Balancing looks to people and to objects during novel word learning

Kyle MacDonald

Committee: Virginia Marchman, Hyowon Gweon, Jay McClleland, & Michael C. Frank

Dissertation overview

Learning words should be hard. Consider that even concrete nouns are often produced in complex contexts with multiple possible referents, which in turn have many conceptually natural properties that a speaker could talk about. This ambiguity creates the potential for an (in principle) unlimited amount of referential uncertainty in the learning task.[1]. Moreover, to find meaning in language requires rapdily establishing reference during real-time interaction where the incoming information is dynamic, multimodal, and transient. Remarkably, children's word learning proceeds despite these challenges, with estimates of adult vocabularies ranging between 50,000 to 100,000 distinct lexical concepts (Bloom 2002). How do learners infer and retain such a large variety of word meanings from data with this kind of ambiguity?

Statistical learning theories offer a solution to this problem by aggregating cross-situational statistics across labeling events to identify underlying word meanings (Yu and Smith 2007; Siskind 1996). Experimental work has shown that both adults and young infants can use word-object co-occurrence statistics to learn words from individually ambiguous naming events (Smith and Yu 2008; Vouloumanos 2008). For example, Smith and Yu (2008) taught 12-month-olds three novel words simply by repeating consistent novel word-object pairings across 10 ambiguous exposure trials. Moreover, computational models suggest that cross-situational learning can scale up to learn adult-sized lexicons, even under conditions of considerable referential uncertainty (Smith, Smith, and Blythe 2011).

While all cross-situational learning models agree that the input is the co-occurrence between words and objects and the output is stable word-object mappings, they disagree about several key points. First, alternative models propose different underlying representaions that support long-term retention of word-object labels. One approach models learning as a process of updating connection strengths between multiple word-object links with the underlying representation being a distributed word-object co-occurrence matrix (McMurray, Horst, and Samuelson 2012). Another approach argues that learners store a single word-object hypothesis, only switching to a new hypothesized link when there is sufficient negative evidence (Trueswell et al. 2013).

In addition to the debate about representation, researchers disagree about the amount of ambiguity in

[1]This problem is a simplified version of Quine's *indeterminacy of reference* (Quine 1960): That there are many possible meanings for a word ("Gavigai") that include the referent ("Rabbit") in their extension, e.g., "white," "rabbit," "dinner." Quine's broader philosophical point was that different meanings ("rabbit" and "undetached rabbit parts") could actually be extensionally identical and thus impossible to tease apart.

the input to cross-situational learning mechanisms. Some studies have shown that the majority (90%) of naming events are ambiguous (Medina et al. 2011), while other work has found a higher proportion of clear naming events (Yurovsky, Smith, and Yu 2013). Moreover, Cartmill et al. (2013) showed that the proportion of unambiguous naming episodes varies across different parent-child dyads, with some parents rarely providing highly informative contexts and others' doing so more often. The key point is that variability in referential uncertainty across naming events exists and should play a role in models of cross-situational word learning.

Thus, cross-situational word learning can appear distributional or discrete, and the input to statistical learning mechanisms can vary along a continuum from low to high ambiguity depending as a function of the communicative context. This point highlights an important gap in the prior experimental work on cross-situational word learning. That is, the majority of this research has used linguistic stimuli that are generated by a disembodied voice coupled with a visual world that consists of pictures of concrete objects. In contrast, real world labeling events occur during face-to-face communicative interactions, which provide the learner with access to a rich set of visual cues (e.g., gestures, facial expressions, mouth movements) that could be used to constrain the ambiguity of input to cross-situational learning mechanisms.

