on average 3.27 sec (12 month olds) and 3.04 sec (14 month olds). On average, infants looked at the left and right sides of each training slide for equal durations ($t < 1.00$ for both 12- and 14-month olds). On 87% of all training slides, the infants looked at both sides (both objects) for at least 1 sec.

Test trials. On average, infants looked at each 8 sec test slides for a total of 5.6 sec for 12 month olds and 6.1 sec for 14 month olds. To examine whether infants *preferentially* looked in the direction of the target object, the object that *across trials* was associated with the auditorally presented label, each infant's looking time to target and distracter on each test trial was submitted to a 2(Age) by 2(Target/Distracter) × 6 (Word) ×2 (Block –first or second test of each target word) analysis of variance for a mixed design. The analysis revealed a highly reliable main effect of looking time to Target/Distracter, $F(1,54) = 35.32$, $\underline{p} < .001$(partial eta squared = .37). As shown in Figure 5, 12- and 14-month old children looked reliably longer to the Target than to the Distracter. The analysis also revealed a reliable interaction between Word and Target/Distracter, $F(5,54) = 3.85$, $p < .05$ (partial eta squared = .19). This result, that infants showed a greater difference in looking time to the target than distracter for some words than for others suggests that some word-picture correspondences were learned better than others. Finally, the analysis revealed an interaction between Age and Target/Distracter that approached significance, $F(5,54) = 3.13$, $p < .08$ (partial eta squared = .04). The older group of children, as can be seen in Figure 5, showed a bigger preference for the target than did the younger children, although the difference in looking times to target and distracter is individually reliable for both age groups (Tukey's hsd, $p < .05$). No other main effects or interactions approached significance.

Post-hoc analyses (Tukey's hsd, $p<.05$) conducted on the difference in looking time to target and distracter for the 6 individual words indicated reliably greater looking time to target than distracter for 4 of the 6 words for the 12 month old group and for 4 of the 6 words for the 14-month old age group. (Three of the individual words were that same at the two age levels, one was different; at neither age level were there reliable differences in the wrong direction for the remaining two words). Since half the children at each age level had different word-object pairings as well as different training orders, and since analyses for effects of slide set yielded no effects or interactions that approached significance, the source of these differences is not obvious. However, the fact that looking times for 4 of the 6 words (67% of the training set) show reliable preferences for the target does indicate that infants can figure out *multiple* word-referent mappings from a *system* of experienced associations. Finally, the group patterns appear to characterize the performance of individual infants in that 46 of the 55 participants infants looked, on average, at the targets more than distracters.

In sum, these results tell us that cross-situational statistical learning is in the repertoire of young word learners. Despite the ambiguity of word-referent mappings on any individual training trial, infants clearly accumulate information across trials and use that information to determine the underlying mappings. In less than four minutes, with six different word forms and six different objects,

infants learned enough to systematically look longer at the objects more strongly associated with the forms than those more weakly associated.

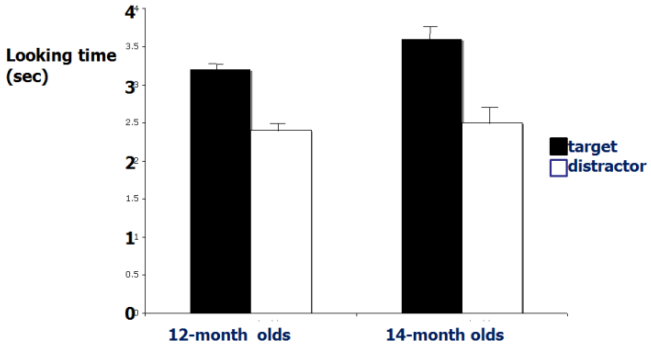

Figure 5. Mean looking time to target and distracter per 8 sec test trial (and standard error of the mean) for younger and older infants.

4. General Discussions

Learning situations such as those used in the present experiments have generally been considered too complex for word learning. Yet the present results show that both adults and young children rapidly discover word-referent mappings in these contexts. The only solution to the mapping problem is the distributional co-occurrence statistics between spoken words and pictures of objects. Our findings in statistical word learning extend those of Saffran, Aslin, & Newport (1996), and Newport and Aslin (2004) in word segmentation, Gomez & Gerken (1999) in syntax learning, and Conway & Christiansen, (2005) in visual and tactile sequence learning by showing that statistical learning broadly characterizes human learning, and that human learners can exploit cross-trial regularities over many potential word and referent pairs. Thus, our work extends Our experimental results showed that both infants and adults can use cross-situational statistics to infer correct word-referent pairings from ambiguous learning environments.

