Balancing looks to people and to objects during cross-situtational word learning

Kyle MacDonald

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

Contents

1 Background

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.

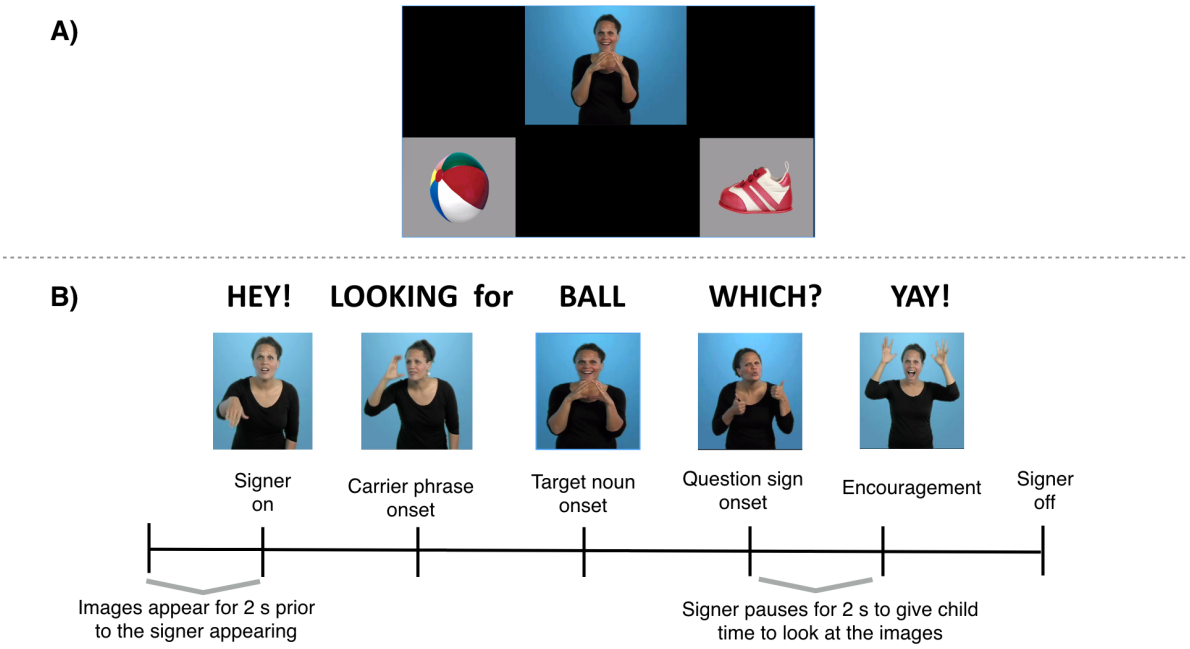

Figure 1: Configuration of visual stimuli (1A) and trial structure (1B) for one question type (sentence final wh-phrase) shown in the central video on the VLP task.

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 (see Fig. 2). 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.

