Balancing looks to people and objects during word learning

Kyle MacDonald

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

Contents

1 Dissertation overview

Learning a new word 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 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

[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.

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.

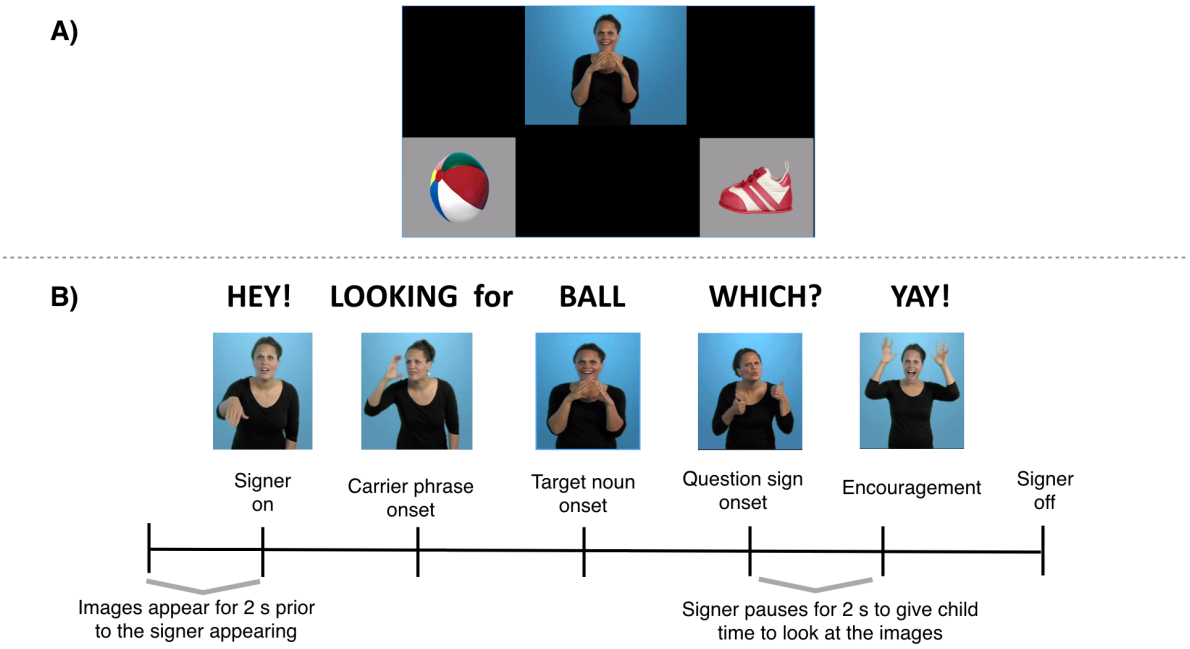

Figure 1: Configuration of visual stimuli (1A) and trial structure (1B) for one question type (sentence final wh-phrase) shown in the central video on the VLP task.

2 Completed work

2.1 Eye movements during real-time American Sign Language comprehension

2.1.1 Study overview

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 processes underlying real-time language comprehension. But how does language influence visual attention when the linguistic signal and the visual world are both processed via the visual channel? In this work, we measured eye movements during real-time comprehension of a visual-manual language, American Sign Language (ASL), by 29 native 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, tending to initiate an eye movement before sign offset. Deaf and hearing ASL-learners showed 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.

