Does Statistical Word Learning Scale? It's a Matter of Perspective

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

Abstract

All computational models of word learning solve the problem of referential ambiguity by integrating information across naming events. This solution is supported by a wealth of empirical evidence from both adults and young children. However, these studies have recently been challenged by new data suggesting that human word learning mechanisms do not scale up to the ambiguity of real naming events. We replicate these experiments, collecting natural naming events both from a tripod-mounted camera and from a head-mounted camera that produced a "child's-eye" view. Although individual naming events were equally ambiguous from both views, significant learning across events occurred only from the child's own view. Thus, statistical word learning scales, but only from the right perspective.

Keywords: word learning; language acquisition; statistical learning; head camera

Introduction

Considerable evidence across domains suggests that infants are adept statistical learners, able to extract the regularities across many individually ambiguous learning instances (Fiser & Aslin, 2002; Saffran, 2003). One such domain in which statistical learning might play an important role in is learning word-referent mappings. Indeed, word-object co-occurrences across labeling instances is the core of every computational model of word learning, regardless of machinery (e.g. neural networks – Li, Farkas, & MacWhinney, 2004; Regier, 2005; statistical hypothesis testers – e.g. Blythe, Smith, & Smith, 2010; Frank, Goodman, & Tenenbaum, 2009; Siskind, 1996; machine translation – Fazly, Alishahi, & Stevenson, 2010; Yu, 2008). This hypothesis is also supported by empirical evidence showing that both adults (e.g. Yu & Smith, 2007; Smith, Smith, & Blythe, 2011) and young children (e.g. Scott & Fischer, in press; Yu & Smith, 2008) can learn word-object mappings by integrating information across ambiguous naming events.

But demonstrations of cross-situational learning in computational models and in simple laboratory experiments do not necessitate an important contribution of these mechanisms to real world word learning. Recently, Medina, Snedker, Trueswell, & Gleitman (in press) have argued that cross-situational learning – as demonstrated in the laboratory – will not scale up to the real world. In a world with many potential novel words, and highly cluttered scenes with many potential referents, learners may simply be unable to track the relevant co-occurrence statistics.

To make their point, Medina et al. followed four young children around their homes and recorded their natural parent-child interactions with a video camera. These videos were used to create short naming vignettes by segmenting out the interactions in which mothers said the names of common objects. The sound in these interactions was muted, and an artificial language label was inserted at the point when the English label was uttered. Following the human simulation paradigm (Gillette, et al., 1999), adult participants watched these vignettes, and guessed the meaning of each word. Even though they saw multiple naming instances for each word, adults were unable to integrate across them to learn new words. In contrast to the results of previous cross-situational word learning experiments (e.g. Yu & Smith, 2007; Smith, Smith, & Blythe, 2011), participants in Medina et al.'s (in press) experiment did not learn across situations. The authors therefore concluded that natural naming events are too ambiguous for cross-situational word learning mechanisms, and that real word learning must occur exclusively in *unambiguous* ostensive naming events.

Their conclusion is both a real possibility and fundamentally important, suggesting that – despite the models and laboratory evidence – word-referent learning is not a form of statistical learning. But should we accept Medina et al.'s conclusion? Here, we consider one limitation of their approach: the learner's perspective. Their participants were exposed to natural naming events, but they saw these events from an unnatural perspective. That is, their participants tried to learn words from the perspective of an adult observer *watching a child interact with a parent*, not from the first person view of a child learner.

Recent data collected from cameras placed on children's foreheads has recorded marked differences between adults' and children's views (Aslin, 2009; Franchak & Adolph, 2010; Smith, Yu, & Pereira, 2011; Yoshida & Smith, 2008). These differences could engage, or be optimal for, very different learning mechanisms (Gibson, 1969). They could also lead to different learning outcomes even from the same learning mechanism (Kuhl, 2004; Perry & Samuelson, 2011; Smith, 2000)

Accordingly, the studies in this paper replicate Medina et al.'s human simulation experiments, asking participants to learn words from ambiguous natural naming events. However, we recorded the same naming events from two perspectives: from a third-person perspective, and from the child's own first-person perspective (Figure 1). We show that cross-situational word learning is a matter of perspective: learning is possible from the child's own view.

Figure 1: Mother-child interactions were recorded from two views: a tripod-mounted camera (left), and a camera low on the child's forehead (right).

