Does social information reduce referential ambiguity in cross-situational word learning?:

Using eye-tracking to study the effect of eye gaze in adult word learners

*Allison Emily Dods*

Symbolic Systems Program

Stanford University

June 8, 2016

To the directors of the Program on Symbolic Systems:

I certify that I have read the thesis of Allison Emily Dods in its final form for submission and have found it to be satisfactory for the degree of Bachelor of Science with Honors.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Michael C. Frank

Psychology

To the directors of the Program on Symbolic Systems:

I certify that I have read the thesis of Allison Emily Dods in its final form for submission and have found it to be satisfactory for the degree of Bachelor of Science with Honors.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Hyowon Gweon

Psychology

# Acknowledgements

I am immensely grateful for the guidance and support of my advisor, Michael Frank, and my mentor, Kyle MacDonald. I would also like to thank my second reader, Hyowon Gweon, as well as Rose Schneider, the manager of the Language and Cognition Lab. Additionally, this thesis is made possible by a generous grant from Undergraduate Advising and Research and a stipend from the Symbolic Systems Program.

# [to add when finished: table of contents]

# Introduction

The study of first language acquisition – that is, the investigation of how young children learn to understand and communicate using language – has implications for a wide variety of disciplines. Among these are cognitive science, linguistics, developmental psychology, cognitive neuroscience, second language acquisition, philosophy of language, and the study of communication disorders. Understanding how children learn language, a uniquely human ability, can give us insight into both the structure of the mind and of language itself.

Language acquisition is, however, a difficult task. To fully understand and be fully understood, a child must be able to parse spoken language into distinct words; understand the meanings of those words (according to one estimate, 60,000 words by the age of 17 (Bloom 2002)); pronounce the sounds of the language(s) she speaks; learn the grammatical rules of the language(s) and be able to form and understand an infinite number of novel sentences; and use language in different social contexts. Finally, a typical first language learner must be able to do all of this and much more with little or no explicit instruction.

This paper focuses on word learning and, in particular, on the question of how first language learners come to understand the meanings of individual words. Unlike adult second language learners, who can study a vocabulary list with translations of words into their native language, first language learners must use other methods and resources to learn the meanings of many words. In this thesis, I will discuss several of these means and offer evidence from the literature for their existence and use. Finally, I present the findings of my own study, which uses an eye tracker to examine their viability as solutions to the word-learning problem of referential uncertainty.

## Referential uncertainty

Word learners face a number of challenges, including speech segmentation (e.g. Saffran et al., 1996) and category generalization. The present study investigates one such challenge, known as “referential uncertainty,” of mapping unknown words onto their referents (Quine, 1960). Specifically, we focus only on mapping words to concrete objects, although word learners must also learn the meanings of more intangible concepts such as verbs, abstract nouns, or adjectives.

To illustrate the problem, imagine a young child who has in front of her an apple, a banana, and an orange. She does not know the names for any of these foods. If her parent says, “Eat your apple,” the child does not know whether the word “apple” refers to the red food, the yellow food, or the orange food. As Quine points out, the word “apple” might even refer to part of a food, like a peal or a stem; or to fruit or food in general; or to the plate or the table; or even to something in another linguistic category, such as the adjective *orange* or the action of eating. How do word learners map a word to its referent object given the problem of referential uncertainty?

The following introductory section traces some of the responses in the literature to the problem of referential uncertainty, focusing on word learners’ tracking of cross-situational co-occurrence information as a potential way to solve this word-learning problem. I first briefly mention some early work that proposed ways in which referential uncertainty might be resolved in single naming events. I then discuss the more recent literature on cross-situational word learning and, in particular, the debate about the specific mechanisms that might underlie a cross-situational solution to referential uncertainty. I consider recent accounts that integrate early social theories of word learning with cross-situational theories to shed light on this debate. Because the present study uses eye-tracking in a novel way to further illuminate the underlying mechanisms of cross-situational word learning, I briefly review the ways in which eye-tracking has been used in other word-learning studies before presenting my hypotheses.

### “In the moment” reduction of referential uncertainty

Many researchers studying word learning between the late 1980s and the early 2000s focused on single naming events. As the term suggests, a single naming event is a singular instance in which a learner hears a novel word and must at that time identify the word’s referent, a process known as “fast mapping” (Yu & Smith, 2007). Though there could be an infinite number of possible referents for a given novel word, some researchers have suggested the existence of constraints that might help to narrow the possibilities and help the learner map the word to its referent. The present study examines specifically the contributions of social information to referential uncertainty reduction, but I here briefly give examples of proposed representational, attentional, and linguistic constraints as well, to illustrate a number of ways in which referential uncertainty may be resolved in-the-moment.

Markman (1990) suggests some possible representational constraints. She posits that even very early word learners might hold conceptual assumptions that bias them towards certain potential referents of novel words. For instance, she argues that the “whole object assumption” leads children to find it more likely that a novel word refers to an entire object instead of to part of that object or to one of its qualities. She also suggests the “mutual exclusivity assumption,” which she claims leads word learners to assume, when possible, that novel words do not map onto objects whose names are already known. These assumptions, Markman explains, work together such that learners first fulfill the whole object assumption and then the mutual exclusivity assumption, with exceptions for names of parts or groupings of things.