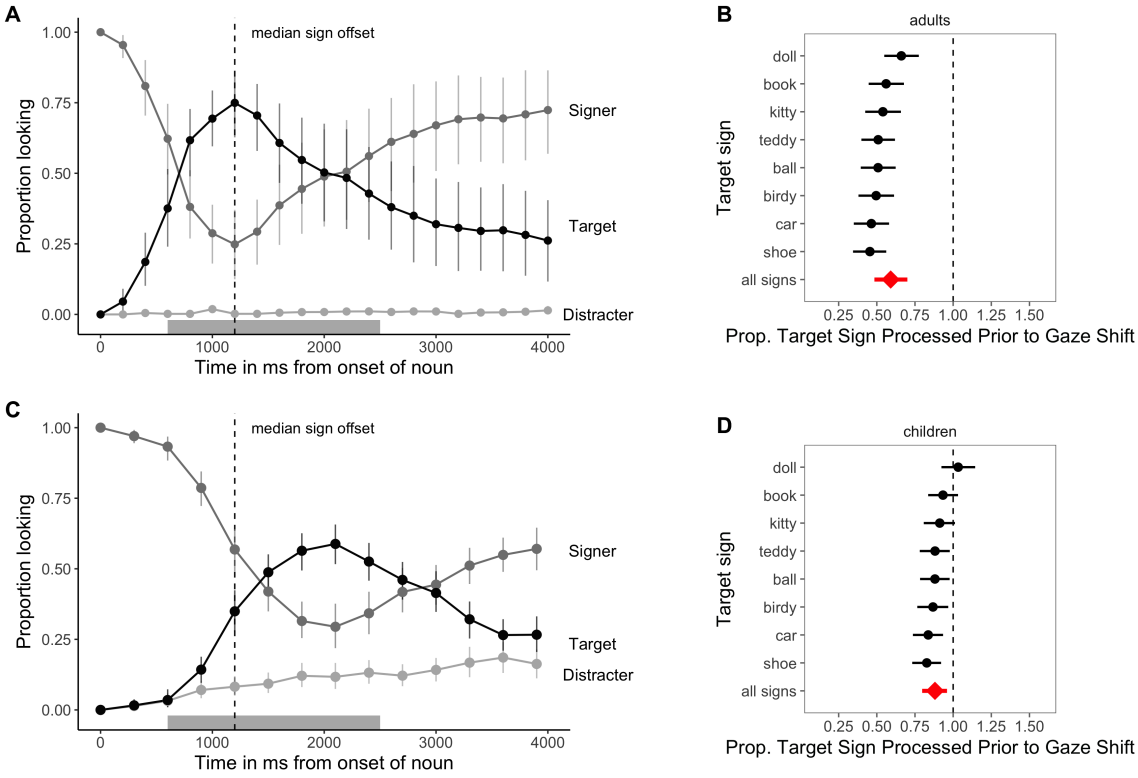

Figure 2: The time course of looking behavior for ASL-proficient adults (2A) and young ASL- learners (2C). The curves show mean proportion looking to the signer (dark grey), the target image (black), and the distracter image (light grey). The grey shaded region marks the analysis window (600-2500 ms); error bars represent 95% CI computed by non-parametric bootstrap. The mean proportion of each target sign length processed prior to shifting visual attention away from the language source to a named object for adults (2B) and children (2D). The diamond indicates the mean estimate for all signs. The dashed vertical line corresponds to a median proportion of 1.0. A median of greater than 1.0 reflects response latencies that occur prior to the offset of the target sign; a median of less than or equal to 1.0 reflects response latencies that occur after target sign offset. Error bars represent 95% Highest Density Intervals.

2.1.2 Method

The task was presented on a 27" monitor. On each trial, pictures of two familiar objects appeared on the screen, a target object corresponding to the target noun, and a distracter object (see Fig. **??**). All picture pairs were matched for visual salience based on prior studies with spoken language (Fernald et al., 2008). Between the two pictures was a central video of an adult female signing the name of one of the pictures. Participants saw 32 test trials with five filler trials (e.g. "YOU LIKE PICTURES? MORE WANT?") interspersed to maintain children's interest. Participants' gaze patterns were video recorded and later coded frame-by-frame at 33-ms resolution by highly-trained coders blind to target side.

