POLITE LANGUAGE REFLECTS COMPETING SOCIAL GOALS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF PSYCHOLOGY
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Erica J. Yoon

March 2019

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Michael C. Frank) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Ellen Markman)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Noah Goodman)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Hyowon Gweon)

Approved for the Stanford University Committee on Graduate Studies

Abstract

We use polite speech on a daily basis. From "thanks" and "please" to "your dress is cute" to "these cookies could use a bit of salt," people often produce polite utterances that are indirect or even false to some degree. Why do people speak politely? This thesis proposes a goal-based account of polite speech, that polite speech arises from a set of competing social goals: the speaker's desire to transfer information in the most truthful and informative manner possible ("informational goal"), and to abide by social norms and expectations and/or maintain the interactants' *face* or public self-image ("pro-social goal") and/or present herself as a particular kind of individual (e.g., kind, helpful person; "self-presentational goal"). In Chapter 1, I provide an overview of this integrative theoretical framework that aims to unify previous theoretical accounts of polite speech and explain existing empirical studies on understanding and production of polite speech in adults and children. In Chapter 2, I provide a computational model that formalizes the notion of goals as utilities that speakers try to maximize through language use, and show that this model successfully captures adults' predictions and judgments for polite lies and indirect speech. Then I present two sets of empirical studies looking at the development of polite language understanding: 2- to 4-year-old children's judgments for polite requests (Chapter 3) and 5- to 8-year-old children's judgments for polite lies versus blunt truths (Chapter 4). Overall, the work presented in this thesis reveals how adults and children's understanding of polite speech reflects their understanding of speaker goals to be informative and social.

Dedication

FIXME

Acknowledgments

FIXME

Contents

List of Tables

List of Figures

Introduction

We use and hear polite speech on a daily basis, ranging from simple words of apology ("sorry") or gratitude ("thanks") to compliments ("I love your dress!") and requests ("Can you please open the window?"). Adults and even young children spontaneously produce requests in polite forms (Axia & Baroni, 1985; H. H. Clark & Schunk, 1980). Speakers exhibit politeness strategies even while arguing, preventing unnecessary offense to their interactants (Holtgraves, 1997). Listeners even attribute ambiguous speech to a polite desire to hide a truth that could hurt another's self-image (e.g., Bonnefon, Feeney, & Villejoubert, 2009). In fact, it is difficult to imagine human speech that efficiently conveys only the truth. Intuitively, politeness is one prominent characteristic that differentiates human speech from stereotyped robotic communication, which may try to follow rules to say "please" or "thank you" yet still lack genuine politeness.

Although language users use polite speech on a daily basis, explaining why we use polite speech or how we understand it is not as straightforward as it first seems. While very simple polite utterances can be produced from straightforward rules (e.g., say "sorry" when you did something bad to someone), When speakers want to tell the listener to "close the window," they often use a more roundabout way and say "can you please close the window?" When people see that their interactant is wearing a new outfit that they think is hideous, they might still say "Your dress looks gorgeous!" As such, polite utterances often seem to misrepresent their intended message or conceal the truth, which shows that polite speech violates a critical principle of cooperative communication: exchanging information efficiently and accurately (Grice, 1975).

If politeness only gets in the way of effective information transfer, why be polite? Clearly, there are social concerns, and most linguistic theories assume utterance choices are motivated by these concerns, couched as either polite maxims (Leech, 1983), social norms (Ide, 1989), or aspects of a speaker and/or listener's identity, known as *face* (P. Brown & Levinson, 1987; Goffman, 1967). All of these theories use different approaches to explain polite language, and some are even framed as counterarguments to existing theories (e.g., see Richard J Watts (2003) and Matsumoto (1988) responding to some issues in P. Brown & Levinson (1987)). One possible commonality among these theories however, is that they all describe ways in which language communication deviates from certain expected utterances or conversations due to speakers' social concerns.

In this thesis, my goal is to offer an integrative theoretical framework that aims to unify these existing theories, and provide empirical evidence in support of this framework. Specifically, I argue for a *goal-based* theory of polite speech: that polite utterances arise from competing social goals that speakers have, such as their desires to convey information as truthfully and efficiently as possible ("informational goal"), to make the listeners feel happy and respected and thereby boost or maintain their face ("prosocial goal"), and to present speakers themselves in a good light (e.g., that they are kind and helpful; "presentational goal"). Speakers then have to consider the tradeoff between these goals, and think about which goal to prioritize and how much to do so to determine their utterance.

For example, imagine that Alice and Bob are having a conversation and Bob asks for Alice's feedback on his cookies that he baked ("How did you like my cookies?") and Alice thinks the cookies tasted bad and salty (Figure 1, top panel). Alice's utterance would differ depending on her goals: whether she wants to prioritize informational goal or telling the truth to Bob; social goal or making Bob feel happy; or presentational goal or presenting Alice herself in a good light that she is kind (Predictions of this specific scenario will be explained in detail in Chapter 2).

The contents of this dissertation will be as follows, as shown in Figure 1: In Chapter 1 (top panel of Figure 1) I present an integrative goal-based framework that aims to explain polite speech based on the idea that it reflects a tradeoff between competing social goals that speakers have. Then using this framework, I will explain existing empirical studies on understanding and production of polite speech in adults and children. Chapters 2-4 describe a set of computational and empirical studies of children and adult's understanding of polite language (bottom panels of Figure 1) . In Chapter 2, I provide a computational model that formalizes the notion of goals as utilities that speakers try to maximize through language use, and show that this model successfully captures adults' predictions and judgments for polite lies and indirect speech. Then I present two sets of empirical studies looking at the development of polite language understanding: Chapter 3 examines 2- to 4-year-old children's judgments for polite requests, and Chapter 4 looks at 5- to 8-year-old children's judgments for polite lies versus blunt truths.