Experiment 1

The key theoretical idea behind cross-situational learning is that it enables learning from individually ambiguous learning events. Thus, we first needed to determine the difficulty of each naming event in order to determine the viability of learning across the most ambiguous events. In Experiment 1, naïve adult participants viewed the entire corpus of naming events. A beep was inserted into each naming vignette instead of an artificial language label, and participants were asked to guess which object the mother was naming. Participants could not learn by accumulating information across trials, giving us a measure of the individual ambiguity of each independent event.

Method

Participants. Four two to two-and-a-half year old children played with toys with their mothers while their interactions were recorded (mean age: 26;15, range: 25;12 – 27;10, 2 female). Each child was compensated with a small gift for their help. Twenty-eight Indiana University undergraduates participated in the human simulation experiment in exchange for course credit. Half viewed the vignettes from the first-person perspective, and half viewed the vignettes from the third-person perspective.

Stimuli. Children and their parents played with twenty-five toys chosen to broadly sample the kinds of toys with which young children are likely to play. These included animals, cars and trucks, colored rings, a telephone, and a baby doll (see Figure 1). Toys were arranged pseuodorandomly in a pile in the center of the room when the play session began.

Vignettes were created by sampling the portions of the mother-child videos corresponding to natural naming events. Each time a mother said the name of one of the toys in the room, a vignette was created spanning from three seconds before the label was uttered to two seconds after. The audio was muted and a beep was inserted into the vignette at the moment at which the real label was uttered. If, in the natural interaction, the name was uttered again in the 2 seconds post-naming, another beep was inserted into the video at this point and a further 2 seconds of silent interaction were appended to the vignette. The entire corpus consisted of 196 such vignettes.

Procedure. Children and their parents were told that we were interested in what toy play looked like from a child's perspective, and that we would be recording their interaction from a tripod-mounted camera as well as from a camera on the child's head. After the child put on the camera and a vest carrying the power supply, they played with the toys for 10 minutes. Their play was unconstrained.

Adults were informed that they would be watching videos of naming events from a set of mother-child interactions. They were told that the beep in each video corresponded to a moment in the real interaction when the mother labeled one the toys, and that they should try to guess the referent in each video. They were told that multiple beeps within a single video always corresponded to a single referent.

Adults watched each vignette in the corpus once, either from the first or the third-person perspective. At the end of each vignette, they were asked to type in the most likely referent in a free-response prompt. Each adult first watched three vignettes from a pilot experiment to ensure that they understood the task. They were prompted to make a guess on each vignette, and to describe the target as unambiguously as possible (e.g. "white stuffed animal") if they could not determine its exact identify (e.g. bunny, sheep). This was done to minimize misidentification errors which may have distorted differences between the two camera views.

We note briefly that these instructions differed from those used by Medina et al (2011), who did not inform participants that the target was an object nor accept guesses which did not exactly identify the target (e.g. "purse" was considered an incorrect guess for "bag.") Pilot data from a group of participants not told that words referred to objects contained a significant proportion of the guesses that were function words (e.g. "the"), pronouns (e.g. "it"), or onomatopoeias (e.g. "whoosh"). We felt that these answers were unlikely to be in the conceptual spaces of young children, and that our instructions increased the tenability of the human simulation hypothesis as a simulation of young learners (Gillette, 1999). Medina and colleagues informed us that such a change in instructions did not qualitatively change the pattern of data they observed (Medina, personal communication).

Results and Discussion

Since we are interested in the population of naming events rather than the populations of adults making guesses, following Medina et al. (in press) we aggregate guesses by vignette rather than guesses by participant. Overall, participants were generally successful at identifying the named target object in each video, with just over half of all guesses being correct. Guessing accuracy did not differ significantly across conditions, suggesting that average ambiguity is the same from each view (M_{1st} = .58, M_{3rd} = .58, $t(195)$ = .26, *n.s.*). However, ambiguity across vignettes was highly variable from both camera views.