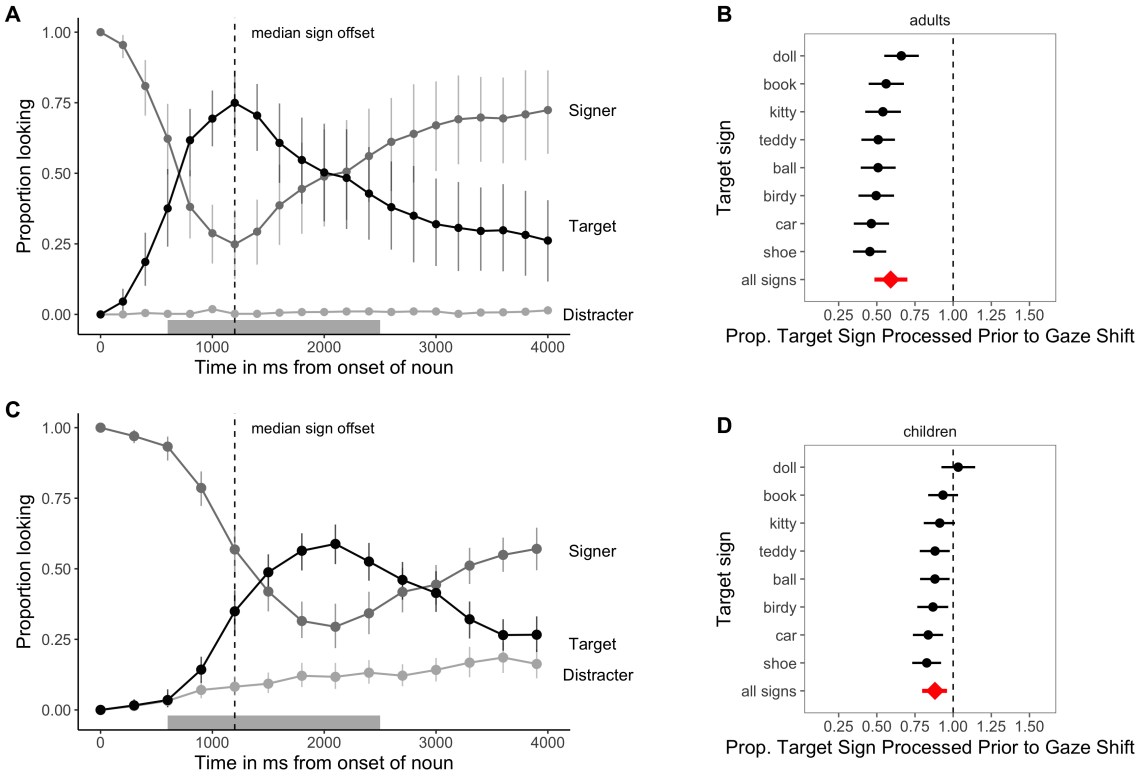

Figure 2: The time course of looking behavior for ASL-proficient adults (2A) and young ASL- learners (2C). The curves show mean proportion looking to the signer (dark grey), the target image (black), and the distracter image (light grey). The grey shaded region marks the analysis window (600-2500 ms); 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. 1). 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.