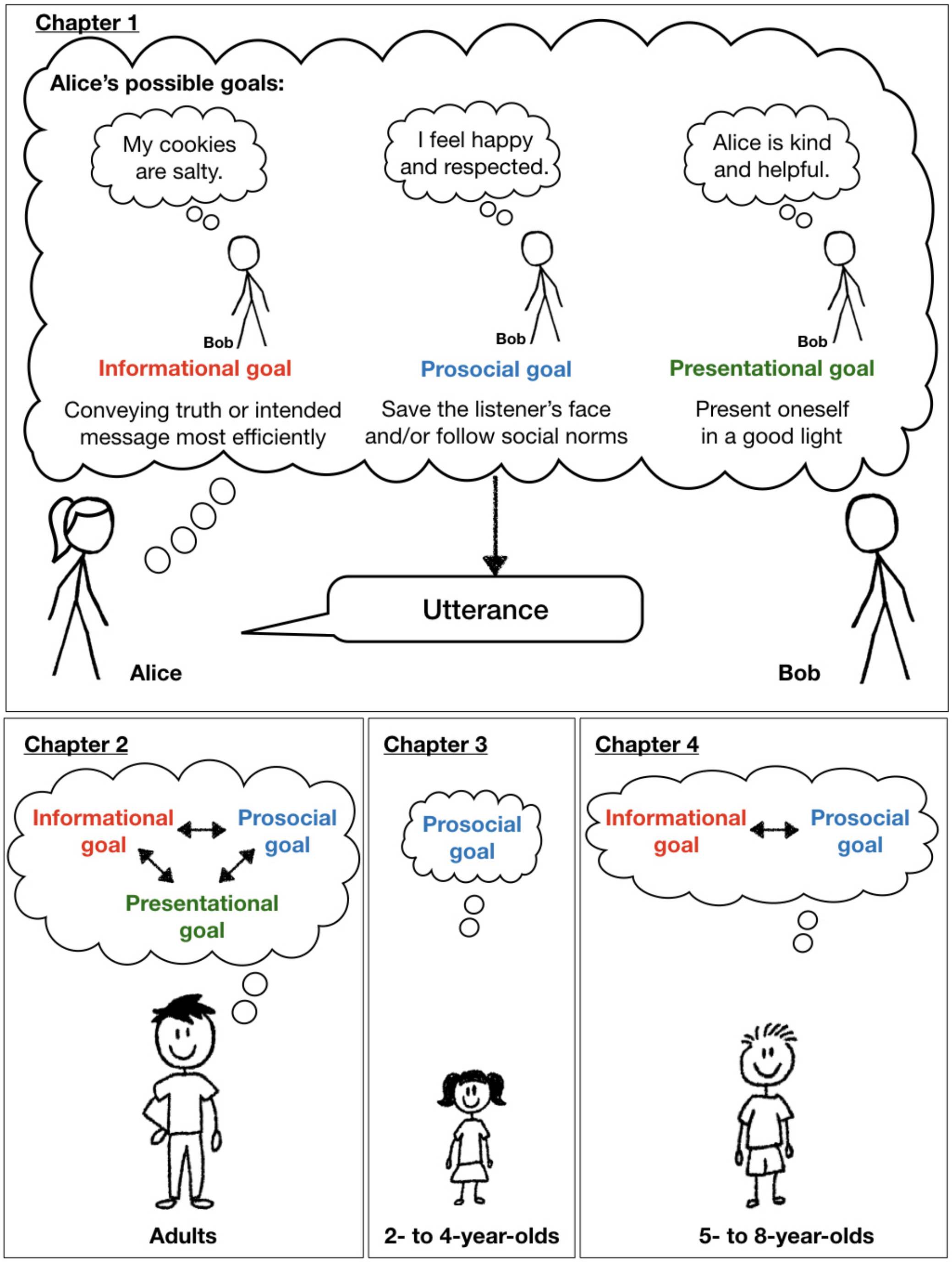


Figure 1: The upper panel shows a schematic overview of an integrative framework of polite language understanding based on competing social goals. The lower panels show different studies examining adults' and children's understanding of different component goals (and possible tradeoffs between them) that correspond to each chapter of the dissertation.

Chapter 1

A goal-based account of polite language

1.1 Introduction

Imagine that a stranger on the street approached you and asked: "I'm sorry to bother you, but can you tell me the way to the city hall by any chance?" Regardless of your answer, you probably would not feel puzzled or offended by the way in which the stranger decided to seek information that he needs. This is in contrast with an alternative situation where the stranger said to you instead: "Tell me the way to the city hall." In such case you would immediately notice the lack of politeness in his utterance, and your response to him may be negatively affected by the irritating oddity of the situation. Similarly, a person who noticed her friend's new, flashy dress could comment: "Your new dress is gorgeous!" even if the speaker actually felt that the dress was merely decent or even ugly. But why? Why are people expected to speak politely, when there are alternative utterances that can convey information about the world or the speaker's intention more directly ("Tell me the way") or truthfully ("Your dress is ugly")?

Language is simultaneously both an informational and social tool. Through language, people communicate information about the world, but also form their social relationships and establish their identity within the society. On one hand, some theories of language functions describe language as a transmission device for transferring from a sender to a receiver information that reflects context or the state of affairs (BÃijhler, 1934; Jakobson, 1960; Shannon, 1948). The importance of informativity is further emphasized in more recent, influential theories on pragmatics of natural language, which explain how meanings beyond literal meanings of utterances arise (Grice, 1975; Searle, 1975). On the other hand, some linguistic theories, especially those with references to language development, identify social roles of language that people use to make contact with others and form relationships (Ervin-Tripp, 1967; Halliday, 1975). These theories underscore that linguistic rules that language users tend to follow represent the norms and structure of the community using the language (Ervin-Tripp, 1969).

Could polite speech reflect both the informational and the social roles of linguistic communication? Previous theoretical accounts for polite speech vary in their focus on informational versus social aspects of polite language. Some theories view polite speech as reflecting social rules and norms (Richard J. Watts, Ide, & Ehlich, 1992), some as abiding by communicative maxims that people are expected to follow in conversations, to be both as informative and as armative of their conversational partner as possible (Lakoff, 1973; Leech, 1983), and others as performing face management, or trying to maintain interactants' good public self-image or reputation (P. Brown & Levinson, 1987).