This gap is important since social-pragmatic theories of language acquisition have long emphasized the role of social context in first language acquisition (Bloom 2002; Clark 2009; Hollich et al. 2000). Moroever, experimental work has shown that even children as young as 16 months prefer to map novel words to objects that are the target of a speaker's gaze and not their own (Baldwin 1993). In an analysis of naturalistic parent-child labeling events, Yu and Smith (2012) found that young learners tended to retain labels that were accompanied by clear referential cues, which served to make a single object dominant in the visual field. And correlational studies have demonstrated links between early intention-reading skills (e.g., gaze following) and later vocabulary growth (Brooks and Meltzoff 2005; Carpenter et al. 1998).

A second open question for models of cross-situational word learning is whether learners might be sensitive to the information processing demands of the input, using this information to flexibly adapt their behaviors to different learning contexts. The majority of prior research has focused on learning words in spoken language within clear listening contexts where the learner has perfect access to the auditory and visual information. This assumption, however, does not capture the variability in the types input that cross-situational mechanisms must operate over. For example, we know relatively little about how children's behavior might adapt to contexts where fixating on another person is critical for language acquisition as in the case of children learning a visual-manual language, like American Sign Language. The sign learning context creates a tradeoff where young signers must decide whether to look at their social partner to gather information about language or to look at the nonlinguistic visual world to gather information about objects. This channel competition potentially complicaties the link between the in-the-moment processes of establishing reference and long-term rentention of object labels.

My dissertation work takes a step towards address these two open questions. Specifically, I have asked how the behaviors that support familiar language comprehension and cross-situational word learning adapt to a wider variety of learning contexts. These contexts include language accompanied by social cues to reference, sign language processing, and comprehending language within noisy auditory contexts. These contexts represent a broad sampling of language environments but share a key feature: The interaction between listener and context modulates the value of gathering and storing certain kinds of information for language comprehension and learning.

In the next three sections, I briefly review the completed dissertation work before motivating the current study. The proposed study aims to connect our prior work on eye movements for information seeking during familiar language comprehension with work on cross-situational word learning. The study will measure how listeners flexibly adapt the dynamics of eye movements away from a language source as word learning unfolds across multiple labeling events. Overall, the results of the proposed study will aim to synthesize ideas from social-pragmatic theories of language acqusition with work on goal-based vision to increase our understanding of how decisions about visual fixation change as learners acquire a new word meaning over time.

Completed work

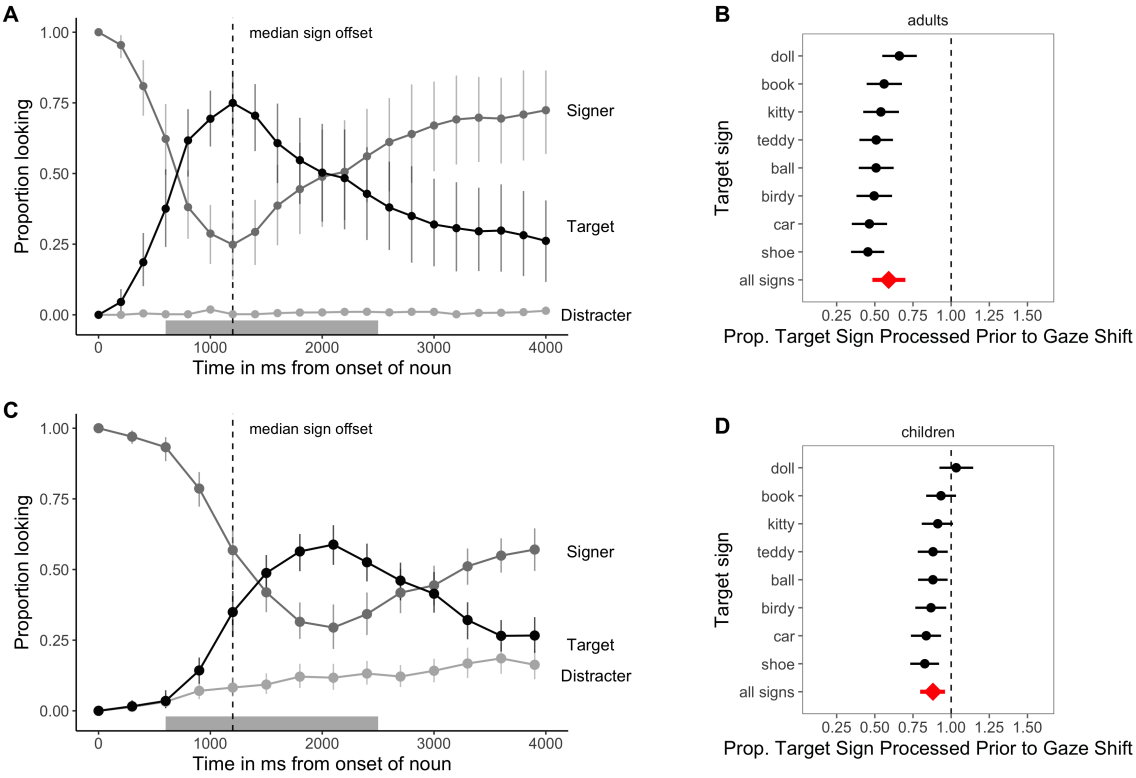
Chapter 1: Eye movements during real-time American Sign Language comprehension