Linguistic information can help eliminate impossible potential referents of a novel word. Gleitman (1990) suggests a process called “syntactic bootstrapping”, through which the syntactic structure of an utterance places constraints upon the space of possible referents. In the above example, a word learner who is familiar with the meaning or use of the word “eat” can infer from the beginning of the command that “apple” is a type of food; this would rule out, for instance, the hypothesis that “apple” referred to the plate, since plates are inedible.

Finally, and most importantly for the current work, social information can also constrain the space of possible referents of a novel word. Recall our earlier example, in which a parent asks a child to “eat [her] apple” over a snack of an apple, a banana, and an orange, none of whose names the child knows. One can imagine the parent looking at, or perhaps even pointing at or touching, the apple as he makes this request. Paul Bloom argues that recognition of speaker intention via an understanding of theory of mind is crucial to children’s success in word learning (2002). In other words, in-the-moment reduction of the child’s referential uncertainty may depend in part on the child’s ability to understand what the parent means by looking at the apple as he pronounces the word “apple.” Additionally, social information, such as that provided by speaker gaze, is thought to have a different effect on word learning than mere salience (Yurovsky & Frank, 2015a). That is, social information may communicate more than simply a signal to look at one object or another.

Research on joint attention indicates that infants encode word-object links during what Baldwin and colleagues call “follow-in labeling” – that is, when both the infant and the adult are looking at an object while the adult labels it (e.g. Baldwin, 1991; Baldwin, 1993; Baldwin et al., 1996). In a paradigm used several times by Baldwin (1991, 1993) and colleagues (1996), experimenters employ “follow-in labeling,” and when tested for comprehension, infants perform better than at chance at identifying that object as the referent of the novel word. Additionally, infants did *not* encode a link between a novel word and object when the experimenter employed “discrepant labeling” – that is, looking at a *different* object than the infant did while pronouncing the novel word (Baldwin, 1991). (To illustrate, this would be as if the child was looking at the banana while the parent looked at the apple and said, “Eat your apple”.) In this situation, infants performed at chance on comprehension tests, indicating that they do not assume a link between the object in their line of sight and the spoken novel word when speaker gaze is directed elsewhere. Speaker gaze, then, can provide information that influences in-the-moment reduction of referential uncertainty.

However, word learners may not be able to *fully* resolve referential uncertainty in the moment. Even taking into account the constraints discussed above, single naming events in real life are often noisy and ambiguous (Yu & Smith, 2007; Smith & Yu, 2008). In addition to there often being many possible referents for a given novel spoken word, one must also take into account referents in a number of categories (not just nouns; also adjectives, verbs, etc.); multi-word referents; and referents that are either not physically present or completely intangible (like *love*) (Medina, Snedeker, Trueswell, & Gleitman, 2011). Additionally, while adults can reference syntactic, lexical, pragmatic, and other contextual clues when learning the meaning of a new word (for example, using Gleitman’s proposed process of *syntactic bootstrapping*), young infants have not all acquired the information needed to rely on such clues (Siskind, 1996; Smith & Yu, 2008; Smith, Suanda, & Yu, 2014).

Some or all of the above limitations of in-the-moment referential uncertainty reduction may be addressed by *cross-situational statistical word learning*, which is the focus of the present study. The following section discusses recent literature on cross-situational word learning.

## Cross-situational statistical word learning

In contrast to single naming events, cross-situational statistical word learning involves the aggregation of information across multiple word learning events or *situations*. Learners might use statistical information about the frequency or the distribution of words and objects to infer word-referent pairs over time. Recall our earlier example, in which a parent asks a child to eat her apple over a snack of an apple, a banana, and an orange. If the child does not know what the word “apple” refers to, this single naming event is ambiguous. Imagine that, the following day, the parents says the word “apple” again, this time having served the child an apple, a kiwi, and a grape. Because the word “apple” has co-occurred with the red food more often than with any others, the child has more information to indicate a link between “apple” and that object. This example is a simple illustration; in reality, word learning events may include hundreds of potential referents, and it may take many more than two naming events before a child learns the referent of a new word. In another subdomain of word learning, infants have been shown to use statistical information to differentiate spoken words within spoken sentences (Saffran et al., 1996), so it is plausible that a similar statistical mechanism might be used to map novel words to referents.

Both adults and children have demonstrated the ability to learn word-object pairs across multiple ambiguous naming situations (Yu & Smith, 2007; Smith & Yu, 2008; Voulousmanos, 2008; Kachergis, Yu, & Shiffrin, 2014; Escudero, Mulak, & Vlach 2015; Yu & Smith 2011). Until recently, formal computation models such as Siskind’s (1996) had suggested that cross-situational statistics were a viable solution to the problem of referential uncertainty, but Smith and Yu’s 2007 adult study was the first to examine cross-situational word learning in human learners. In their paradigm, participants saw training trials consisting of several pictures of objects and simultaneously heard several novel words corresponding to the number of objects. The words and objects were repeated over subsequent trials, with each word presented always in conjunction with the same object. No other information was given that might indicate the object to which a word mapped. Test trials at the end of the experiments measured whether participants had learned which words referred to which objects via forced-choice tests for adults (2007) and manually-coded looking times for children (2008). Both age groups learned more word-object relationships than would have been expected by chance.