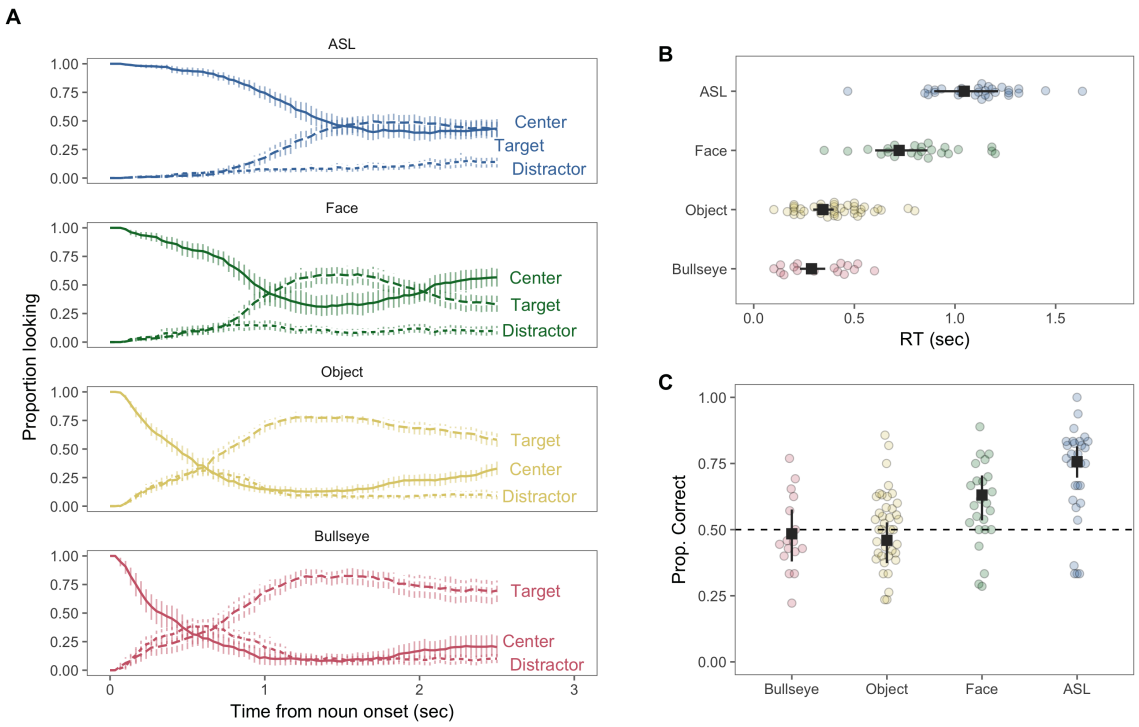

Figure 3: Timecourse looking and first shift Reaction Time (RT) and Accuracy results. Panel A shows the overall looking to the center, target, and distracter stimulus for each context. Panel B shows the distribution of RTs for each participant. Each point represents a participant's average RT. Color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

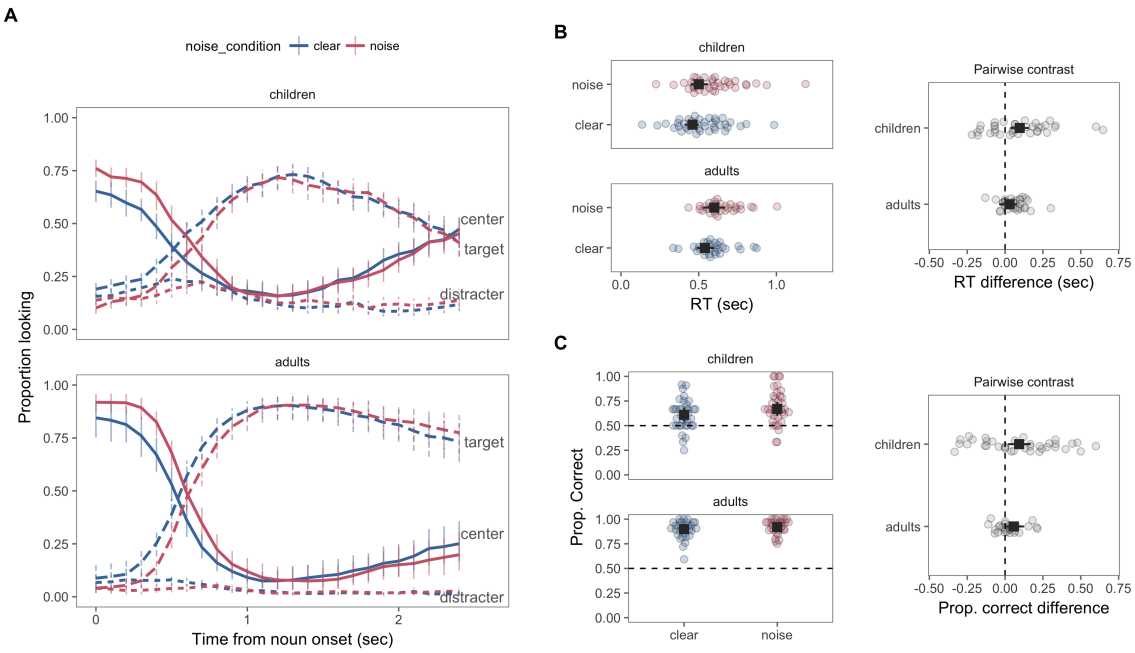

Figure 4: Timecourse looking and first shift Reaction Time (RT) and Accuracy results. Panel A shows the overall looking to the center, target, and distracter stimulus for each processing condition and age group. Panel B shows the distribution of RTs for each participant and the pairwise contrast between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

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. 3). 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. 4).

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.

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).

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. 5). 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.