Study overview

In our previous work, we characterized how children and adults choose to allocate visual attention between a social partner and objects during familiar American Sign Language (ASL) comprehension (Chapter 1).

When children interpret spoken language in real time, linguistic information drives rapid shifts in visual attention to objects in the visual world, which can provide insights into the development of efficiency in lexical access. But how does language influence visual attention when the linguistic signal and the visual world are both processed via the visual channel? We developed precise measures of eye movements during real-time comprehension of a visual-manual language, American Sign Language (ASL), by 29 native, monolingual ASL-learning children (16-53 mos, 16 deaf, 13 hearing) and 16 fluent deaf adult signers. All signers showed evidence of rapid, incremental language comprehension, initiating eye movements prior to sign offset. Deaf and hearing ASL-learners showed remarkably similar gaze patterns, suggesting that the in-the-moment dynamics of eye movements during ASL processing are shaped by the constraints of processing a visual language in real time and not by differential access to auditory information in day-to-day life. Finally, variation in children's ASL processing was positively correlated with age and vocabulary size. Thus, despite channel competition, allocation of visual attention during ASL comprehension reflects information processing skills that are fundamental for language acquisition regardless of language modality.

Key takeaway



Chapter 2: Comparing eye movements during real-time spoken and signed language processing: An information seeking account