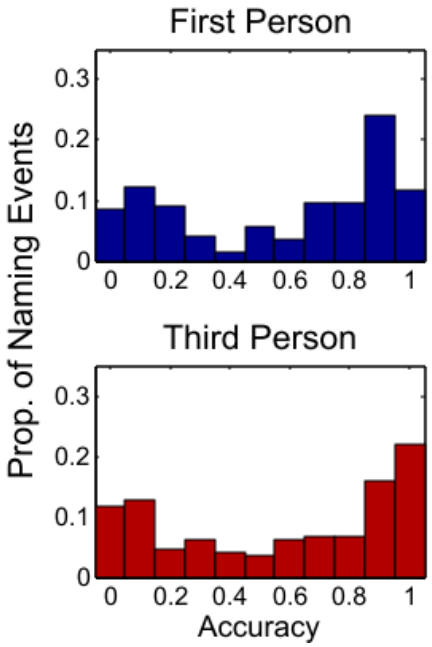

Figure 2: Naming event accuracy was highly bimodal from both views. Only the most ambiguous events, as measured in Experiment 1, were used in Experiment 2.

As shown in Figure 2, the distribution of difficulty across vignettes was bimodal from both views. Approximately half of naming the events were either highly ambiguous (≤10% accuracy) or highly unambiguous (≥90% accuracy). This ambiguity distribution suggests that while many naming events may be unambiguous, a large proportion are likely to be opaque to single-instance learning mechanisms. These ambiguous naming events are exactly the kind of input over which cross-situational learning mechanisms are hypothesized to operate. But can humans learn by integrating information across these difficult naming events?

Experiment 2

In Experiment 1, participants gave their best guess on each individual vignette, providing a measure of ambiguity for each individual naming event. In Experiment 2, participants were exposed to the most ambiguous vignettes from Experiment 1, but the beep in each was replaced by an artificial language label. Since participants saw multiple vignettes for each label, they could, in principle, learn by aggregating information across these instances If cross-situational word learning does not scale, as argued by Medina et al. (in press) then guess accuracy should not increase across vignettes. If, in contrast, inability to learn across ambiguous instances is peculiar to the third-person view, then participants who saw the same naming events from the child's perspective should show better guess accuracy after multiple vignettes.

Method

Participants. Forty-eight Indiana University undergraduate students participated in exchange for course credit. Twenty-four participants watched vignettes from the first-person perspective, and twenty-four from the third-person perspective. None previously participated in Experiment 1.

Stimuli. Stimuli for Experiment 2 were a subset of the naming event vignettes used in Experiment 1. However, instead of replacing each naming utterance with a beep as in Experiment 1, each naming utterance was replaced with an artificial language label. Labels were all produced by a female native speaker of English. Seven unique labels were produced, one for each of five toys that participants were asked to learn cross-situationally, and two that were used in example vignettes to acquaint participants with the task.

Vignettes were chosen for Experiment 2 were those for which participant guessing accuracy was low in Experiment 1. Four vignettes for each of five objects were chosen under the constraints that all could not come from the same parent-child interaction, no two vignettes for the same object could overlap in the original interaction, and participant guessing accuracy for each event must have been less than 30% in Experiment 1. Mean Experiment 1 guess accuracy for these vignettes is shown in Figure 3.

Procedure. Participants were again instructed that they would be watching vignettes constructed from natural parent-child interactions, and that each vignette would correspond to a moment in the real interaction at which the mother labeled one of the toys in the room. They were asked to guess the toy most likely to correspond to the artificial language label in each video, and that each individual label would always refer to the same toy.

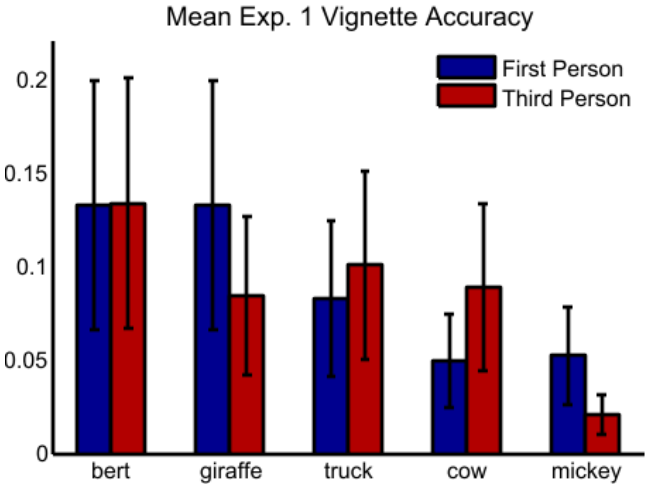

Figure 3: Vignettes for Experiment 2 were chosen to be comparably difficult across views. Solid bars show mean Experiment 1 guess accuracy for the four vignettes used for each object in Experiment 2. Error bars indicate +/- 1 standard error.