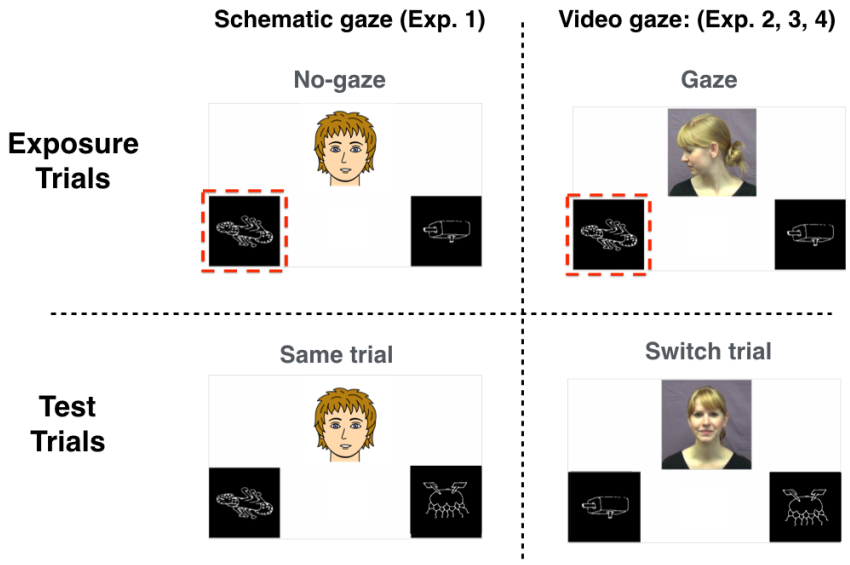

Figure 5: Exposure and test trials from Experiments 1-4. The top left panel shows an exposure trial in the No-gaze condition using the schematic gaze cue (Experiment 1). The top right panel shows an exposure trial in the Gaze condition using the video gaze cue (Experiments 2-4). Participants saw either 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.

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 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).

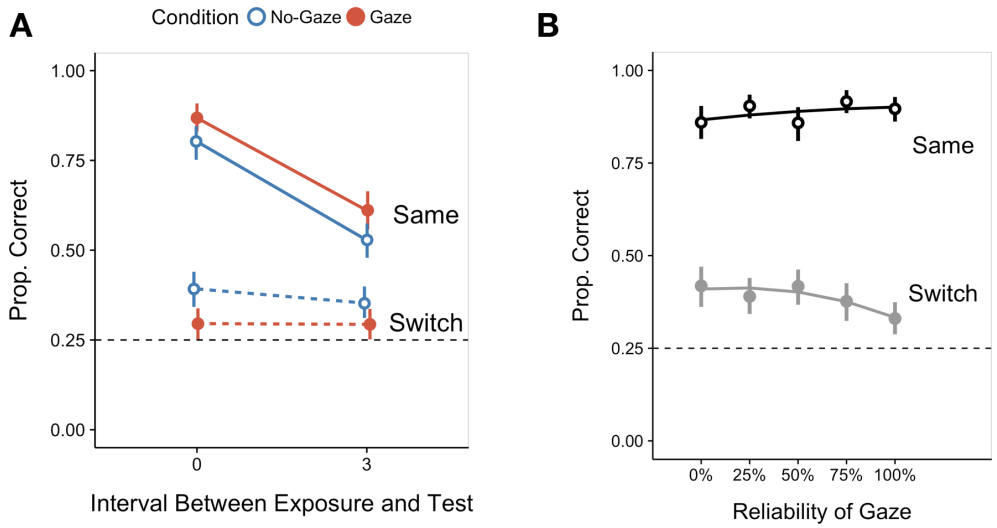

Figure 6: Panel A shows accuracy on Same and Switch test trials as a function of the interval between exposure and test trials. Panel B shows memory performance as a function of the reliability of the speaker's prior gaze cues. Error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

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 bring together aspects of social and statistical accounts of word learning to better understand how statistical learning shapes decisions about visual fixation within grounded, social interactions. We will measure how the dynamics of children's eye movements during object labeling change as a function of their uncertainty over the correct word-object mappinsg and the value of different fixation locations (i.e., speaker vs. objects). Our hypothesis is that looks to a speaker are most useful early in learning, when uncertainty over word meanings is high. As the learner builds stronger word-object links via repeated exposures to word-object 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.

The study aims to answer the following research questions:

- How does access to visual information from a speaker change in-the-moment decisions about visual fixation?
- How do children decide when to gather visual information about objects vs. people?