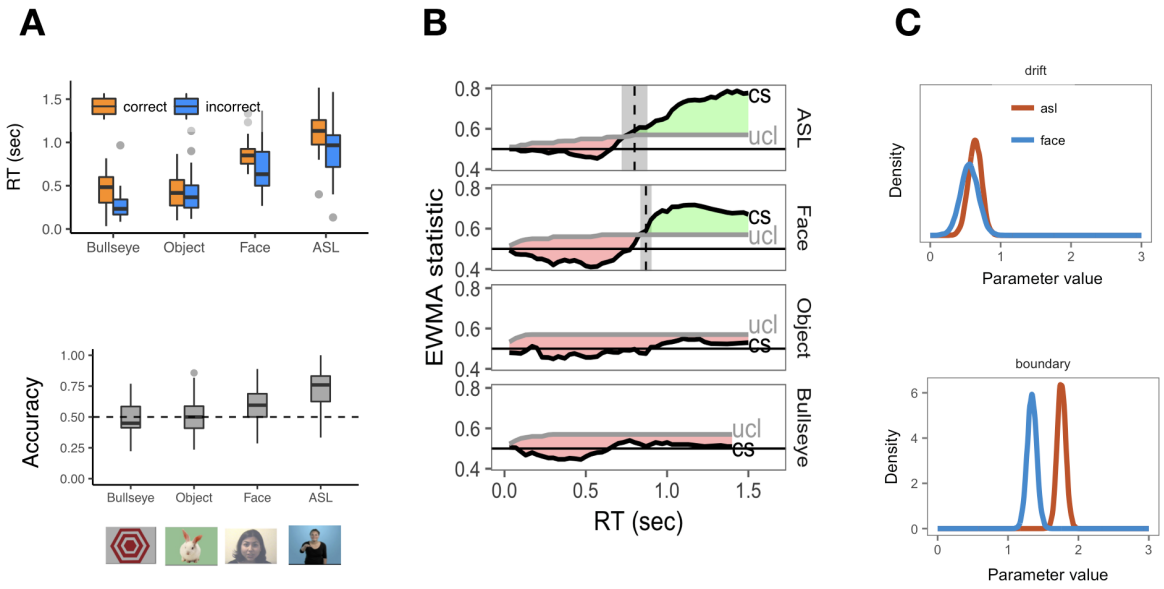
Study overview

The study of eye movements during language comprehension has provided fundamental insights into the interaction between conceptual representations of the world and the incoming linguistic signal. For example, research shows that adults and children will rapidly shift visual attention upon hearing the name of an object in the visual scene, with a high proportion of shifts occurring prior to the offset of the word (Allopenna, Magnuson, and Tanenhaus 1998; Tanenhaus et al. 1995). Moreover, researchers have found that conceptual representations activated by fixations to the visual world can modulate subsequent eye movements during language processing (Altmann and Kamide 2007).

The majority of this work has used eye movements as a measure of the output of the underlying language comprehension process, often using linguistic stimuli that come from a disembodied voice. But in real world contexts, people also gather information about the linguistic signal by fixating on the language source. Consider a speaker who asks you to "Pass the salt" but you are in a noisy room, making it difficult to understand the request. Here, comprehension can be facilitated by gathering information via (a) fixations to the nonlinguistic visual world (i.e., encoding the objects that are present in the scene) or (b) fixations to

the speaker (i.e., reading lips or perhaps the direction of gaze). But, this situation creates a tradeoff where the listener must decide what kind of information to gather and at what time. How do we decide where to look? We propose that people modulate their eye movements during language comprehension in response to tradeoffs in the value of gathering different kinds of information.

We test this adaptive tradeoff account using two case studies that manipulate the value of different fixation locations for language understanding: a) a comparison of processing sign vs. spoken language in children (E1), and b) a comparison of processing printed text vs. spoken language in adults (E2). Our key prediction is that competition for visual attention will make gaze shifts away from the language source less valuable than fixating the source of the linguistic signal, leading people to generate fewer exploratory, nonlanguage-driven eye movements.



Some of our work has explored how the presence of another person changes the set of information seeking behaviors available (MacDonald et al. 2017). Inspired by theories of natural vision that characterize eye movements as an information seeking mechanism, we asked whether children and adults would allocate more visual attention to a speaker when the linguistic signal was noisy to support the goal of rapid language understanding. We used an eye-tracking task to measure participants' gaze patterns while they processed clear or degraded speech (speech with brown noise added). Both children and adults spent more time fixating on the speaker in the degraded speech context. Interestingly, children and adults were also more accurate in word recognition even though the speech was noisy and difficult to process. This result suggests that listeners were compensating for the uncertainty in the auditory channel by gathering visual information from the speaker. Critically, listeners would not have been able to gather this information if the speaker was not present (e.g., listening to a noisy recording) and in clear view.

Key takeway

We then compared the dynamics of eye movements during familiar ASL processing to those of spoken language learners, showing that ASL-learners gather more information about the linguistic signal before shifting away from a language source compared to spoken language learners. We proposed an information-seeking account to explain these modality-based differences and tested predictions of our account across a variety of language comprehension contexts. We found the same pattern of eye movements for English-speaking adults processing displays of printed text and for both children and adults processing speech in noisy auditory environments (Chapter 2). These results suggest that listeners flexibly adapt eye movements to the value of seeking higher value visual information to support their goal of rapid langauge comprehension.