There remain important questions about the potential use of cross-situational word learning. Recently, for instance, Smith, Suanda, and Yu (2014) questioned whether infants are able to learn words cross-situationally outside of laboratory settings, in noisier contexts. In this study, we focus on one particular question: how exactly cross-situational word learning works in a learner’s mind.

Over time, a debate has emerged between two ideas for the mechanisms that learners use to decide a word’s referent across situations (Smith, Suanda, & Yu, 2014).

The first idea, here referred to as *multiple-alternative tracking*, is that learners track multiple alternatives for the possible referent (e.g. McMurray, Horst, & Samuelson, 2012; Yu, 2008), and the second, *single-hypothesis tracking*, is that learners store a single strong hypothesis about the object a word refers to (Woodard, Gleitman, & Trueswell, 2016; Trueswell, Medina, Hafri, & Gleitman, 2013; Medina, Snedeker, Trueswell, & Gleitman, 2011). Much of the literature since then, including the present study, has focused on distinguishing between these processes, with recent attention directed toward the ways in which they might interact. In the immediately following section, I will explain some of the arguments and evidence that have been put forth first for multiple-alternative tracking, then for single-hypothesis tracking, and finally for integrative accounts of both ideas.

## Competing accounts of cross-situational word learning processes

### Multiple-alternative tracking

Multiple-alternative tracking, also known as “associative learning” (Smith, Suanda, & Yu, 2014), posits that learners keep track of multiple possible referents for a given word across naming situations. Specifically, the learner might track information about the statistical distribution of the possible referents, including how often the object occurs in the same setting as the spoken word. The account predicts that after some number of naming situations, one object will emerge as having co-occurred with the word more often than other objects; when asked about a word’s referent, learners will then choose this object.

To illustrate, consider again our example of the child and the fruit. If the child is tracking multiple alternatives, she implicitly keeps track of all of the possible things “apple” might refer to, and the frequency with which those objects co-occur with the word “apple.” The first day, the alternatives might include an apple, a banana, and an orange; the second, a kiwi, an apple, and a grape; and the third, a banana, an apple, and a grape. All of these fruits are possible referents of “apple”, but some are more likely than others. Because this is a toy example, we can calculate explicitly that the apple has appeared three times, the banana and grape twice, and the orange and kiwi once, so the apple is the most likely to be the referent of “apple”. In reality, the process might take place over much longer periods of time, with many more possible referents, and on an implicit level.

The multiple-alternative account predicts that word learners will demonstrate memory of multiple potential referents of a given word, even those that are not the most likely referent. Consistent with this prediction, Vouloumanos (2008) found that adult word learners differentiated between objects that were, for instance, 20% likely to be the referent of a word versus objects that were 10% likely to be the referent, even if neither object was statistically the *most* likely. This finding suggests that learners retain information about objects beyond simply the most likely referent. Kachergis, Yu, and Shiffrin (2014) found a similar phenomenon, noting that adult learners kept track of co-occurrences between the word and multiple objects, although this result was stronger when participants were explicitly asked to learn the meanings of words compared to an implicit task that did not contain that instruction. Associative cross-situational word-learning models have been proposed by McMurray, Horst, and Samuelson (2012) and by Yu (2008).

### Single-hypothesis tracking

Proponents of single-hypothesis tracking, also known as “propose-but-verify” (Woodard, Gleitman, & Trueswell, 2016; Trueswell et al., 2013), argue that learners form a single hypothesis about a word’s referent on the first naming events. Over subsequent naming events – that is, cross-situationally – the hypothesis is either strengthened by the continued co-occurrence of the word and hypothesized referent, or rejected and replaced with a new hypothesis. In the fruit example, if the child is tracking a single hypothesis, she might hypothesize the first time she hears the word “apple” that it refers to the orange food. If the orange food is present when she next hears “apple”, her hypothesis is strengthened. If it is absent, she must abandon this hypothesis and form a new one.

Single-hypothesis tracking bears a strong resemblance to the “fast mapping” approach discussed above, which research suggests is used in single naming events. To clarify, the single-hypothesis tracking account *combines* fast mapping on the first naming event (the “proposal” of the hypothesis) with amendments to the mapping on subsequent naming events (the “verification” of the hypothesis). As mentioned earlier in this introduction, a number of studies have shown that word learners use various conceptual, pragmatic, and linguistic constraints to correctly map some words to referents on single naming events (e.g. Carey, 1978; Baldwin, 1993). Trueswell and colleagues (2013) claim that the ability to fast-map a word to a referent on the first trial of multiple influences the storage of a single strong hypothesis across trials, until that hypothesis is disconfirmed by a subsequent trial.

Other arguments for the use of single-hypothesis tracking in cross-situational word learning come from Medina et al. (2011). In an attempt to more faithfully represent real-world word learning situations than in previous word-learning studies, they asked adults to watch muted vignettes of parents speaking to children, with one word in the vignette replaced by a novel nonsense word. Participants were asked what they thought the meaning of each nonsense word was after each of five vignettes, to track whether their hypotheses changed over time. Medina et al. found, contrary to what might be predicted by associative accounts, that participants’ accuracy in mapping words to referents did not improve across trials. They also observed that accuracy on the final guess depended on how informative the *first* trial had been, a finding consistent with the fast mapping process supposedly involved in single-hypothesis tracking.