In this Chapter, I propose that that polite speech highlights both informational and social uses of language: Polite speech reflects a principled tradeoff between the informational, epistemic content a speaker wants to convey (e.g. "I want to know the way to the city hall") and other social concerns, such as prosocial or self-presentational goal that the speaker wants to express for herself and others ("I'm not rudely commanding you to tell me the answer, but making a request in a respectful way"). Thus, my goal is to unify previous theoretical frameworks in one, goal-driven account of polite speech. In what follows, I will describe the goal-based account of polite speech in detail (Part I), and summarize previous models and theories of polite speech and situate them within the framework of the current goal-driven account (Part II). Then I will examine empirical evidence for goal-driven approach to polite speech: In Part III, I will focus on empirical work on adult production and comprehension of polite speech which show that adults reason about speaker goals in polite speech; and in Part IV, I will probe empirical evidence from developmental work that children's production and understanding of polite speech advance as they grow older. In doing this I will show that children's production and comprehension of polite speech is related to the relative complexity of polite speech based on its goal tradeoff.

1.2 Part I: A Goal-based account of politeness

What does it mean to speak politely? Common instances of polite speech that occur to one's mind probably include the simplest politeness markers, such as "please," "thanks," and "sorry." More complicated examples would involve ways in which, for example, a person make a request: under normally conceivable circumstances, it would certainly be more polite to ask "Would it be too much trouble to ask you to complete this survey when you're not too busy?" than to say "Do this survey now." The word "polite" can sometimes carry a negative undertone in its meaning, as in "she was just being polite," which is likely to mean the speaker was hiding her genuine intentions to make the listener feel good. From these examples, we can identify a few aspects that a polite utterance may exhibit: observance of social rules, relatively high degree of elaborateness or indirectness, and dishonesty or disingenuousness in the interest of others' feelings or reputations.

More formal definitions of the term politeness also reveal common features that polite speech shows. Cambridge and Oxford Dictionaries respectively define what it means to be polite: "behaving in a way that is socially correct and shows respect for other people's feelings"; and "courteous, behaving in a manner that is respectful or considerate of others; well-mannered" ("Polite," 2017a, 2017b). Similar to the previous examples of polite speech, these definitions suggest politeness involves (1) observance of social expectations and (2) respect for others. Boyer (1702)'s The English Theophrastus: of the Manners of the Age, compilation of texts describing the English life in the early eighteenth century, identifies a purpose in trying to be polite: "Politeness may be defined as a dextrous management of our Words and Actions whereby *men make other people have a better Opinion of us and themselves* [emphasis added]." Thus, according to the Theophrastus definition, speakers speak politely in order to boost the self-images of the interactants (both the speakers themselves and their addressees).

These common themes of politeness have been identified by previous theoretical accounts of polite speech (reviewed in detail in Part II). But each of the accounts only focused on certain aspects of polite speech but disregarded others, and their explanations for politeness have been viewed as largely disparate. Here I make a unification proposal, where these existing approaches to polite speech can be united under a single goal-based account of polite speech.

I propose that polite speech reflects some degree of tradeoff between three main communicative goals that a speaker has: informational, prosocial, and (self-)presentational. *Informational goal* has to do with the speaker's desire to convey the most accurate information in the most efficient manner. *Prosocial goal* is about the speaker's desire to retain the listener's acceptable self-image as a decent individual and as a reputable member of society. *Presentational goal* reflects the speaker's desire to present the speaker herself in a good light, to appear to be a kind and helpful individual. Below I describe each goal in more detail.

1.2.1 Informational goal.

A speaker's informational goal, i.e. to prioritize information transfer in communication, may involve two closely related notions: informativity and truthfulness. Informativity is the notion of conveying the intended meaning in the most efficient and precise manner possible. The idea of informativity here is similar to Grice (1975)'s notion of cooperativity: A speaker will cooperatively choose utterances such that the listener can understand her intended message. Thus, the current notion of informativity that I adopt will encompass the whole Cooperative Principle (CP) that Grice posited ("Make your contribution such as required by the purposes of the conversation at the moment"), and especially the Maxim of Quantity ("Make your contribution as informative as is required"), though Maxims of Relevance ("Be relevant") and Manner ("Be perspicuous"; i.e. be brief, orderly, and unambiguous) can also be relevant to the notion of informativity as I discuss in this Chapter.

The idea of informativity has been formalized in probabilistic (Bayesian) models as a utility function of a speaker with particular goals in mind. The "rational speech act" (RSA) theory of language understanding (see N. D. Goodman & Frank (2016a) for a review) assumes that listeners expect speakers to aim to be helpful yet parsimonious, choosing their utterances approximately optimally based on a communicative goal (e.g., inform the listener) and interpret an utterance by inferring what the helpful speaker meant based on the utterance and any other relevant information about the world. The theory defines a standard, informative utility as the amount of information a literal listener (L_0) would still not know about world state (s) after hearing a speaker's utterance (w):

$$U_{epistemic}(w; s) = ln(P_{L_0}(s|w))$$

where the utterance choice is approximately rational (i.e., in proportion to the expected utility gain) and w is chosen from a set of alternative, relevant utterances.

For example, if Bob asked Alice "How was my cookie?" and Alice said to Bob "It was good," with only the goal to be informative in mind, Bob would think that Alice was being maximally informative by using the word "good" instead of another relevant, stronger word such as "amazing," and infer that Ann meant "good but not amazing" because otherwise Ann would have used the word "amazing" instead.

I note here that Ann's speech act could be analyzed as having observed the Gricean Maxim of Quantity, by making her utterance maximally informative, and Bob's inference as being based on such assumption that Ann's utterance is as informative as is required to meet Bob's needs. But as the comparison between the former goal-directed analysis and this latter maxim-based analysis may reveal, the maxim-based account is difficult to formalize (Hirschberg, 1985) whereas the goal-directed view allows for quantitative account of factors contributing to the linguistic phenomena at hand (N. D. Goodman & Frank, 2016a). From here on, I take the goal-directed view rather than the Gricean maxim-based view of the speaker; thus, analyses of speakers and their utterances will be based on speakers' communicative goals to convey information, etc., rather than their observation or violation of Gricean maxims.