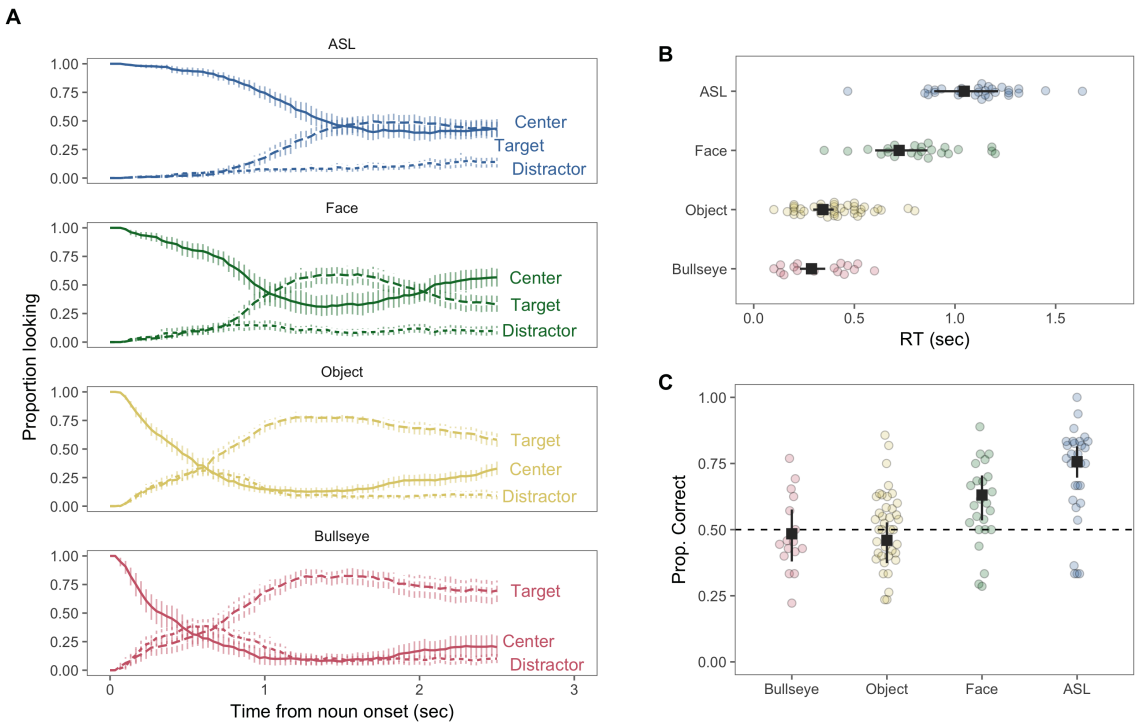

Figure 3: Timecourse looking and first shift Reaction Time (RT) and Accuracy results. Panel A shows the overall looking to the center, target, and distracter stimulus for each context. Panel B shows the distribution of RTs for each participant. Each point represents a participant's average RT. Color represents the processing context. Panel C shows the same information but for first shift accuracy.

2.1.3 Key takeaway

This study shows that, despite competition for attention within a single modality, both children and adults rapidly shifted visual attention away from a social partner and towards objects prior to sign offset. This result suggests that there is aa robust link between processing an object label and seeking that object in the visual world. However,

2.2 Comparing children's eye movements during real-time spoken and signed language processing: An information seeking account

2.2.1 Study overview

Language comprehension in grounded, social contexts involves extracting meaning from the linguistic signal and mapping it to the surrounding world. But how should listeners prioritize integrating information from the linguistic and visual signals? In this work, we proposed that listeners flexibly adapt their gaze behaviors in response to features of the language processing context, seeking visual information from their social partners that supports language comprehension. We present evidence for our account using three case studies, sampled

from a diverse set of language processing contexts. First, compared to children learning spoken English (n=80) and adults (n=25), young ASL-learners (n=30) and adults (n= 16) delayed their gaze shifts away from a language source, were more accurate with these shifts, and produced a smaller proportion of random shifting behavior (see Fig. **??**). Next, English-speaking adults produced fewer random gaze shifts when processing dynamic displays of printed text compared to processing spoken language. Finally, 3-5 year-olds (n=39) and adults (n=31) delayed the timing of gaze shifts away from a speaker's face when processing speech in a noisy environment, which resulted in fewer random eye movements, and more accurate gaze shifts, despite the noisier processing context (see Fig. **??**).

2.2.2 Methods

The design and procedure was parallel to the work on children's eye movements reviewed above. We compared the timing and accuracy of eye movements for children learning ASL to children learning a spoken language using parallel real-time language comprehension tasks where participants processed familiar sentences (e.g., "Where's the ball?") while looking at a simplified visual world with 3 fixation targets (a center stimulus that varied by condition, a target picture, and a distracter picture). In this work, we analyzed the timing and accuracy of children's initial gaze shifts after the onset of the target noun. The timescale of this analysis is milliseconds and focused on a single decision within a series of decisions about where to look during sentence processing. We made thie decision because first shifts provide a window onto changes in the underlying dynamics of how listeners integrate linguistic information in the unfolding word with the decision process that generates eye movements.

We predicted that, compared to spoken language processing, processing ASL would increase the value of fixating on the language source and decrease the value of generating exploratory, nonlanguage-driven shifts after the target linguistic item began unfolding in time. We also hypothesized that a noisy auditory environment would increase the value of fixating a speaker to gather visual information. Thus we predicted that adults and children in noisy contexts would delay generating an eye movement away from a speaker until they have accumulated additional visual information about the identity of the named referent.

2.2.3 Key takeway

Taken together, these results provide evidence that young listeners, like adults, will adapt their gaze patterns to the demands of different processing environments by seeking out visual information from social partners to support language comprehension.