Both adults and children have been shown to engage in behaviors expected of a single-hypothesis tracker (Medina et al., 2011; Trueswell et al., 2013; Woodard, Gleitman, & Trueswell, 2016). Trueswell et al. (2013) and Woodard, Gleitman, and Trueswell (2016) make use of an additional metric that is also used in the current study to identify behavior consistent with single-hypothesis tracking. [these studies use a setup that’s similar to the same/switch one in MacDonald & in the present study – not sure if I should explain them here?] A word learner who forms a single strong hypothesis might be predicted to fail to remember the other possible referents. Indeed, these studies found that both adults and children performed at chance when asked to identify the referent of a word from a set of objects that did not include their (incorrectly) hypothesized referent.

### Integrative accounts

While earlier research advocated for one or the other underlying mechanism of cross-situational word learning, recent work has focused on integrative accounts of the two. Yurovsky and Frank (2015b) found not only that learners appear to use both mechanisms during cross-situational word learning, but also that certain variables can be manipulated to make learners look more like either single-hypothesis or multiple-alternative trackers. In particular, the study manipulated both the number of potential referents for a given word (either 2, 3, 4, or 8) and the intervening trials between the learning and the test trial for a word (either 1, 2, 4, or 8). When more potential referents were presented, learners had worse memory for the objects they had not hypothesized as a word’s referent, having had to distribute their attention more widely during the learning phase. In contrast, when fewer objects were presented, learners could concentrate their attention among this smaller set and were better able to track multiple alternative potential referents.

MacDonald, Yurovsky, and Frank (2015) altered the paradigm used in Yurovsky and Frank (2015b) to include the presence of a social cue, namely the speaker’s eye gaze. As noted above, social information has been shown to modulate uncertainty in single naming events (Baldwin, 1991; Baldwin & Moses, 2001). MacDonald, Yurovsky, and Frank found that when adult participants watched a face turn its head toward one of the potential referent objects while hearing a nonsense word spoken, they were less likely to track multiple alternative possible referents and more likely to behave as if they had stored a single hypothesis.

## Current Study

The present study aims to replicate the key finding from MacDonald, Yurovsky, and Frank (2015) – namely, that the presence of eye gaze causes adult word learners to behave more like single-hypothesis trackers and less like multiple-alternative trackers. The current study uses the paradigm developed by Yurovsky and Frank (2015b) and amended to include social information in MacDonald, Yurovsky, and Frank (2015). I describe it briefly here and in more detail later in this paper.

Participants were shown a series of ambiguous word-learning trials. The trials consisted of a exposure phase, in which a speaker pronounced a novel word while participants saw two novel objects, and a subsequent test phase, in which the speaker again pronounced the word while participants saw a different set of two objects, one of which had appeared in the original set and one of which was new. Participants assigned to the *gaze* condition saw a woman’s face turn towards one of the objects during exposure trials, while the woman’s face was uninformative to participants in the *no-gaze* condition.

Unlike previous studies with similar designs the present study uses an eye-tracker both to examine participants’ behavior while learning words and to identify their confidence in the referent of a word during a testing phase. An eye-tracker non-invasively records the coordinates of participants’ eye gaze on a screen. Eye-tracking data is very precise (Huettig & Altmann, 2005), allowing for timecourse analyses of participant gaze. Further, using eye gaze as the indicator of the participant’s hypothesized referent removes the need for participants to make a forced choice, which obscures the participant’s potential uncertainty about the object she thinks the word refers to. Finally, eye-tracking allows for the study of younger populations for whom the word-learning question is more relevant (Fernald, Zangl, Portillo, & Marchman, 2008; Halberda, 2006). Eye-tracking has recently been used reliably to study cross-situational word learning in young children (Yu & Smith, 2011), although it has not yet been used to study the presence of social cue in cross-situational word learning, nor has it been used with the paradigm described in the section on integrative accounts above.

As mentioned above, the present study examines only adult word learning, with the hope that it will inform future work that seeks to use eye-tracking data to analyze the interaction between social information and cross-situational word learning in children. There is a strong precedent for testing cross-situational word learning paradigms on adult participants to identify an effect before later recruiting child participants. For instance, Yu and Smith first observed the potential ability to learn word-object links cross-situationally in adults (2007) before subsequently testing for the same effect, using a modified paradigm, in children (Smith & Yu, 2008). Other studies have worked only with adult participants on questions that are also relevant to child word learning (e.g. Medina et al., 2011; Yurovsky & Frank, 2015b).

I predict that the presence of a speaker’s eye gaze will reduce referential uncertainty and cause participants to behave more like single-hypothesis trackers. Empirically, I predict to see (1) that participants who see the social cue on learning trials will allocate more attention to the target of the speaker’s gaze than to other potential referents, and that participants who do not see the social cue will allocate attention equally across potential referent objects; (2) a quantitative relation between a learner’s time spent looking at a word’s referent during the testing phase and their allocation of attention to that object during the learning phase; and (3) a difference in the way word-object links are encoded across the two conditions.

Hypothesis (1) draws directly from theories of social information and joint attention in word learning, discussed above (e.g. Baldwin, 1991). According to these accounts, we can expect participants who observe speaker gaze to follow that gaze. Unlike in previous similar studies, the eye-tracker allows us to observe the proportion of time a participant spends looking at each object and to test whether the distribution of attention differs between participants who see the social cue and those who do not.