Informational goal can also involve truthfulness: the meaning that accessed by the listener should match the true state of the world as closely as possible. In other words, the speaker will want to convey what is true (to the extent of her knowledge), not what is false. For example, if Bob baked some cookies and they tasted terrible, and Alice had the goal to be truthful, she would want to convey what matches the true state of the world as closely as possible ("Your cookies tasted terrible"). On the other hand, if Alice remarked to Bob about his cookies "Your cookies tasted good," with goal to be informative and truthful, the true state of the world must be such that Bob's presentation was truly good (but probably not amazing, because otherwise Alice would have said that it was amazing), and not bad or terrible.

As for when speakers decide not to tell the truth, there is a difference between violating versus flouting of truthfulness (which Grice also discusses; Grice (1975), p. 49, 53). Flouting involves contradicting common knowledge shared between the speaker and listener about the true state of the world, such that the listener notices that the utterance meaning does not match the true state. For example, after Alice and Bob together watch a movie that was obviously gory and disturbing to both of them, if Bob says "well, that was a really happy, fluffy movie!" then his utterance would be flouting truthfulness goal, and Ann would recognize Bob's utterance as an ironic one. On the other hand, violating truthfulness does not hold assumption of such common knowledge of the true state of the world, and thus the listener may not notice the mismatch between the utterance meaning and true state of the world, although the listener can potentially access it through other means (e.g. realizing that the speaker had reasons to lie). Here I will mainly focus on violation, not flouting, of truthfulness, for example when speakers tell white lies (i.e., when the speakers do not intend that the listeners know what the truth is).

Speakers' informational goal to be informative and truthful may encompass communicative co-operation in both locutionary and perlocutionary senses. A locutionary goal deals with conveying the intended meaning of the utterance within a conversation, whereas a perlocutionary goal involves achieving the speaker's ultimate goals toward the listener (Attardo, 1997). What could be an informational goal in its perlocutionary sense? In being truthful, speakers may ultimately want to maintain their moral obligation to tell the truth to others. This obligation is in line with Western philosophers' argument throughout the history that it is morally wrong to lie (Augustine, 1952; Kant, 1949), although there have been debates on whether the degree of wrongness may depend on context (e.g., if the speaker was telling a white lie; Sweetser, 1987). For example, if Alice said to Bob "Your cookies were good," Alice's locutionary goal would be to convey to Bob her intended meaning that his presentation was good (but perhaps not amazing), whereas her perlocutionary goal would be for Bob to think that the presentation was good, which was (apparently) the truth; Alice thereby upholds her moral obligation to tell the truth to Bob. Alice's goal to be truthful, then, is achieved in both locutionary and perlocutionary senses.

In the next section, I present an overview of the empirical work in the dissertation, linking it to the integrative framework discussed in this chapter. I also highlight the relevant pieces of the broader framework that the empirical work addresses. Importantly, these case studies tend to focus on eye movements that gather task-related information to guide future action, as opposed to exploratory eye movements in a novel environment. This feature makes them a good candidate for the active-social framework inspired by OED since these models were developed to explain information seeking in the context of well-defined tasks and clear learning goals.

1.2.2 Links between case studies and the broader framework

Chapter 2 investigates children's decisions about where to look while comprehending a visual-manual language (American Sign Language: ASL). In the grounded sign processing context, children must use visual fixations to gather information about both the referents and the linguistic signal, creating competition for visual attention that is not present in spoken language comprehension. This competition could fundamentally change how children decide to gather visual information during real-time language comprehension. Using our active-social theoretical framework, we might hypothesize that ASL-learners query the visual world in different ways such as fixating longer on social partners before seeking a named object since this behavior provides access to language, which can help guide future actions. Alternatively, it could be that ASL learners show parallel looking patterns to children acquiring spoken language, shifting their gaze to gather visual information about the objects in the scene soon after they have sufficient information about the identity of the named object. Chapter 2 explores these questions by measuring the timing and location of ASL learners' eye movements as they process sentences that contain familiar signs referring to objects in the visual scene.

Chapter 3 builds on the sign language work by directly comparing the gaze dynamics of children learning ASL to children learning spoken English during familiar language processing. We measure when children decide to stop fixating on a social partner to seek named referents in the visual world. Using our integrative framework, we hypothesized that looking to a signer returns higher value information for the specific task of identifying the named object, which would lead signers to generate slower but more consistent gaze shifts away from the language source and to the rest of the visual world. Chapter 3 also describes a study of English-learning children and adult's eye movements in noisy auditory contexts. We hypothesized that looking to a speaker becomes more useful when the auditory signal is less reliable, which, as in the ASL case, would make children fixate on the language source longer to gather more information, leading them to produce more language-consistent gaze shifts.

Chapters 4 and 5 explore how the information seeking framework generalizes to processing words in the context of social cues to reference. Social cues are an interesting test case for our active-social account because communicative partners who produce reliable cues to reference might become more valuable targets for children's looking. Thus, we hypothesized that in the context of reliable social cues, children should fixate longer on their social partner, modulating the information stored from a labeling event. In Chapter 4, we present a series of large-scale word learning experiments, showing that the presence of a social cue reduces the number of word-object hypotheses that adults remember from a labeling event. Chapter 5 describes a set of eye-tracking studies that measure when children decide to stop fixating on a social partner as a function of (1) whether the social partner provides a cue to reference and (2) their knowledge of the word-object mappings.

Overall, the goal of the empirical work is to ask how children's gaze patterns adapt to a diverse set of contexts that change the value of seeking information from social partners. The common thread across this research is that children's real-time visual information seeking is quite flexible and can adapt to the informational structure of the environment to support language processing.

Chapter 2

Children's distribution of visual attention during real-time American Sign Language comprehension[1]

This chapter presents a study of eye movements in a visual-manual language (American Sign Language: ASL). ASL is a compelling case because children use visual fixation to gather information about referents and the linguistic signal. In contrast, children learning spoken language can look away from their social partners while still gathering more linguistic information through the auditory channel. Within the broader active-social framework, this study asks whether ASL-learners' visual information seeking actions (i.e., queries) look dramatically different given the constraints of processing a visual language in real time or by having differential access to auditory information in day-to-day life.

To answer this question, we measured eye movements during real-time ASL comprehension of 29 native ASL-learning children (16-53 mos, 16 deaf, 13 hearing) and 16 fluent deaf adult signers. All

[1]This chapter is published in MacDonald, LaMarr, Corina, Marchman, & Fernald (2018). Real-time lexical comprehension in young children learning American Sign Language. *Developmental science*, e12672.