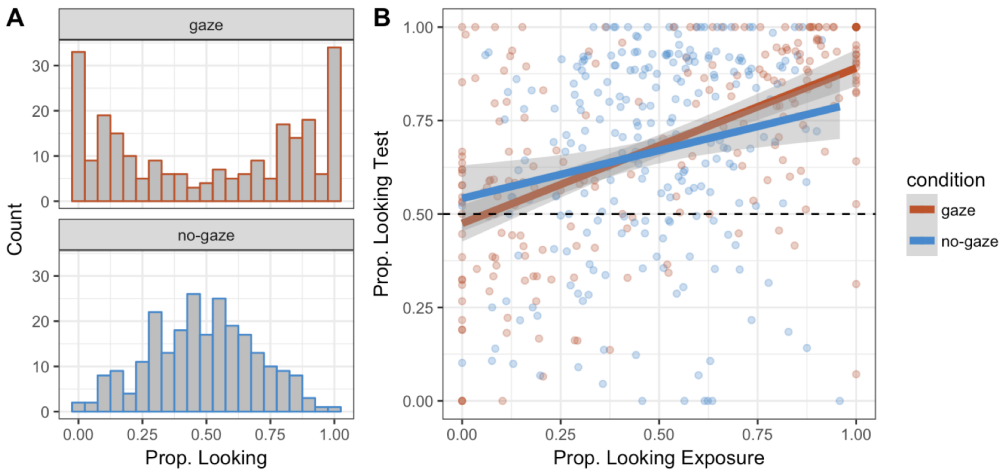

Figure 7: Panel A shows adults' allocation of visual attention during learning trials. Panel B shows adults stronger memory for novel word-object links when their attention was cued by gaze compared to when a gaze cue was absent, even when they had fixated on the target objects for similar amounts of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

- How do decisions about where to look change over the course of learning new words?

3.1 Pilot

Our prior work showed that the presence of a social cue (speaker gaze) shifted adults away from storing multiple word-object links and towards tracking a single hypothesis. However, those experiments relied on an offline measurement of word learning (button press at test) and an indirect measure of attention during learning (self-paced inspection time of the visual scene). To address these limitations, we adapted the paradgim to use eye-tracking methods, allowing us to ask two questions:

1. How does the presence of gaze alter the distribution of visual attention during labeling?
2. Does gaze cueing change the strength of the inferences about word-object links?

3.1.1 Method

Adults' (n=30) eye movements were tracked while they watched a series of ambiguous word-learning events (16 novel words) organized into pairs of exposure and test trials (32 trials total). All trials consisted of a set of two novel objects and one novel word. Participants were randomly assigned to either the Gaze condition in which a speaker looked at one of the objects on exposure trials or the No-Gaze condition in which a speaker looked straight on exposure trials. Every exposure trial was followed by a test trial, where participants heard the same novel word paired with a new set of two novel objects. One of the objects in the set had appeared in the exposure trial ("kept" object), while the other object had not previously appeared in the experiment

("novel" object).

3.1.2 Key takeway

We found that the presence of social cues focuses visual attention on a single object (Panel A of Fig. 7) and that gaze-cued attention lead to stronger inferences for word-referent pairings, with adults allocating more attention to the correct object at test in the gaze condition despite fixating on the target objects for similar amounts of time during learning (Panel B of Fig. 7). 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.

3.1.3 Limitations

The key limitation of our pilot paradigm was the timing of the linguistic stimulus with respect to the images appearing in the visual world. That is, the language started as soon as the images and the speaker appeared on the screen, making it difficult to analyze the timing/accuracy of first shifts decisions away from the speaker and to the objects. Moreover, this trial structure does not allow us to measure decisions about visual fixation that occur prior to the start of language, i.e., while learners are first exploring the visual world. Finally, the linguistic stimuli consisted of sixteen pseudowords recorded by a speech synthesizer and presented in isolation and thus removed from any sentential context. This mode of presentation is unlikely to work with younger age groups, and does not allow us to separate decisions about fixations during language processing more broadly from decisions that occur after the onset of the target noun.

3.2 Study proposal

3.2.1 Participants

We will collect data from an adult sample (n=30) using the the Stanford Psychology Credit Pool. We will recruit data from children ages 3-5 years (n=50) from the Bing Nursery school.