Results and Discussion

Figure 4 shows participant guess accuracy on each situation from both views. Accuracy on the first instance of each vignette was quite low, and not significantly different across views (M_{1st} = .12, M_{3rd} = .10, $t(46)$ = .34, *n.s.*). This validates the difficulty measure from Experiment 1, and verifies that participants saw highly ambiguous vignettes in Experiment 2. But, while accuracies were comparable across views on the first instance of each vignette, they rapidly diverged after more instances were encountered.

From the third person view, accuracy did not increase significantly from the first vignette to any of the successive vignettes (M_2 = .11, $t(23)$ = .46, *n.s.*; M_3 = .16, $t(23)$ = 1.43, *n.s.*; M_4 = .15, $t(23)$ = .97, *n.s.*). Further, guessing accuracy was uncorrelated with vignette number, indicating that participants were not learning across instances (r = .12, *n.s.*). Thus, Experiment 2 replicates Medina et al.'s (in press) results, showing no evidence of learning across ambiguous instances from the third-person perspective.

The results from the first-person view, however, tell a different story: accuracy increased marginally from the first to the second vignette (M = .22, $t(23)$ = 1.90, p = .07), and was significantly higher on the third (M = .25, $t(23)$ = 2.50, $p < .05$) and fourth vignettes (M = .26, $t(23)$ = 3.09, $p < .05$). Further, vignette number and guess accuracy were significantly correlated (r = .27, $p < .01$). Thus, guessing accuracy increased over exposure to multiple ambiguous instances, indicating that participants were integrating information across vignettes to learn the target of each label.

What explains these marked differences across views? One possibility is that differences in learning are explained by different accessibility of the children's own knowledge. If children knew the English labels spoken in each vignette, they may have shifted their gaze in response to labeling, and these gaze shifts could have been easier to access from the child's view. Since accuracy was comparable across views in Experiment 1, and for the first vignette for each label in Experiment 2, it is not likely to have been the main driver of learning differences, but it could nonetheless have contributed. To test this possibility, post-referential gaze shift behavior was recorded by a naïve coder for each of the 20 naming events participants saw Experiment 2. This coder was asked to identify the target of the first gaze shift after the beep in each vignette, or alternatively to indicate that no shift occurred. On 12 of the 20 naming events, children shifted their gaze in the 2-seconds after the label was heard. However, only 5 of these gaze-shifts were directed at the named object, and the remaining 7 were shifts to other toys in the room. Thus, post-referential gaze shifts were not generally a good source of information in these vignettes.

But differences in learning rates could nevertheless reflect use of this information. Each vignette was assigned one of three values: no gaze shift (0), shift to correct object (1), or shift to incorrect object (-1). Average shift information of the four vignettes for each label were then used to predict differences in final gaze accuracy across views, but were not found to be correlated (r = .01, *n.s.*). Thus, we can conclude

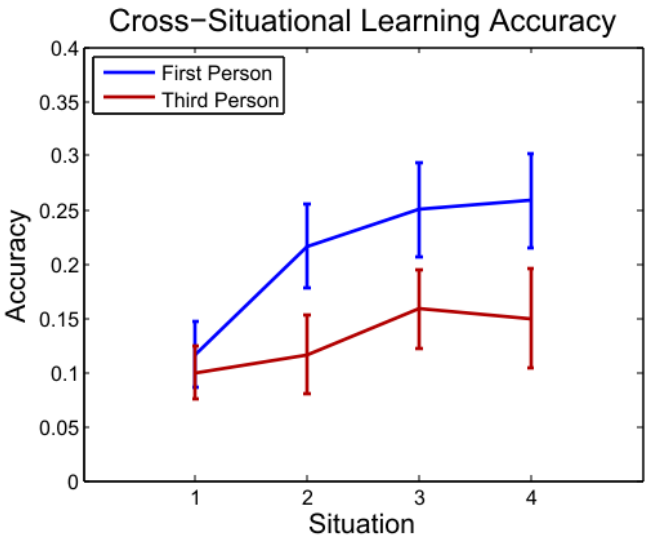

Figure 4: Naming event accuracy across instances from both views. Significant learning across instances was found only from the first-person view

that children's own knowledge embodied in the videos does not explain difference in learning from the two views.