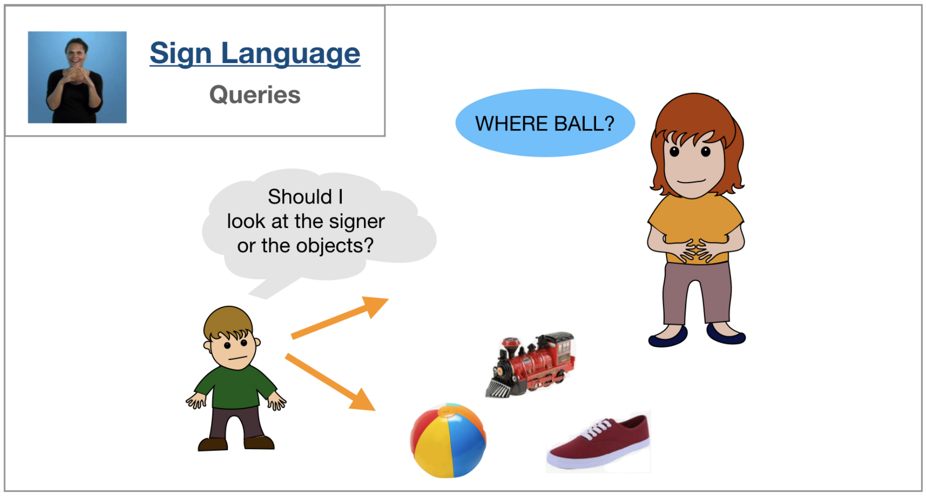


Figure 2.1: A schematic showing the components of the active-social learning framework addressed by the case study in Chapter 2.

signers showed evidence of incremental language comprehension, tending to initiate an eye movement before sign offset. Moreover, Deaf and hearing ASL-learners showed remarkably similar gaze patterns. Finally, variation in children's ASL processing was positively correlated with age and vocabulary size. These results suggest that, despite competition for attention within a single modality, signers will use visual attention to rapidly gather information about signs and seek named objects in ways that parallel spoken language processing. Moreover, these results suggest that the timing and accuracy of visual fixations in ASL reflect language-relevant information processing skills.

2.1 Introduction

Finding meaning in a spoken or a signed language requires learning to establish reference during real-time interaction – relying on audition to interpret spoken words, or on vision to interpret manual signs. Starting in infancy, children learning spoken language make dramatic gains in their efficiency in linking acoustic signals representing lexical forms to objects in the visual world. Studies of spoken language comprehension using the looking-while-listening (LWL) procedure have tracked developmental gains in language processing efficiency by measuring the timing and accuracy of young children's gaze shifts as they look at familiar objects and listen to simple sentences (e.g., "Where's the ball?") naming one of the objects (Fernald, Zangl, Portillo, & Marchman, 2008; Law & Edwards, 2014; Venker, Eernisse, Saffran, & Ellis Weismer, 2013). Such research finds that eye movements to named objects occur soon after the auditory information is sufficient to enable referent identification, and often prior to the offset of the spoken word (Allopenna, Magnuson, & Tanenhaus, 1998). Moreover, individual differences in the speed and accuracy of eye movements in response to familiar words predict vocabulary growth and later language and cognitive outcomes (Fernald, Perfors & Marchman, 2006; Marchman & Fernald, 2008). Together, these results suggest that gaze shifts to objects in response to spoken language reflect a rapid integration of linguistic and visual information, and that variability in the timing of these gaze shifts provides researchers a way to measure the efficiency of the underlying integration process.

Much less is known about how language influences visual attention during sign language comprehension, especially in young learners. Given the many surface-level differences between signed and spoken languages, it is not immediately clear whether the findings from spoken language will generalize to signed languages or whether they are specific to mechanisms of language comprehension in the auditory modality. In particular, studies with children learning spoken languages find that these skills undergo dramatic developmental changes over the 2nd and 3rd years of life. Moreover, there are significant relations between variation in efficiency in online language processing, as indexed by language-driven eye movements, and measures of linguistic achievement, such as vocabulary size and scores on standardized tests (Fernald et al., 2006; Marchman & Fernald, 2008). Will individual variation in language processing among children learning a signed language also be related to their age and vocabulary outcomes, as observed in children learning a spoken language?

Here we address this question by developing precise measures of speed and accuracy in real-time sign language comprehension by children learning American Sign Language (ASL). First, we estimate the extent to which adults and children tend to shift visual attention to a referent and away from the language source prior to the offset of a sign naming an object in the visual scene. Will signers wait until the end of the signed utterance, perhaps to reduce the probability of missing upcoming linguistic information? Or will signers shift gaze incrementally as the signs unfold in time, initiating saccades soon after there is enough information in the signal to identify the referent, similar to children and adults processing spoken language? Another related possibility is that signers would produce incremental gaze shifts to the named objects while still monitoring the linguistic signal in the periphery. This analysis provides an important first step towards validating the linking hypothesis that eye movements generated in our task reflect efficiency of sign recognition, rather than some other process, such as attending to the objects after the process of sign comprehension is complete. If children and adults produce rapid gaze shifts prior to target sign offset, this would provide positive evidence of incremental ASL processing.

Next, we compare the time course of ASL processing in deaf and hearing native ASL-learners to ask whether having the potential to access auditory information in their day-to-day lives would change the dynamics of eye movements during ASL processing. Do deaf and hearing native signers show parallel patterns of looking behavior driven by their similar language background experiences and the in-the-moment constraints of interpreting a sign language (i.e., fixating on a speaker as a necessary requirement for gathering information about language)? Or would the massive experience deaf children have in relying on vision to monitor both the linguistic signal and the potential referents in the visual world result in a qualitatively different pattern of performance compared to hearing ASL learning, e.g., waiting until the end of the sentence to disengage from the signer? This analysis is motivated by prior work that has used comparisons between native hearing and deaf signers to dissociate the effects of learning a visual-manual language from the effects of lacking access to auditory information (e.g., Bavelier, Dye, & Hauser, 2006).