Hypothesis (2) is consistent with findings from MacDonald, Yurovsky, and Frank (2015) and Yurovsky and Frank (2015b). Namely, these studies found that decreased demands on attention during the learning phase of a similar paradigm, represented by fewer potential referents in Yurovsky and Frank’s (2015) study and by the presence of a social cue in MacDonald, Yurovsky, and Frank’s (2015b) study, were correlated with better memory for multiple alternative referents. In a similar vein, we expect to see a difference in test performance between participants in the two conditions, modulated by the difference in attention allocation predicted in hypothesis (1).

Finally, hypothesis (3) claims that there is something “special” about social information that affects the strength of word-object mappings in the learner’s mind beyond simply a difference in allocation of attention. This hypothesis is based on findings that suggest that social information communicates more to a learner than simply making an object more salient (Yurovsky & Frank, 2015a). We can measure this hypothesis by testing to see whether participants in different conditions who allocate the same amount of attention to an object perform differently; if they do, it will suggest that speaker gaze affects storage of word-object links beyond only influencing allocation of attention.

# Methods

## Participants

34 undergraduate students were recruited from the Stanford Psychology 1 credit pool. Of the participants, 17 were females and 17 were males. 31 were between the ages of 18 and 21, two were between the ages of 22 and 25, and one was below the age of 18. The students received course credit for participation. Four participants were excluded during analysis because of equipment error – the eye-tracker did not properly record their gaze coordinates. The final sample included 30 participants.

## Stimuli and Apparatus

Eye-tracking software from SensoMotoric Instruments (SMI) was used to design the study and to collect the coordinates of participants’ eye gazes. The design of the video was based on a paradigm introduced by Yurovsky and Frank (2015) and further developed by MacDonald, Yurovsky, and Frank (under review) to include the presence of a social cue. The participants viewed the video on a 1920x1080 laptop screen.

The experiment featured sixteen pseudowords recorded by an AT&T Natural VoicesTM speech synthesizer using the “Crystal” voice (a woman’s voice with an American English accent), as well as 48 novel objects represented by black-and-white drawings of fictional objects from Kanwisher, Woods, Iacoboni, and Mazziotta (1997). Sixteen words were used so that the experiment would be sufficiently long to make within-subject comparisons across trials, and 48 objects were used so that objects would not be repeated across trials. The trial design is explained in the subsection immediately below.

## Design and Procedure

Participants were seated with their faces about a foot away from a monitor and told they would watch a very short video, during which their eye movements would be recorded. They were asked to stay still and to keep their eyes on the screen. The experimenter then began the video, which lasted 2.6 minutes, and stepped away from the screen until the experiment was over.

The experiment consisted of a series of paired *exposure* and *test* trials, such that each exposure trial was immediately followed by a test trial. The first two pairs of trials were invariably *training* trials, followed by sixteen novel trials, whose order was randomized by the eye-tracking software, for a total of eighteen pairs of exposure and test trials. Furthermore, participants were randomly sorted into either the *gaze* or the *no-gaze* condition, explained in detail below.

On exposure trials, a woman’s face appeared on the screen above two novel objects, while a woman’s voice pronounced a pseudoword. In the *no-gaze* condition, the woman’s face looked straight ahead while the word was pronounced. In the *gaze* condition, the woman’s face turned toward one of the objects and then back to straight-ahead. The direction in which the woman’s face turned on a given trial in the gaze condition was random, but counterbalanced so that she turned towards the leftmost object eight times and the rightmost the other eight. A participant in the *gaze* condition was said to “follow gaze” if she spent a larger proportion of time on a given exposure trial looking at the object that was the target of gaze than at the other object.

On test trials, the woman’s face looked straight ahead in both conditions, while the voice repeated the same word from the immediately preceding exposure trial. One of the two objects from the exposure trial remained on the screen, while the other of the two objects was replaced by a third object that had not previously appeared in the video. To illustrate, if Objects A and B were displayed during an exposure trial, either Objects A and C or Objects B and C would be displayed during the following test trial. In what follows, the object that remained on the screen across both trials (in the illustrated case, Object A) is referred to as the “kept” object for convenience. A participant was said to “succeed” on a test trial if she spent a larger proportion of time looking at the kept object than at the new (non-kept) object.

The first two exposure/test trial pairs were *training* pairs. The objects in training trials were commonly recognizable objects, such as a squirrel or a cup, while the corresponding words were common English words that corresponded to an object on the screen, such as “squirrel” or “tomato”. The training trials were meant both to signal to the participant that the face was “labeling” objects on the screen and to check that participants were following the face’s gaze (in the *gaze* condition). Participants who did not reliably follow gaze on both training trials were excluded from the study. On the exposure training trials, the woman’s face looked straight ahead as the objects were labeled in the *no-gaze* condition, while in the *gaze* condition, she looked at one of the two objects, consistent with the later exposure trials. On the test training trials, the woman’s face looked straight ahead on both conditions.

The placements of the “kept” objects in both the exposure and test trials were counterbalanced throughout the experiment. Of the sixteen non-training test trials, eight “kept” the object that had been the leftmost of the two objects in the corresponding exposure trial. The other eight non-training test trials “kept” the object that had been on the right in the exposure trial. Additionally, the position of the “kept” object was counterbalanced such that it appeared on the left of the screen for half of the test trials and on the right for the other half of the test trials, independent of its position in the corresponding exposure trial. Finally, in the *gaze* condition, half of the test trials kept the object that was the target of the face’s gaze in the exposure trial, while the other half kept the object that had not been the target of gaze.