3.2.2 Stimuli and Design

We will compare the timing and accuracy of eye movements during a real-time cross-situational learning task where participants process sentences that contain a novel word (e.g., "Where's the dax?") while looking at a simplified visual world with 3 fixation targets (a video of a speaker and two images of unfamiliar objects). The trial structure will be identical to our work on eye movements during familiar language comprehension reviewed in section 2.2. This parallel structure will allow us to take advantage of the techniques for analyzing shifts in the dynmacis of participants' initial gaze shifts after the onset of the target noun.

Participants will watch a series of ambiguous word-learning events (four novel words) organized into pairs of exposure and test trials. All trials will consist of of a set of two unfamiliar objects and one novel

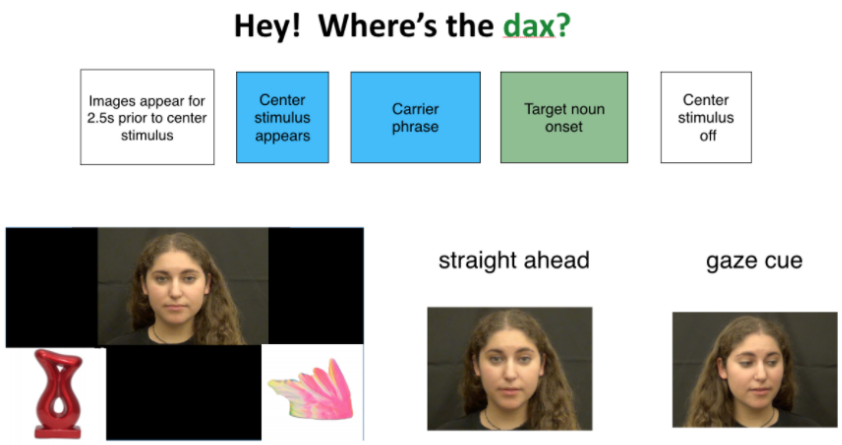

Figure 8: Proposed stimuli information, including the trial structure, fixation targets in the visual world, and the gaze cue manipulation.

word. Each exposure trial will be followed by a test trial, where participants will hear the same novel word but paired with a new set of two novel objects. One of the objects in the set will have appeared in the exposure trial (target object), while the other object will be new, i.e., not previously shown in the experiment (distractor object). Critically, participants will see four exposure trials for each of the four novel word-object pairings over the course of the entire experiment. By including multiple exposures, we can analyze changes decisions about visual fixation as participants build up a stronger word-object link in memory. In sum, each participant will see: 4 novel words X 2 trial types X 4 exposures = 32 trials total

We will manipulate the value of the speaker as a target for visual fixation by varying whether the speaker provides a post-nominal gaze cue. Participants will be randomly assigned to either a Gaze condition or No-Gaze condition. In the Gaze condition, the speaker will look at one of the objects after producing the novel label, thus fully disambiguating reference. In the No-Gaze condition, the speaker will continue to look straight-ahead at the participant throughout the exposure trial. This will be a within-participants manipulation with adults (blocked design) and a between-participants manipulation with children.

3.2.3 Analysis plan

We will follow the analysis plan developed in our prior work on familiar language comprehension and described in section 2.2. We will analyze First Shift Accuracy and Reaction Time (RT). RT corresponds to the latency to shift away from the central stimulus to either picture measured from the onset of the target noun. Accuracy corresponds to whether participants' first gaze shift landed on the target or the distracter picture.

Next, we will use two model-based analyses to link observable behavior (accuracy and RT) to underlying psychological constructs. We will use an exponentially weighted moving average (EWMA) method (Vandekerckhove and Tuerlinckx 2007) to classify gaze shifts as language-driven or random. In contrast to the standard RT/accuracy analysis, the EMWA approach allows us to quantify participants' willingness to generate gaze shifts after noun onset but before collecting sufficient information to seek the named referent.

Concretely, the EWMA models changes in random shifting behavior as a function of RT. For each RT, the model generates two values: a "control statistic" (CS, which captures the running average accuracy of first shifts) and an "upper control limit" (UCL, which captures the pre-defined limit of when accuracy would be categorized as above chance level). Here, the CS is an expectation of random shifting to either the target or the distracter image (nonlanguage-driven shifts), or a Bernoulli process with probability of success 0.5. As RTs get slower, we assume that participants have gathered more information and should become more accurate (language-driven), or a Bernoulli process with probability success > 0.5. Using this model, we can quantify the proportion of gaze shifts that were language-driven as opposed to random responding.