Finally, we compare timing and accuracy of the eye movements of young ASL-learners to those of adult signers, and ask whether there are age-related increases in processing efficiency that parallel those found in spoken languages. We also examine the links between variability in children's ASL processing skills and their expressive vocabulary development. A positive association between these two aspects of language proficiency, as previously shown in children learning spoken languages, provides important evidence that skill in lexical processing efficiency is a language-general phenomenon that develops rapidly in early childhood, regardless of language modality.

2.1.1 ASL processing in adults

Research with adults shows that language processing in signed and spoken languages is similar in many ways. As in spoken language, sign recognition is thought to unfold at both the lexical and sub-lexical levels. Moreover, sign processing is influenced by both lexicality and frequency; non-signs are identified more slowly than real signs (Corina & Emmorey, 1993) and high frequency signs are recognized faster than low frequency signs (Carreiras, GutiÃľrrez-Sigut, Baquero, & Corina, 2008). Recent work using eye-tracking methods found that adult signers produce gaze shifts to phonological competitors, showing sensitivity to sub-lexical features, and that these shifts were initiated prior to the offset of the sign, showing evidence of incremental processing (Lieberman, Borovsky, Hatrak, & Mayberry, 2015). In addition, Caselli and Cohen-Goldberg (2014) adapted a computational model, developed for spoken language (Chen & Mirman, 2012), to explain patterns of lexical access in sign languages, suggesting that the languages share a common processing architecture.

However, differences between spoken and signed languages in both sub-lexical and surface features of lexical forms could affect the time course of sign recognition (for reviews, see Carreiras, 2010 and Corina & Knapp, 2006). For example, Emmorey and Corina (1990) showed deaf adults repeated video presentations of increasingly longer segments of signs in isolation and asked them to identify the signs in an open-ended response format. In the same study, English-speaking adults heard repeated presentations of increasingly longer segments of spoken words. Accurate identification of signs required seeing a smaller proportion of the total sign length compared to words (see also Morford & Carlsen, 2011), suggesting that features of visual-manual languages, such as simultaneous presentation of phonological information, might increase speed of sign recognition. Moreover, Gutierrez and colleagues (2012) used EEG measures to provide evidence that semantic and phonological information might be more tightly linked in the sign language lexicon than in the spoken language lexicon. Thus there is evidence for both similarities and dissimilarities in the processes underlying spoken-word and manual-sign recognition. However, with a few exceptions (e.g. Lieberman et al., 2015, 2017), most of this work has relied on offline methods that do not capture lexical processing as it unfolds in time during naturalistic language comprehension. In addition, no previous studies have characterized how young ASL-learners choose to divide visual attention between a language source and the nonlinguistic visual world during real-time language comprehension.

2.1.2 Lexical development in ASL

Diary studies show that ASL acquisition follows a similar developmental trajectory to that of spoken language (Lillo-Martin, 1999; Mayberry & Squires, 2006). For example, young signers typically produce recognizable signs before the end of the first year and two-sign sentences by their 2nd birthday (Newport & Meier, 1985). And as in many spoken languages (Waxman et al., 2013), young ASL-learners tend first to learn more nouns than verbs or other predicates (Anderson & Reilly, 2002).

However, because children learning ASL must rely on vision to process linguistic information and to look at named objects, it is possible that basic learning processes, such as the coordination of joint visual attention, might differ in how they support lexical development (Harris & Mohay, 1997). For example, in a study of book reading in deaf and hearing dyads, Lieberman, Hatrak, and Mayberry (2015) found that deaf children frequently shifted gaze to caregivers in order to maintain contact with the signed signal. Hearing children, in contrast, tended to look continuously at the book, rarely shifting gaze while their caregiver was speaking. This finding suggests that the modality of the linguistic signal may affect how young language learners negotiate the demands of processing a visual language while simultaneously trying to fixate on the referents of that language.

This competition for visual attention in ASL could lead to qualitatively different looking behavior during real-time ASL comprehension, making the link between eye movements and efficiency of language comprehension in ASL less transparent. On the one hand, demands of relying on vision to monitor both the linguistic signal and the named referent might cause signers to delay gaze shifts to named objects in the world until the end of the target sign, or even the entire utterance. In this case, eye movements would be less likely to reflect the rapid, incremental influence of language on visual attention that is characteristic of spoken language processing. Another possibility is that ASL-learners, like spoken language learners, will shift visual attention as soon as they have enough linguistic information to do so, producing saccades prior to the offset of the target sign. Evidence for incremental language processing would further predict that eye movements during ASL processing could index individual differences in speed of incremental comprehension, as previously shown in spoken languages.

2.1.3 Research questions

Adapting the LWL procedure for ASL enables us to address four questions. First, to what extent do children and adult signers shift their gaze away from the language source and to a named referent prior to the offset of the target sign? Second, how do deaf and hearing ASL-learners compare in the time course of real-time lexical processing? Third, how do patterns of eye movements during real-time language comprehension in ASL-learners compare to those of adult signers? Finally, are individual differences in ASL-learners' processing skill related to age and to expressive vocabulary development?

2.2 Study

2.2.1 Methods

Participants were 29 native, deaf and hearing ASL-learning children (17 females, 12 males) and 16 fluent adult signers (all deaf), as shown in Table 1. Since the goal of the current study was to document developmental changes in processing efficiency in native ASL-learners, we set strict inclusion criteria. The sample consisted of both deaf children of deaf adults and hearing Children of Deaf Adults (CODAs), across a similar age range. It is important to note that all children, regardless of hearing status, were exposed to ASL from birth through extensive interaction with at least one caregiver fluent in ASL and were reported to experience at least 80% ASL in their daily lives. Twenty-five of the 29 children lived in households with two deaf caregivers, both fluent in ASL. Although the hearing children could access linguistic information in the auditory signal, we selected only ASL-dominant learners who used ASL as their primary mode of communication both within and outside the home (10 out of 13 hearing children had two deaf caregivers). Adult participants were all deaf, fluent signers who reported using ASL as their primary method of communication on a daily basis. Thirteen of the 16 adults acquired ASL from their parents and three learned ASL while at school.