Chapter 3: Social cues to reference modulate attention and memory during cross-situational word learning

Study overview

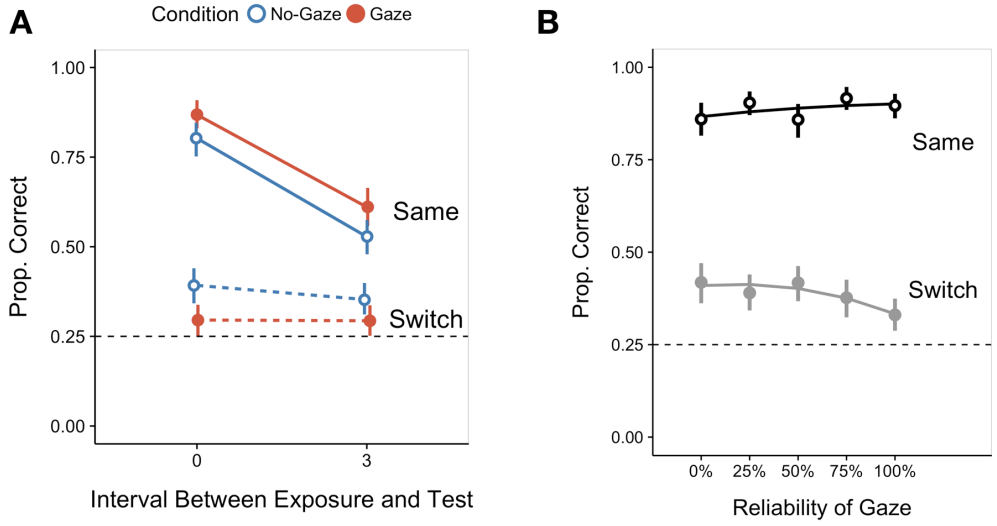
Our prior work provides evidence that the social context can modulate the content of the learner's hypothesis space (MacDonald, Yurovsky, and Frank 2017). Inspired by ideas from Social-pragmatic theories of language acquisition that emphasize the importance of social cues for word learning (Clark 2009; Hollich et al. 2000; Bloom 2002), we showed adults a series of word learning contexts that varied in ambiguity depending on whether there was a useful social cue to reference available (a speaker's gaze). We then measured learners' memory for alternative word-object links. People flexibly responded to the amount of ambiguity in the input, and as uncertainty increased, they tended to store more word-object hypotheses. Moreover, we found that learners stored representations with different levels of fidelity as a function of the reliability of the social cue. When the speaker was a less reliable source of information, learners distributed attention and memory broadly, storing more hypotheses.

These results provide evidence that the content of learners' hypothesis spaces changed as a function of social information. Further suppport for this idea comes from experimental work showing that even children as young as 16 months prefer to map novel words to objects that are the target of a speaker's gaze and not their own (Baldwin 1993), and analyses of naturalistic parent-child labeling events shows that young learners tended to retain labels accompanied by clear referential cues, which served to make a single object dominant in the visual field (Yu and Smith 2012). One important direction for future research is to measure the full causal pathway from variation in social information through children's hypothesis spaces to their information seeking behaviors. For example, it would be interesting to know whether learners' subsequent questions or decisions about where to allocate attention would be affected by the social context in which they were first exposed to a new word.

Key takeaway

In a separate line of work, we have asked how the presence of a social cue to reference – a speaker's gaze – could change the representations that support novel word learning (Chapter 3). Our results suggest that

word learneers stored representations with different levels of fidelity depending on the amount of ambiguity present during learning. In the absence of a referential cue to word meaning, learners tended to store more alternative word-object links. In contrast, when gaze was present learners stored less information, showing behavior consistent with tracking a single hypothesis. Thus, word learners flexibly respond to the amount of ambiguity in the input, and as referential uncertainty increases, they tend to store more information.