Finally, we use drift-diffusion models (DDMs) (Ratcliff and Childers 2015) to ask whether any measured behavioral differences in accuracy and RT are driven by a shift towards a more cautious responding strategy (i.e., gathering more visual information) or by more efficient information processing of the linguistic stimuli. We will follow Vandekerckhove and Tuerlinckx (2007) and select shifts categorized as language-driven by the EWMA and fit a hierarchical Bayesian drift-diffusion model (HDDM). The DDM quantifies differences in the underlying decision process that lead to different patterns of behavior. The model assumes that people accumulate noisy evidence in favor of one alternative with a response generated when the evidence crosses a pre-defined decision threshold. Here, we focus on two parameters of interest: *boundary separation*, which indexes the amount of evidence gathered before generating a response (higher values indicate more cautious responding) and *drift rate*, which indexes the amount of evidence accumulated per unit time (higher values indicate more efficient processing).

We will also analyze particpants' proportion looking to the objects (target vs. distractor) vs. the speaker during the different phases of exposure trials, e.g., the 2.5 seconds prior to the speaker appearing on the screen and two seconds after noun onset. In contrast to the first shift analyses, the proportion looking analysis is on the timescale of seconds and collapses across a series of decisions about where to look over the course of the trial.

For all statistical models, we will use the `rstanarm` (Gabry and Goodrich 2016) package to fit Bayesian mixed-effects regression models. The mixed-effects approach will allow us to model any nested structure in our data – multiple trials for each participant and item, and a within-participants manipulation of gaze cue.

3.2.4 Predictions

Our key behavioral prediction is that the distribution of attention to speakers vs. objects will shift over the course of learning. Early in learning, participants will allocate fixations that prioritize gathering visual information about the objects or about disambiguating reference in-the-moment of hearing a new word. After experiencing multiple exposures to a word-object pairing, learners will start to distribute fixations such that they support the goal of rapid comprehension of the incoming speech, showing the behavioral signatures of eye movements during familiar language processing that we observed in our prior work.

Prediction	Trial type	Analysis window	Measure(s)
Broader distribution of attention across objects early in learning, which declines over learning	learning	Prior to speaker appearing	Proportion looking and entropy
Early random first shifts, changing to language-driven shifts later in learning	learning	After noun onset	First shift accuracy; EWMA (higher proportion guessing)
Slower first shifts early in learning and in the gaze condition	learning	After noun onset	First shift RT
Show behavioral signatures of eye movements for familiar language comprehension later in the task, but earlier for the gaze condition	learning	After noun onset	First shift Accuracy with RT; HDDM (lower drift and boundary separation)
Higher proportion looking to the target object early in learning in the gaze condition	test	After noun onset	Proportion looking
Increase in target looking in the no-gaze condition later in the task	test	After noun onset	Proportion looking

Figure 9: Predictions for learning and test trials for different measures and analyses.

3.2.5 Open questions for design and analysis

- Should the test trials also include video of a speaker? This would allow us to do the first shift analyses on test trials, but it would create less of a separation between exposure and test.
- When should the gaze cue occur in the sentence: pre- vs. post-nominal? Pre-nominal might be more natural, but post-nominal makes certain analyses possible.
- Should we include a speaker-absent condition?
- Should we manipulate the informativeness or complexity of the objects?

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.

Gabry, Jonah, and Ben Goodrich. 2016. "Rstanarm: Bayesian Applied Regression Modeling via Stan." *R Package Version* 2 (1).

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.

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

Ratcliff, Roger, and Russ Childers. 2015. "Individual Differences and Fitting Methods for the Two-Choice Diffusion Model of Decision Making." *Decision* 2 (4). Educational Publishing Foundation:237–79.

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.

Vandekerckhove, Joachim, and Francis Tuerlinckx. 2007. "Fitting the Ratcliff Diffusion Model to Experimental Data." *Psychonomic Bulletin & Review* 14 (6). Springer:1011–26.

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.