For the sake of analysis, the terms “Same trials” and “Switch trials” are used to dichotomize the test trials; the terms are used in a slightly different way than in MacDonald, Yurovsky, and Frank (2015), so a brief explanation is included here. “Same” refers to the test trials in which the kept object was also the object that the participant spent a larger proportion of time looking at during exposure; “Switch” refers to the test trials in which the object that was kept was the same object that the participant had looked at less during exposure. We hypothesized that participants in the *gaze* condition would allocate more attention and look more to the target of the face’s gaze on exposure; that they would thus be less likely to encode a link between the word and the non-target object; and that they would then perform well (that is, “succeed” often) on “Same” test trials and perform at chance (“succeed” 50% of the time) on “Switch” test trials. We predicted on the other hand that participants in the *no-gaze* condition would allocate equal attention to both objects on exposure and thus perform better than *gaze* condition participants on “Switch” trials.

# Results

The following analysis consists of two broad components. First, I examine performance on exposure trials in both the *gaze* and *no*-*gaze* conditions to demonstrate the effectiveness of the social cue; that is, to ensure that participants in the *gaze* condition reliably followed the face’s gaze. I then examine performance on test trials to identify any effects of condition. Recall that we hypothesized that participants in the *gaze* condition would look more to the target of the face’s gaze on exposure (“follow gaze”), thus allocating less attention to the non-target object, while participants in the *no-gaze* condition would divide attention more equally between the two objects on exposure and therefore be more likely than *gaze* condition participants to encode multiple potential word-object links. We expected to see evidence of these phenom­ena in a comparison of participants’ accuracy on test trials across conditions. In particular, we predicted a difference in performance on “Switch” test trials, in which the object the participant looked at less on exposure was the one “kept” on the test trial; we expected participants in the *gaze* condition to perform at chance in this situation, having failed to encode a word-object link for the “kept” object, and participants in the *no-gaze* condition to perform better than participants in the *gaze* condition, having tracked the “kept” object as an alternative referent of the word.

We used mixed-effects regression models to test some of our predictions. Namely, we predicted a quantitative relation between a participant’s amount of looking at the target object during exposure and his or her looking at the target object on the following test trial. We also predicted that condition would have a further effect on test performance; even when two participants looked the same amount at an object during exposure, we expected to see that the participant in the *gaze* condition would perform worse than the participant in the *no-gaze* condition on “Switch” test trials. Code for these models, along with the rest of our analysis code, is located at <https://github.com/langcog/gaze-xsit>.

## Exposure Trials

First, we checked to see whether participants in the *gaze* condition followed the face’s gaze by looking at the object that was the target of gaze. Figure 1 is a histogram showing the percentage of time on each individual exposure trial a participant spent following gaze. The distribution is skewed heavily right, and the mean proportion of gaze following is 77%, suggesting that participants did, for the most part, follow gaze on exposure.

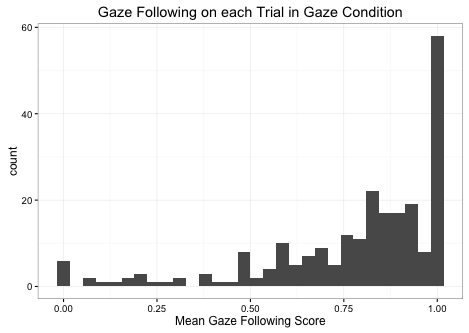


Figure 1. Proportion of time spent looking at the target of gaze on exposure trials.

The side-by-side histograms in Figure 2 display accuracy on each exposure trial. Accuracy is defined as the amount of time spent looking at the object that would subsequently be “kept” in the following test trial. As explained above in the Methods section of this paper, the “kept” object appeared 50% of the time as the leftmost object on exposure trials and 50% of the time as the rightmost object. Furthermore, on the *gaze* condition, the “kept” object was 50% of the time the target of the face’s gaze and 50% of the time the non-target object. The order of trials was randomized for each participant. The *gaze* condition histogram in Figure 2 has a bimodal distribution, which is consistent with predicted gaze-following behavior. The exposure trials in which participants had near 100% accuracy reflect the situation in which participants followed gaze to the “kept” object, allocating almost 100% of their attention to that object. The trials in which participants had near 0% accuracy reflect the situation in which the participants followed gaze to the object that was not “kept”, allocating almost none of their attention to the “kept” object. The *no-gaze* condition histogram acts as a control, displaying a roughly normal distribution, consistent with our prediction that participants in the *no-gaze* condition would distribute their attention roughly equally between the two objects on exposure. While allocating nearly 100% or 0% of one’s attention to the “kept” object is common in the *gaze* condition, it is rare in the *no-gaze* condition; on the other hand, it is common for participants in the *no-gaze* condition to allocate closer to 50% of attention to the “kept” object, and rare for participants in the *gaze* condition to do so.