Proposed work

The goal of the proposed work is to understand how children use the presence of social information to help solve the problem of mapping concrete nouns to their referents amidst referential uncertainty. We will test an information-theoretic account of eye movements within a context where the child has uncertainty over word-object links. Our hypothesis is that gathering visual information from a speaker becomes more useful when uncertainty over word meanings is high and the goal is to learn word-object links. As the learner builds stronger word-object links via repeated exposures to co-occurence information, we predict that the value of allocating fixations to the speaker should decrease while the value of looking to the objects should increase.

Balancing looks to people and to objects during word learning

This project aims to answer the following research questions:

How does access to social cues shape in-the-moment decisions about visual fixation? How do children balance looks to people and to objects over the course of learning a new concept? Do children use prior knowledge to select fixation behaviors that best support word learning?

The word learning context is an interesting case because learners are working towards multiple goals: comprehending speech in the moment (a dynamic intergration of linguistic and visual signal with prior

knowledge) and figuring out what the new word refers to in the visual scene. Thus, eye movements during concept formation can be used to gather visual information from the speaker (e.g., eye gaze or mouth movements) or about the nonlinguistic visual world (encoding objects). How do we explain where children look as they acquire more information in-the-moment of language comprehension and as they build a learning history about the correct word-object mapping? This question can be formalized as a sequential decision making problem where children make fixation choice based on (1) their knowledge of the target concept, (2) the value of fixating a speaker for linguistic processing, and (3) the cost of each eye movement.

Framing fixation behaviors as a goal-based decision-making problem allows us to connect to formal models of action selection developed to explain

A growing body of psychological research has used the OED framework as a metaphor for active learning. The idea is that when people make decisions, they engage in a similar process of evaluating the "usefulness" of different actions relative to their learning goals. And they select behaviors that maximize the potential for gaining information. A success of the OED account is that it can capture a wide range of information seeking behaviors, including verbal question asking (Ruggeri and Lombrozo 2015), planning interventions in causal learning tasks (Cook, Goodman, and Schulz 2011), and decisions about where to look during scene understanding (Najemnik and Geisler 2005). Figures 1 and 2 present schematic overviews of how OED principles could shape the learning process for two of these domains – causal learning (Figure 1) and word learning (Figure 2).

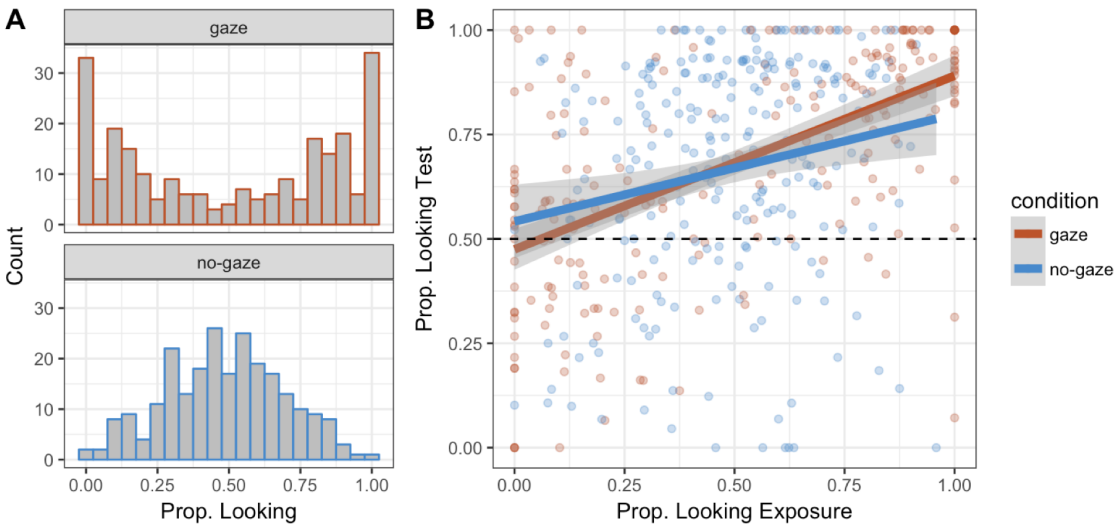
One compelling use case of OED metaphor as a model of human behavior comes from Nelson (2005) study of eye movements during novel concept learning. Their model combined Bayesian probabilistic learning, which represents current knowledge as a probability distribution over concepts, with an OED model that calculated the usefulness of different patterns of eye movements. Here, eye movements were modeled as a type of question-asking behavior that gathered visual information about the target concept. Nelson (2005) found that participants' eye movements aligned with predictions from the OED model. Specifically, participants changed the dynamics of eye movements depending on how well they learned the target concepts. Early in learning, when the concepts were unfamiliar, the model generated a broader, less efficient distribution of fixations to explore all candidate features that could be used to categorize the stimulus. However, after the model began to learn the target concepts, eye movement patterns shifted to become more efficient and focused on a single stimulus dimension to maximize accuracy. This shift from exploratory to efficient eye movements matched adult performance on the task, suggesting that people's behavior was sensible given the structure of the learning problem and the uncertainty in the context.