There are still several open questions in this new statistical learning paradigm. In particular, what are underlying mechanisms that support statistical cross-situational learning? Traditionally, two classes of cross-situational learning mechanisms have been considered. One is associative learning. Across trials, the learner could accrue associations between words and their potential referents by strengthening and weakening associative links between experiences of names and objects (see Plunkett, 1997, for review and discussion). Building on the "ball/bat" example above, the learner could on trial 1, equally associate "ball" with BALL and BAT. But after trial 1, and based on the experience of "ball" in the context of BALL and DOG, the association between "ball" and BALL

would be much stronger than that between "ball" and BAT. Over enough trials, these association strengths would converge on the real world statistics and yield the right word-referent pairs. The success of such a learning mechanism would seem to depend heavily on constraints on the kind of associations potentially formed given the logically infinite number of referents in any scene (e.g. Regier, 2003). An alternative way that learners could use cross-trial information is through hypothesis testing, by formulating and evaluating hypotheses about which names map to which referents (e.g., Siskind, 1996; Tennenbaum & Xu, 2000). Building on the "ball/bat" example above, the learner could wrongly hypothesize on the initial trial that "ball" refers to BAT but correct that hypothesis on trial 2 which presents disconfirming evidence. Across trials, the co-occurrence probabilities would support the "right" hypotheses for the language over others. These kinds of learning mechanisms also require constraints on the kinds of hypotheses that can be formed, given the infinite number of logically correct hypotheses true of any single datum.

Our ongoing empirical and computational studies intend to systematically investigate which mechanism is more plausible, what kinds of constraints are encoded in statistical cross-situational word learning mechanism and how prior knowledge is used in subsequent statistical learning. We are also interested in potential developmental changes in statistical learning. Moreover, we argue that social cues can be seamlessly integrated in this statistical learning paradigm and by doing so make the same statistical learning mechanism more effective. Overall, we suggest that statistical cross-situational learning may be fundamental to word learning and further studies on this topic have potentials to shed lights on the nature of human learning in general and on word learning in particular.

Acknowledgments: We thank Samantha Brandfon and Char Wozniak for collection of the data. This research was supported by National Institutes of Health R01 HD056029 and National Science Foundation Grant BCS0544995.

References

Baldwin, D. (1993). Early referential understanding: Infant's ability to recognize referential acts for what they are. *Developmental psychology* (29), 832-843.

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

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

Clark, E.V. (1987). The Principle of Contrast: a constraint on language acquisition. In B. MacWinney (Ed.), *Mechanisms of language acquisition* (pp. 1-33): Hillsdale, NJ: Lawrence Erlbaum Associates.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj II (Ed.), *Language development* (Vol. 2). Hillsdale, NJ: Erlbaum.

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

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1997). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of child language, 14*(1), 23-45.

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

Markman, E. M. (1990). Constraints Children Place on Word Learning. *Cognitive Science, 14*, 57-77.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*(2), 127-162.

Plunkett, K. (1997). Theories of early word learning. *Trends in Cognitive Sciences*, 1, 146-153.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926-1928.

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

Smith, L. B. (2000). How to learn words: An Associative Crane. In R. Golinkoff & K. Hirsh-Pasek (Eds.), *Breaking the word learning barrier* (pp. 51-80): Oxford: Oxford University Press.

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

Tenenbaum, J., & Xu, F. (2000). Word learning as Bayesian inference. In L. Gleitman & A. Joshi (Eds.), *Proceeding 22nd annual conference of cognitive science society* (p. 517-522). Mahwah.NJ: ErIBaum.

Tomasello, M. (2000). Perceiving intentions and learning words in the second year of life. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 111-128): Cambridge University Press.

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

Yu, C., Smith, L. B., Klein, K., Shiffrin, R.M. (2007). Hypothesis Testing and Associative Learning in Cross-Situational Word Learning: Are They One and the Same? In McNamara & Trafton (Eds.), *Proceeding 29nd annual conference of cognitive science society* (p 737-742). Mahwah.NJ: ErIBaum.