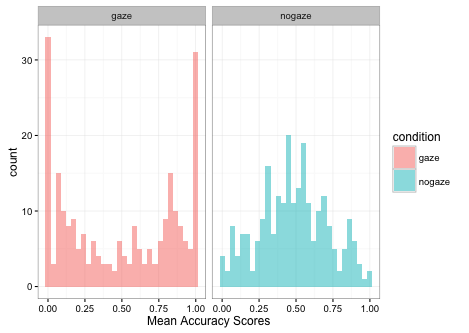


Figure 2. Proportion of time spent looking at the "kept" object on exposure trials in both conditions.

## Test Trials

Accuracy is defined for test trial analysis as the proportion of time spent looking at the “kept” object. Figure 3 shows the mean accuracy in both conditions across all test trials and participants, with the dotted line representing chance performance. The mean accuracy was .689 for the gaze condition and .659 for the no-gaze condition; there was no significant difference between the two means (t = 1.168).

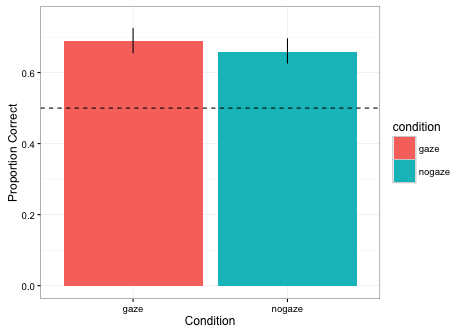


Figure 3. Mean accuracy on test trials in both conditions.

Figure 4 displays accuracy on test trials plotted as a function of accuracy on the corresponding exposure trial. For clarity, a linear regression is fitted to the data and a dotted line representing chance performance is shown. Figure 5 displays the same data as does Figure 4, but fitted with a local polynomial regression instead of a linear regression. The distribution of data points differs between conditions. As expected given the bimodal distribution of accuracy during exposure in the *gaze* condition and the roughly normal distribution of accuracy during exposure in the *no-gaze* condition, the data points for the *gaze* condition are concentrated towards the sides of the plot, while those for the *no-gaze* condition are concentrated towards the center.

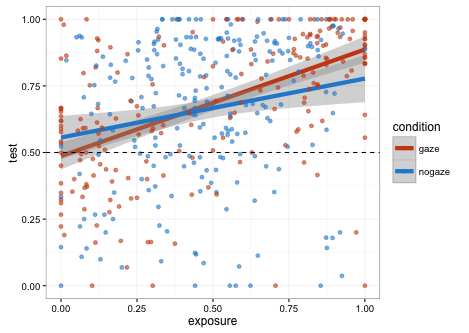


Figure 4. Performance on test trials as a function of performance on exposure trials, fitted with a linear regression.

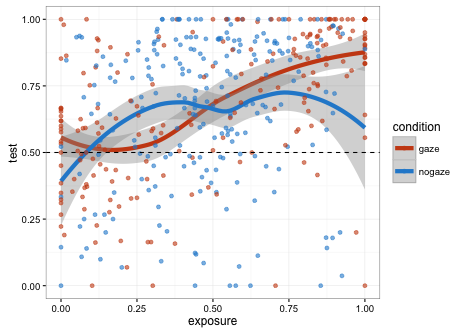


Figure 5. Performance on test trials as a function of performance on exposure trials, fitted with a local polynomial regression.

Both conditions show positive slopes overall, indicating a positive correlation between looking time at the “kept” object during exposure and looking time at the “kept” object during testing. Notably, in both conditions, complete failure to look at the “kept” object during exposure results in a mean performance around chance during testing. We quantified this relationship using the following mixed-effects linear regression model: **test ~ exposure \* condition + (1|subid)**. The output of the model is shown in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Predictor | Estimate | Standard Error | *t*-value |
| (Intercept) | 0.4725122 | 0.0373224 | 12.660299 |
| Exposure accuracy | 0.4160538 | 0.0437836 | 9.502501 |
| *No-gaze* condition | 0.0912118 | 0.0572694 | 1.592680 |
| Exposure accuracy\*  *No-gaze* condition | -0.2148968 | 0.0804835 | -2.670072 |

Table 1. Output of a linear mixed-effects model examining the relationship between exposure performance, condition, and test performance.

Table 1 reveals that there is a strong positive significant correlation between performance during exposure and performance during testing (*t* = 9.503), such that as participants look more at the “kept” object during exposure, they tend to look more at the “kept” object during testing. We also found a weak significant interaction between condition and performance during exposure (*t* = -2.670). This interaction indicates that, though the slopes of both the red and blue lines in Figure 4 are positive, the slope of the red line (representing the *gaze* condition) is slightly steeper than the slope of the blue line (representing the *no-gaze* condition). In other words, participants in the *gaze* condition demonstrated a slightly stronger correlation between looking at the “kept” object during exposure and looking at the “kept” object during testing than did participants in the *no-gaze* condition.

This interaction can be better understood by looking at Figure 6, which displays mean test accuracy in both conditions for each quartile of exposure accuracy. A dotted line at .5 represents chance performance during testing. In all but the lowest quartile, performance during testing was above chance in both conditions, and, most notably, performance during testing is slightly higher for the *gaze* condition than for the *no-gaze* condition only in the highest quartile (looking to the “kept” object during exposure more than 75% of the time).