The intuition is that people balance fixating a speaker and fixating objects to support concept learning. The question is whether models of Bayesian concept learning and Optimal Experiment Design (Neslon & Cottrell, 2007) provide a good explanation of children's eye movements. How far can we get using a purely computational information seeking decision model?

Pilot

When gaze cued adults' visual attention, they showed stronger memory for the word-object link compared to when a gaze cue was absent (Figure 2). This result suggest that social information does more than modulate

how people allocate their visual attention (more than a filter); instead, social cues change the strength of the inference.



Design

Predictions

The prediction is that the dynamics of eye movement will shift over the course of learning. In the beginning of the task, learners will distribute fixations to prioritize gathering information about the objects or about disambiguating reference (e.g., gathering a gaze cue). After learning the word-object links, people will shift and start to distribute more fixations to the speaker to gather visual information that supports comprehension of the speech, showing the behavioral signatures measured in the familiar language comprehension task.

References

Allopenna, Paul D, James S Magnuson, and Michael K Tanenhaus. 1998. "Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models." *Journal of Memory and Language* 38 (4). Elsevier:419–39.

Altmann, Gerry, and Yuki Kamide. 2007. "The Real-Time Mediation of Visual Attention by Language and World Knowledge: Linking Anticipatory (and Other) Eye Movements to Linguistic Processing." *Journal of Memory and Language* 57 (4). Elsevier:502–18.

Baldwin, Dare A. 1993. "Infants' Ability to Consult the Speaker for Clues to Word Reference." *Journal of Child Language* 20 (02). Cambridge Univ Press:395–418.

Bloom, Paul. 2002. *How Children Learn the Meaning of Words*. The MIT Press.

Brooks, Rechele, and Andrew N Meltzoff. 2005. "The Development of Gaze Following and Its Relation to Language." *Developmental Science* 8 (6). Wiley Online Library:535–43.

Carpenter, Malinda, Katherine Nagell, Michael Tomasello, George Butterworth, and Chris Moore. 1998. "Social Cognition, Joint Attention, and Communicative Competence from 9 to 15 Months of Age." *Monographs of the Society for Research in Child Development*. JSTOR, i–174.

Cartmill, Erica A, Benjamin F Armstrong, Lila R Gleitman, Susan Goldin-Meadow, Tamara N Medina, and John C Trueswell. 2013. "Quality of Early Parent Input Predicts Child Vocabulary 3 Years Later." *Proceedings of the National Academy of Sciences* 110 (28). National Acad Sciences:11278–83.

Clark, Eve V. 2009. *First Language Acquisition*. Cambridge University Press.

Cook, Claire, Noah D Goodman, and Laura E Schulz. 2011. "Where Science Starts: Spontaneous Experiments in Preschoolers' Exploratory Play." *Cognition* 120 (3). Elsevier:341–49.