Another possibility is that differences in learning from the two views can be explained by differences in memory. Because of the complexity of the third-person view, it is possible that participants were forgetting correct guesses and thus effectively being prevented from learning (Medina et al., in press). To test this possibility, we calculated the conditional probability of making a correct guess on the current vignette given that a correct guess was made on the previous vignette for the current word. For instance, if a participant correctly guessed that 'humbi' referred to the cow on the second vignette, would that participant give the same correct guess on the third vignette for 'humbi?' After six participants who made no correct guesses were excluded, conditional probability of making a correct guess did not differ significantly across views (M_{1st} = .43, M_{3rd} = .33, $t(39)$ = .79, *n.s.*), although the first-person view did show a slight advantage. Thus, differences in learning across the views do not seem to stem from differences in memorability of guesses.

A third possibility is that the different views gave rise to genuinely different learning strategies. In each vignette participants saw a number of different toys which could have been the target of the label. But, in contrast to standard cross-situational word learning experiment, each object was not equally salient (Yu & Smith, 2007; Smith, Smith, & Bythe, 2011). Each vignette presented a natural interaction with many kinds of cues to reference other than co-occurrence frequency, and the different views may have made this information differentially available. If these cues are highly salient, they may be given high weight even when they conflict with co-occurrence information. One way to measure this is examine how exposure to multiple vignettes for a label changed the dispersion of participants' guesses within a single vignette. If participants integrate information

across vignettes, then dispersion of guesses within a single vignette should increase. This is because participants will explore portions of possible object space cued not just by the properties of the current vignette, but also by co-occurrence information from previous vignettes. In contrast, if participants give high weight to other cues, co-occurrence information should have little effect on the guesses they entertain.

To measure dispersion, we use the entropy of the set of guesses made by all participants. Entropy integrates both the number of unique objects guessed and the relative frequency of each object. Entropy is maximized when many objects are chosen with equal frequency, and minimized when all participants guess the same object. If participants are changing their guessing strategy over exposure to multiple vignettes, then guess entropy should be higher in Experiment 2 than it was in Experiment 1. In contrast, if participants are guessing based only on the current vignette, perhaps having their attention divided by many potential cues to reference, entropy should be the same across the experiments. Guess entropy for the four vignettes for each word seen by participants in Experiment 2 were submitted to a 2 (Experiment) x 2 (View) mixed ANOVA. Results showed a significant main effect of view ($F(1,8) = 28.38$, $p < .001$) moderated by a significant interaction between view and experiment ($F(1,8) = 12.59$, $p < .01$). Follow-up tests showed that entropy did not differ between views in Experiment 1 ($t(8) = .72$, *n.s.*), and that entropy increased significantly from Experiment 1 to Experiment 2 for the first-person view ($t(4) = 13.71$, $p < .001$), but not the third-person view ($t(4) = .94$, *n.s.*). Thus, participants in the first-person condition changed their guessing strategy when they had multiple vignettes for each video, but participants in the third-person condition did not. Figure 5 shows guess entropies across views and experiments.

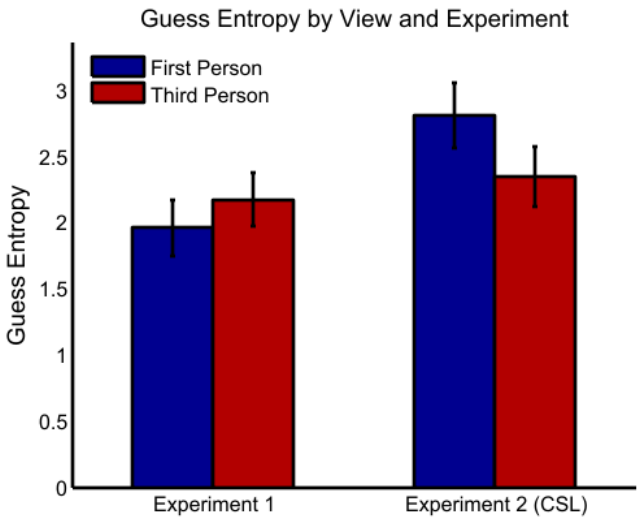

Figure 5: Entropy of participant guesses for vignettes from Experiments 1 and 2. There was no difference in dispersion for the third-person view, suggesting that participants were not using co-occurrence information.