Our final sample size was determined by our success over a two-year funding period in recruiting and testing children who met our strict inclusion criteria – receiving primarily ASL language input. It is important to note that native ASL-learners are a small population. The incidence of deafness at birth in the US is less than .003%, and only 10% of the 2-3 per 1000 children born with hearing loss have a deaf parent who is likely to be fluent in ASL (Mitchell & Karchmer, 2004). In addition to the 29 child participants who met our inclusion criteria and contributed adequate data, we also recruited and tested 17 more ASL-learning children who were not included in the analyses, either because it was later determined that they did not meet our stringent criterion of exposure to ASL from birth (n = 12), or because they did not complete the real-time language assessment due to inattentiveness or parental interference (n = 5).

Table 2.1: Age (in months) of hearing and deaf ASL-learning participants

Hearing status	n	Mean	SD	Min	Max
deaf	16	28.0	7.5	16	42
hearing	13	29.4	11.2	18	53
all children	29	28.6	9.2	16	53

Measures

Expressive vocabulary size: Parents completed a 90-item vocabulary checklist, adapted from Anderson and Reilly (2002), and developed specifically for this project to be appropriate for children between 1.5 and 4 years of age. Vocabulary size was computed as the number of signs reported to be produced by the child.

ASL Processing: Efficiency in online comprehension was assessed using a version of the LWL procedure adapted for ASL learners, which we call the Visual Language Processing (VLP) task. The VLP task yields two measures of language processing efficiency, reaction time (RT) and accuracy. Since this was the first study to develop measures of online ASL processing efficiency in children of this age, several important modifications to the procedure were made, as described below.

Procedure

The VLP task was presented on a MacBook Pro laptop connected to a 27” monitor. The child sat on the caregiver’s lap approximately 60 cm from the screen, and the child’s gaze was recorded using a digital camcorder mounted behind the monitor. To minimize visual distractions, testing occurred in a 5’ x 5’ booth with cloth sides. On each trial, pictures of two familiar objects appeared on the screen, a target object corresponding to the target noun, and a distracter object. 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.

Table 2.2: Iconicity scores (1 = not iconic at all; 7 = very iconic) and degree of phonological overlap (out of 5 features) for each sign item-pair. Values were taken from ASL-LEX, a database of lexical and phonological properties of signs in ASL.

Item Pair (iconicity score 1-7)	**Number of matched features**	**Matched features**
bear (3.0) – doll (1.2)	1	Movement
cat (4.6) – bird (4.5)	3	Selected Fingers, Major Location, Sign Type
car (6.2) – book (6.7)	4	Selected Fingers, Major Location, Movement, Sign Type
ball (5.7) – shoe (1.5)	4	Selected Fingers, Major Location, Movement, Sign Type

Coding and Reliability. Participants' gaze patterns were video recorded and later coded frame-by-frame at 33-ms resolution by highly-trained coders blind to target side. On each trial, coders indicated whether the eyes were fixated on the central signer, one of the images, shifting between pictures, or away (off), yielding a high-resolution record of eye movements aligned with target noun onset. Prior to coding, all trials were pre-screened to exclude those few trials on which the participant was inattentive or there was external interference. To assess inter-coder reliability, 25% of the videos were re-coded. Agreement was scored at the level of individual frames of video and averaged 98% on these reliability assessments.

Stimuli

Linguistic stimuli. To allow for generalization beyond characteristics of a specific signer and sentence structure, we recorded two separate sets of ASL stimuli. These were recorded with two native ASL signers, using a different alternative grammatical ASL sentence structures for asking questions (see Petronio and Lillo-Martin, 1997):

- Sentence-initial wh-phrase: "HEY! WHERE [target noun]?"
- Sentence-final wh-phrase: "HEY! [target noun] WHERE?"

Each participant saw one stimulus set which consisted of one ASL question structure, with roughly an even distribution of children across the two stimulus sets (16 saw sentence-initial wh-phrase structure; 13 saw the sentence-final wh-phrase structure). To prepare the stimuli, two female native ASL users recorded several tokens of each sentence in a child-directed register. Before each sentence, the signer made a hand-wave gesture commonly used in ASL to gain an interlocutor's attention before initiating an utterance. These candidate stimuli were digitized, analyzed, and edited using Final

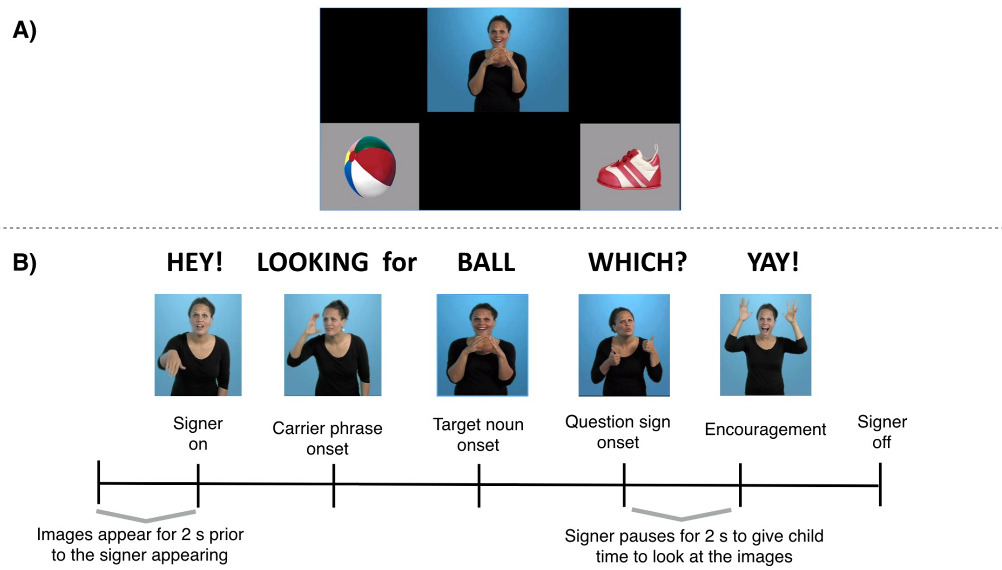


Figure 2.2: Configuration of visual stimuli (1A) and trial structure (1B) for one question type (sentence final wh-phrase) shown in the central video on the VLP task.