Hollich, George J, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, Rebecca J Brand, Ellie Brown, He Len Chung, Elizabeth Hennon, Camille Rocroi, and Lois Bloom. 2000. "Breaking the Language Barrier: An Emergentist Coalition Model for the Origins of Word Learning." *Monographs of the Society for Research in Child Development*. JSTOR, i–135.

MacDonald, Kyle, Aviva Blonder, Virgnia and Marchman, Anne Fernald, and Michael C Frank. 2017. "An Information-Seeking Account of Eye Movements During Spoken and Signed Language Comprehension." In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

MacDonald, Kyle, Daniel Yurovsky, and Michael C Frank. 2017. "Social Cues Modulate the Representations Underlying Cross-Situational Learning." *Cognitive Psychology* 94. Elsevier:67–84.

McMurray, Bob, Jessica S Horst, and Larissa K Samuelson. 2012. "Word Learning Emerges from the Interaction of Online Referent Selection and Slow Associative Learning." *Psychological Review* 119 (4). American Psychological Association:831.

Medina, Tamara Nicol, Jesse Snedeker, John C Trueswell, and Lila R Gleitman. 2011. "How Words Can and Cannot Be Learned by Observation." *Proceedings of the National Academy of Sciences* 108 (22). National Acad Sciences:9014–9.

Najemnik, Jiri, and Wilson S Geisler. 2005. "Optimal Eye Movement Strategies in Visual Search." *Nature* 434 (7031). Nature Publishing Group:387.

Nelson, Jonathan D. 2005. "Finding Useful Questions: On Bayesian Diagnosticity, Probability, Impact, and Information Gain." *Psychological Review* 112 (4). AMER PSYCHOLOGICAL ASSOC/EDUCATIONAL PUBLISHING FOUNDATION.

Quine, Willard V. 1960. "0. Word and Object." *111e MIT Press*.

Ruggeri, Azzurra, and Tania Lombrozo. 2015. "Children Adapt Their Questions to Achieve Efficient Search." *Cognition* 143. Elsevier:203–16.

Siskind, Jeffrey Mark. 1996. "A Computational Study of Cross-Situational Techniques for Learning Word-to-Meaning Mappings." *Cognition* 61 (1). Elsevier:39–91.

Smith, Kenny, Andrew DM Smith, and Richard A Blythe. 2011. "Cross-Situational Learning: An Experimental Study of Word-Learning Mechanisms." *Cognitive Science* 35 (3). Wiley Online Library:480–98.

Smith, Linda B, and Chen Yu. 2008. "Infants Rapidly Learn Word-Referent Mappings via Cross-Situational Statistics." *Cognition* 106 (3). Elsevier:1558–68.

Tanenhaus, Michael K, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. 1995. "Integration of Visual and Linguistic Information in Spoken Language Comprehension." *Science* 268 (5217). The American Association for the Advancement of Science:1632.

Trueswell, John C, Tamara Nicol Medina, Alon Hafri, and Lila Gleitman. 2013. "Propose but Verify: Fast Mapping Meets Cross-Situational Word Learning." *Cognitive Psychology* 66 (1). Elsevier:126–56.

Vouloumanos, Athena. 2008. "Fine-Grained Sensitivity to Statistical Information in Adult Word Learning." *Cognition* 107 (2). Elsevier:729–42.

Yu, Chen, and Linda B Smith. 2007. "Rapid Word Learning Under Uncertainty via Cross-Situational Statistics." *Psychological Science* 18 (5). SAGE Publications:414–20.

———. 2012. "Embodied Attention and Word Learning by Toddlers." *Cognition*. Elsevier.

Yurovsky, Daniel, Linda B Smith, and Chen Yu. 2013. "Statistical Word Learning at Scale: The Baby's View Is Better." *Developmental Science* 16 (6). Wiley Online Library:959–66.