General Discussion

Children could learn words only from unambiguous naming events. Alternatively, they could learn the meanings of words from a broader array of potentially noisy data, tracking the co-occurrence statistics of words and potential referents they encounter. Because children learn words so rapidly, acquiring more than 1300 words by 30 months of age (Mayor & Plunkett, 2010), many have argued that this learning cannot emerge from just unambiguous naming events, but must also reflect the integration of information from less informative events (e.g. Blythe, Smith, & Smith, 2011; Siskind, 1996; Yu & Smith, 2007). This claim has been supported empirically by evidence that adults and infants are sensitive to the statistics of co-occurrence between words and objects (Smith & Yu, 2008; Vouloumanos, 2008; Smith, Smith, & Blythe, 2011). However, whereas ambiguity in these experiments has been manipulated by presenting multiple isolated objects on the screen, the ambiguity of real world naming events may be of a totally different character. Medina et al. (in press) exposed participants to ambiguous natural naming events and found no learning over multiple naming events for the same label. Consequently, they argued that statistical learning is a laboratory phenomenon.

But participants in these experiments saw natural naming from an unnatural perspective: that of an adult observer. In this paper, we replicate Medina et al.'s (in press) results, but show that participants do integrate information over the same exact naming events when they are viewed from the perspective of the child to whom they are directed. Although more work must be done to determine the exact origins of this difference, one likely possibility is that it arises from differential access to other cues to reference provided by the two views. While the information in the child's view is not the same as the information available in standard cross-situational laboratory experiments, the child's view may be more like these laboratory tasks than it is like the third-person view. Thus, if we wish to study the ambiguity structure of natural naming events, we stress the importance of considering the visual information they actually provide.

None of this is intended to discount the importance of clear, unambiguous naming events. Mounting evidence, for instance, suggests that statistical speech segmentation is bootstrapped by exposure to isolated words (Brent & Siskind, 2001; Lew-Williams, Pelucchi, & Saffran, 2011). It is likely that word learning operates similarly, and that information from ambiguous events is integrated with unambiguous events, perhaps weighted in proportion to its uncertainty (e.g. Fazly; Alishahi, & Stevenson, 2010; Frank, Goodman, & Tenenbaum, 2009; Yu, 2008). We only wish to point out that the even ambiguous events contain structure, that humans are sensitive to this structure, and that this structure is likely to contribute to word learning. Since there are so many words to learn, children are likely to use all of the information they can get (Recchia & Jones, 2009).

Acknowledgments

We are grateful to Amanda Favata who helped design the head-camera apparatus and collect the mother-child interaction videos, to Amber Matthew and Becca Baker who helped collect the data, and to Collin Caudell for help coding the videos. We also thank Mike Frank, Dick Aslin, Hanako Yoshida, and all of the members of the Smith, Yu, and Shiffrin labs for comments and discussion. This work was supported by a National Science Foundation Graduate Research Fellowship to DY and National Institute of Health Grant R01HD056029 to CY.

References

Aslin, R. N. (2009). How infants view natural scenes gathered from a head-mounted camera. *Optometry and Vision Science*, *86*, 561-565.

Blythe, R. A., Smith, K., & Smith, A. D. M. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*, 620-642.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, B33-B44.

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

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Science*, *99*, 15822-15826.

Franchak, J. M. & Adolph, K. E. (2010). Visually guided navigation: Head-mounted eye-tracking of natural locomotion in children and adults. *Vision Research*, *50*, 2766-2774.

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

Gibson, E. J. (1969). *Principles of perceptual learning and development*. Englewood Cliffs, NJ: Prentice Hall.

Gilette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*, 145-176.

Kuhl, P. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*, *831-843*.

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, *14*, 1323-1329.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, *17*, 1345-1362.

Mayor, J., & Plunkett, K. (2011). A statistical estimate of vocabulary size from CDI analysis. *Developmental Science*, *14*, 769-785.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (in press). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*.

Perry, L. K., & Samuelson, L. K. (2011). The shape of vocabulary predicts the shape of the bias. *Frontiers in Developmental Psychology*, *2*, 1-12.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*, 647-656.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *6*, 819-865.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*, 110-114.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to meaning mappings. *Cognition*, *61*, 1-38.

Scott, R. M., & Fischer, C. (in press). 2.5-year-olds use cross-situational consistency to learn verbs under referential uncertainty. *Cognition*.

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

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

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

Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother's view: The dynamics of toddler visual experience. *Developmental Science*, *14*, 9-17.

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

Yoshida, H. & Smith, L. B. (2008) What's in view for toddlers? Using a head camera to study visual experience. *Infancy*, *13*, 229-248.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Acquisition*, *4*, 32-62.

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