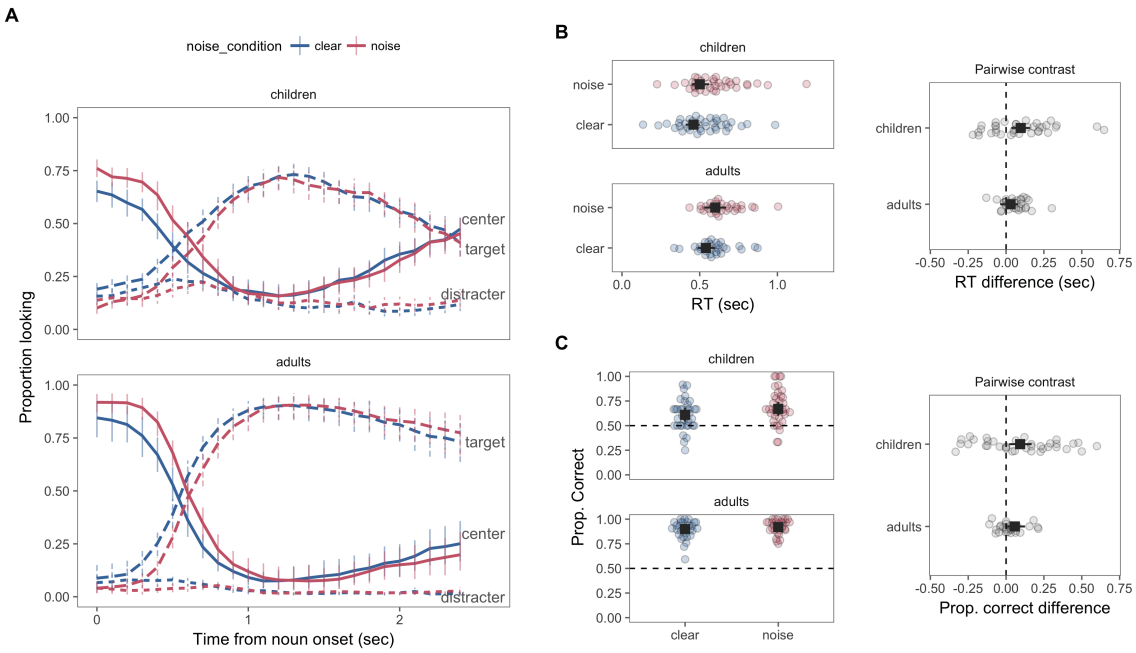

Figure 4: Timecourse looking and first shift Reaction Time (RT) and Accuracy results. Panel A shows the overall looking to the center, target, and distracter stimulus for each processing condition and age group. Panel B shows the distribution of RTs for each participant and the pairwise contrast between the noise and clear conditions. Panel C shows the same information but for first shift accuracy.

2.3 Social cues to reference modulate adults attention and memory during cross-situational word learning

2.3.1 Study overview

Because children hear language in environments that contain many things to talk about, learning the meaning of even the simplest word requires making inferences under uncertainty. A cross- situational statistical learner can aggregate across naming events to form stable word-referent mappings, but this approach neglects an important source of information that can reduce referential uncertainty: social cues from speakers (e.g., eye gaze). In four large-scale experiments with adults, we tested the effects of varying referential uncertainty in cross-situational word learning using social cues. Social cues shifted learners away from tracking multiple hypotheses and towards storing only a single hypothesis (Experiments 1 and 2; Panel A of Fig. 6). In addition, learners were sensitive to graded changes in the strength of a social cue, and when it became less reliable, they were more likely to store multiple hypotheses (Experiment 3; Panel B of Fig. 6). Finally, learners stored fewer word-referent mappings in the presence of a social cue even when given the opportunity to visually inspect the objects for the same amount of time (Experiment 4).