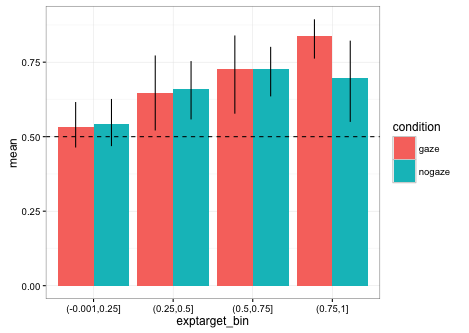


Figure 6. Mean test accuracy for both conditions, binned by quartile of exposure performance.

This relationship is demonstrated also in Table 2, which shows the output of the following mixed-effects linear regression model: **Test performance ~ Exposure performance binned by quartiles \* Condition + (1|subject)**. Note that this model is almost identical to the one whose output is displayed in Table 1, except the continuous “exposure” variable has been replaced by the discrete binned exposure variable.

|  |  |  |  |
| --- | --- | --- | --- |
| Predictor | Estimate | Standard Error | *t*-value |
| (Intercept) | 0.5086340 | 0.0367145 | 13.8537574 |
| (0.25, 0.5] (2nd quartile)  exposure accuracy | 0.0786732 | 0.0538247 | 1.4616573 |
| (0.5, 0.75] (3rd quartile)  exposure accuracy | 0.2643982 | 0.0523670 | 5.0489428 |
| (0.75, 1] (4th quartile)  exposure accuracy | 0.3379407 | 0.0373142 | 9.0566312 |
| *No-gaze* condition | 0.0345847 | 0.0604173 | 0.5724294 |
| *No-gaze* condition \* 2nd quartile exposure accuracy | 0.0401951 | 0.0740278 | 0.5429734 |
| *No-gaze* condition \* 3rd quartile exposure accuracy | -0.1108209 | 0.0732261 | -1.5134065 |
| *No-gaze* condition \* 4th quartile exposure accuracy | -0.1798464 | 0.0695677 | -2.5852011 |

Table 2. Output of a linear mixed-effects model examining the relationship between binned exposure performance, condition, and test performance.

Table 2 reveals significant effects of exposure accuracy between 0.5 and 0.75 (*t* = 5.049) and between 0.75 and 1 (*t* = 9.057) on test accuracy. That is, when participants looked at the “kept” object during exposure between 50% and 75% of the time, there existed a weak positive correlation between looking time to the “kept” object during exposure and during test, and when participants looked at the “kept” object during exposure between 75% and 100% of the time, there existed a moderate positive correlation between looking time to the “kept” object during exposure and during test. We can also observe a significant weak interaction between condition and performance in the upper quartile of exposure accuracy (*t* = -2.585), such that when participants looked at the “kept” object during exposure more than 75% of the time, *gaze* participants performed slightly better during testing than did *no-gaze* participants.

# Discussion

This study tested the primary hypothesis that the presence of social information during word learning reduces learners’ referential uncertainty, causing them to be more likely to track a single strong hypothesis for a word’s referent in cross-situational contexts. Unlike previous studies, it uses eye-tracking to examine the interaction between social information and cross-situational learning in resolving referential uncertainty. In this section, I discuss both the broader implications of our findings as well as ways in which the current study might address some limitations of previous investigations.

We found evidence to support the first of our hypotheses; namely, that allocation of attention between the two objects during exposure trials would differ for participants in different conditions. Specifically, as predicted, we found that participants in the *gaze* condition largely followed the speaker’s gaze, spending a disproportionate amount of time looking to the target of gaze and creating the bimodal distribution shown in Figure 2 for the *gaze* condition. Meanwhile, also as predicted, we found that participants in the *no-gaze* condition tended to divide attention more equally between the two objects, creating the roughly normal distribution shown in Figure 2 for the *no-gaze* condition.

Our second prediction was a quantitative relation between exposure accuracy and test accuracy. Evidence to support this prediction can be found in both Figure 4 and Table 1 of the Results section. In particular, the significant strong positive correlation between exposure accuracy and test accuracy across conditions is consistent with prior findings that attention allocated to an object during exposure is related to memory for that object later.

Finally, we hypothesized that participants in the *gaze* condition would perform at chance on “Switch” test trials, in which the object the participant looked at less during exposure was the one “kept” during testing, and that participants in the *no-gaze* condition would perform better than participants in the *gaze* condition. **[posthoc analysis]**

The current study does, of course, have certain potential limitations, some of which may have influenced our results and can hopefully be addressed in future work. Perhaps the most salient difference between this study and its immediate predecessors is the current study’s lack of explicit instruction and forced choice. That is, unlike Yurovsky and Frank (2015b) and MacDonald, Yurovsky, and Frank (2015), we did not tell participants to attempt to learn the meanings of words, or to make explicit guesses about a word’s referent via clicking on an object. Instead, participants were asked only to watch the video. On the one hand, the lack of forced choice makes it admittedly difficult to analyze the current results using the same framework as was used for previous similar studies; on the other hand, it avoids forcing the participant to make a hypothesis about a word’s referent. The lack of discrete choice allowed us to see, for instance, participant uncertainty reflected in divided attention to both objects during testing. Smith, Suanda, and Yu also point out that paradigms like the one used in the present study may be too simple to generalize to real-world word learning (2014).

[implications – to do once hypothesis section has been filled out entirely]

Future work, we hope, will take the framework used in this study in order to conduct more fine-grained tests of infant referential uncertainty reduction via a combination of cross-situational statistics and social information, leading to a better understanding of the ways in which word-object links are stored in word learners’ minds.