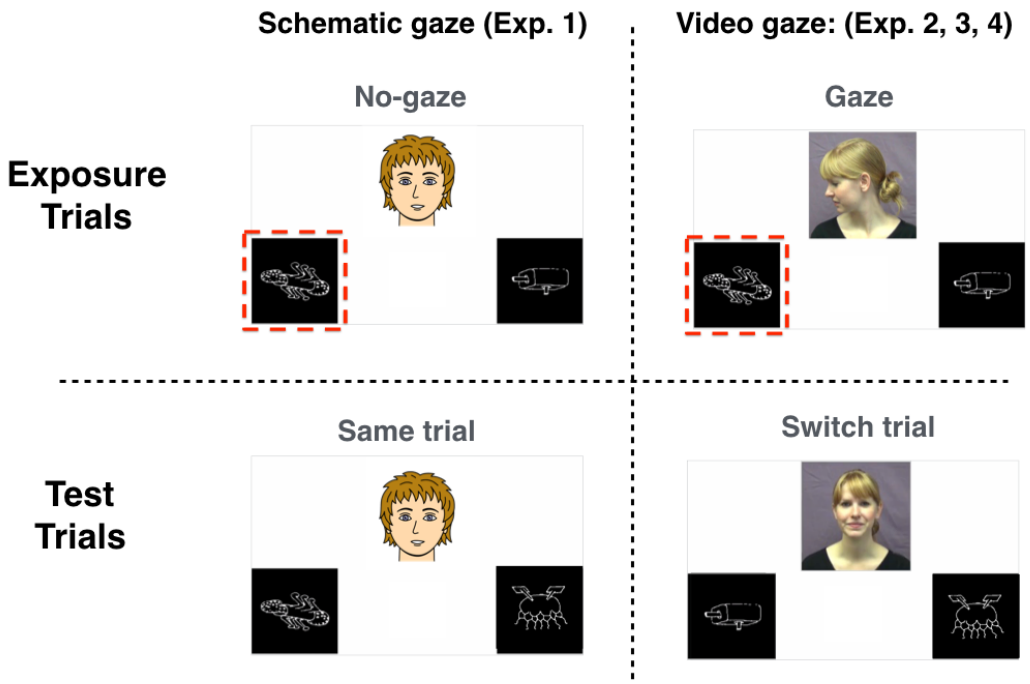

Figure 5: Exposure and test trials from Experiments 1-4. The top left panel shows an exposure trial in the No-gaze condition using the schematic gaze cue (Experiment 1). The top right panel shows an exposure trial in the Gaze condition using the video gaze cue (Experiments 2-4). Participants saw either Gaze or No-gaze exposure trials depending on condition assignment, and participants saw both types of test trials: Same (bottom left panel) and Switch (bottom right panel). On Same trials, the object that participants chose during exposure appeared with a new novel object. On Switch trials the object that participants did not choose appeared with a new novel object.

2.3.2 Method

Experiments 1-4 followed a similar design and procedure. We posted a set of Human Intelligence Tasks (HITs) to Amazon Mechanical Turk. Adults saw a total of 16 trials: eight exposure trials and eight test trials. On each trial, they heard one novel word, saw a set of novel objects, and were asked to guess which object went with the word (see Fig. **??**). Before seeing exposure and test trials, participants completed four practice trials with familiar words and objects. These trials familiarized participants to the task and allowed us to exclude participants who were unlikely to perform the task as directed, either because of inattention or because their computer audio was turned off.

After the practice trials, participants were told that they would now hear novel words and see novel objects and that their task was to select the referent that "goes with each word." Over the course of the experiment, participants heard eight novel words two times, with one exposure trial and one test trial for each word. Four of the test trials were Same trials in which the object that participants selected on the exposure

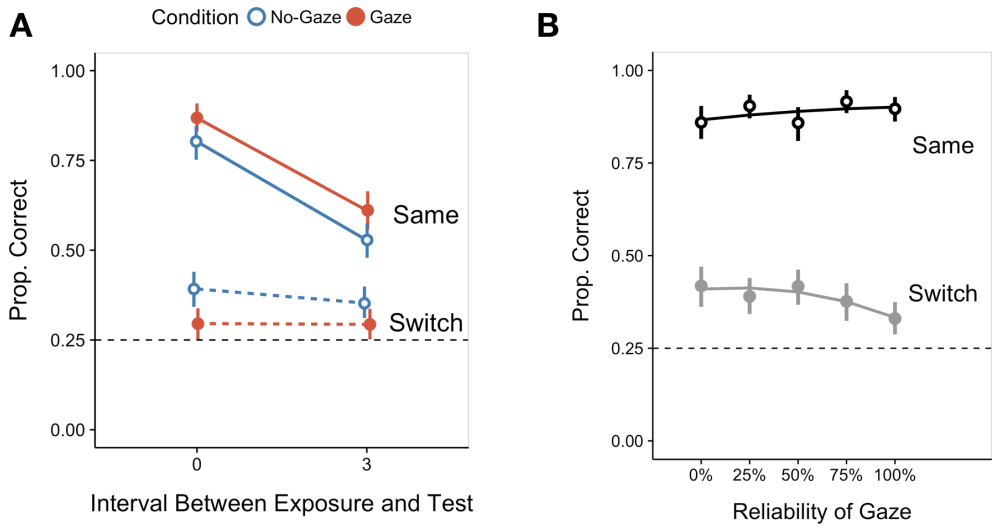

Figure 6: Panel A shows accuracy on Same and Switch test trials as a function of the interval between exposure and test trials. Panel B shows memory performance as a function of the reliability of the speaker's prior gaze cues. Error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

trial was shown with a set of new novel objects. The other four test trials were Switch trials in which one of the objects was chosen at random from the set of objects that the participant did not select on exposure.

In Experiment 3, we modified the cross-situational learning paradigm to include a block of 16 familiarization trials (8 exposure trials and 8 test trials) at the beginning of the experiment. These trials served to establish the reliability of the speaker's gaze. To establish reliability, we varied the proportion of Same/Switch trials that occurred during the familiarization block. Recall that on Switch trials the gaze target did not show up at test, which provided evidence that the speaker's gaze was not a reliable cue to reference. Reliability was a between-subjects manipulation such that participants either saw 8, 6, 4, 2, or 0 Switch trials during familiarization, which created the 0%, 25%, 50%, 75%, and 100% reliability conditions. After the familiarization block, participants completed another block of 16 trials (8 exposure trials and 8 test trials).

2.3.3 Key takeaway

Taken together, our data suggest that the representations underlying cross-situational word learning of concrete object labels are quite flexible: In conditions of greater uncertainty, learners store a broader range of information.

3 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.

3.1 Balancing looks to people and to objects during word learning

This study 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

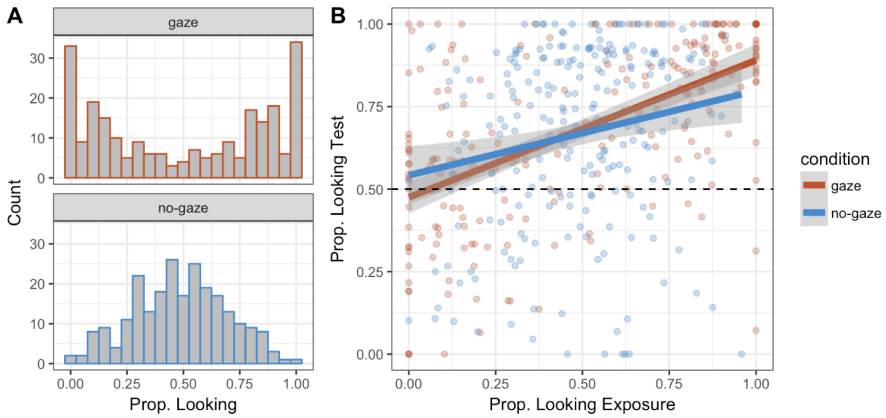

Figure 7: Panel A shows adults' allocation of visual attention during learning trials. Panel B shows adults stronger memory for novel word-object links when their attention was cued by gaze compared to when a gaze cue was absent, even when they had fixated on the target objects for similar amounts of time during learning.

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?

3.1.1 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.

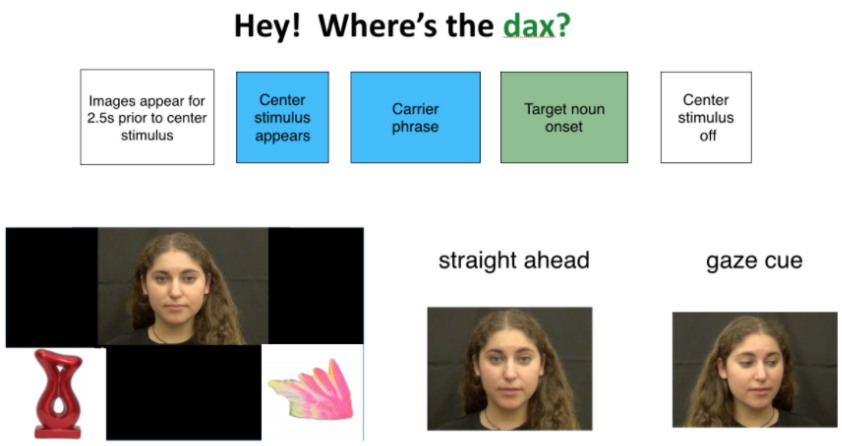

Figure 8: Proposed stimuli information, including the trial structure, fixation targets in the visual world, and the gaze cue manipuulation.

3.1.2 Design

3.1.3 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.

4 References

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.

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.

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.