Cut Pro software, and two native signers selected the final tokens. The target nouns consisted of eight object names familiar to most children learning ASL at this age.

Visual stimuli. The visual stimuli consisted of colorful digitized pictures of objects corresponding to the target nouns presented in four fixed pairs (cat—bird, car—book, bear—doll, ball—shoe). See Table 2 for information about the degree of phonological overlap in each item-pair and the degree of iconicity for each sign (values were taken from ASL-LEX [Caselli et al., 2017]).[2] Images were digitized pictures presented in fixed pairs, matched for visual salience with 3–4 tokens of each object type. Each object served as target four times and as distracter four times for a total of 32 trials. Side of target picture was counterbalanced across trials.

Trial Structure

Figure 1 shows the structure of a trial with a sentence-final wh-phrase, one of the two question types in the VLP task. On each trial, children saw two images of familiar objects on the screen for 2 s before the signer appeared, allowing time for children to inspect both images. Next, children

[2]We did not find evidence that these features were related to the speed or accuracy of participants' eye movements in our task. However, this study was not designed to vary these features systematically. This analysis is presented in the Appendix for this chapter.

saw a still frame of the signer for one second, so they could orient to the signer prior to sentence onset. The target sentence was then presented, followed by a question and 2-s hold, followed by an exclamation to encourage attention to the task. This structure is nearly identical to the auditory LWL task, differing only in the addition of the 2-s hold. The hold was included to give participants additional time to shift gaze from the signer to the objects.

Calculating measures of language processing efficiency

Computing target sign onset and offset. In studies of spoken language processing, target word onset is typically identified as the first moment in the auditory signal when there is acoustic evidence of the target word. However, in signed languages like ASL, phonological information is present in several components of the visual signal simultaneously – for example, in one or both hands as well as in the face of the signer - making it difficult to determine precisely the beginning of the target sign. Because sign onset is critical to operationalizing speed of ASL comprehension in this task, we applied an empirical approach to defining target-sign onset. We used a gating task in which adult signers viewed short videos of randomly presented tokens that varied in length. Two native signers first selected a sequence of six candidate frames for each token, and then 10 fluent adult signers unfamiliar with the stimuli watched videos of the target signs in real-time while viewing the same picture pairs as in the VLP task. Participants indicated their response with a button press. For each sign token, the onset of the target noun was operationalized as the earliest video frame? at which adults selected the correct picture with 100% agreement. To determine sign offset, two native signers independently marked the final frame at which the handshape of each target sign was no longer identifiable. Agreements were resolved by discussion. Sign length was defined as sign offset minus sign onset (Median sign length was 1204 ms, ranging from 693-1980 ms).

Reaction Time. Reaction time (RT) corresponds to the latency to shift from the central signer to the target picture on all signer-to-target shifts, measured from target-noun onset. We chose cutoffs for the window of relevant responses based on the distribution of children's RTs in the VLP task, including the middle 90% (600-2500 ms) (see Ratcliff, 1993). Incorrect shifts (signer-to-distracter [19%], signer-to-away [14%], no shift [8%]) were not included in the computation of median RT. The RT measure was reliable within participants (Cronbach's $\alpha = 0.8$).

Target Accuracy. Accuracy was the mean proportion of time spent looking at the target picture out of the total time looking at either target or distracter picture over the 600 to 2500 ms window from target noun onset. We chose this window to be consistent with the choice of the RT analysis window. This measure of accuracy reflects the tendency both to shift quickly from the signer to the target picture in response to the target sign and to maintain fixation on the target picture. Mean proportion looking to target was calculated for each participant for all trials on which the participant was fixating on the center image at target-sign onset. To make accuracy proportion scores more suitable for modeling on a linear scale, all analyses were based on scores that were scaled in log space using a logistic transformation. The Accuracy measure was reliable within participants (Cronbach's $\alpha = 0.92$)

Proportion Sign Length Processed Prior to Shifting. As a measure of incremental processing, we used the mean proportion of the target sign that children and adults saw before generating an initial eye movement away from the central signer. Because target signs differed in length across trials, we divided each RT value by the length of the corresponding target sign. Previous research on spoken language suggests that at least 200 ms is required to program an eye-movement (Salverda, Kleinschmidt, & Tanenhaus, 2014), so we subtracted 200 ms from each RT to account for eye movements that were initiated during the end of the target sign (proportion target sign = $\frac{(RT - 200\text{ms})}{\text{Sign Length}}$). Mean proportion of sign processed was computed for each token of each target sign and then averaged over all target signs within participants, reflecting the amount of information signers processed before generating an eye movement, on average. A score of ≥ 1.0 indicates that a signer tended to initiate eye movements to the target pictures after sign offset. An average < 1.0 indicates eye-movements were planned during the target sign, reflecting the degree to which signers showed evidence of incremental language processing.

2.2.2 Analysis Plan

We used Bayesian methods to estimate the associations between hearing status, age, vocabulary, and RT and accuracy in the VLP task. Bayesian methods are desirable for two reasons: First, Bayesian methods allowed us to quantify support in favor of a null hypothesis of interest – in this case, the absence of a difference in real-time processing skills between age-matched deaf and hearing ASL learners. Second, since native ASL learners are rare, we wanted to use a statistical approach that allowed us to incorporate relevant prior knowledge to constrain our estimates of the strength of association between RT/accuracy on the VLP task and age/vocabulary.

Concretely, we used prior work on the development of real-time processing efficiency in children learning spoken language (Fernald et al., 2008) to consider only plausible linear associations between age/vocabulary and RT/accuracy, thus making our alternative hypotheses more precise. In studies with adults, the common use of eye movements as a processing measure is based on the assumption that the timing of the first shift reflects the speed of their word recognition (Tanenhaus, Magnuson, Dahan, & Chambers, 2000).[3] However, studies with children have shown that early shifts are more likely to be random than later shifts (Fernald et al., 2008), suggesting that some children's shifting behavior may be unrelated to real-time ASL comprehension. We use a mixture-model to quantify the probability that each child participant's response is unrelated to their real-time sign recognition (i.e., that the participant is responding randomly, or is "guessing"), creating an analysis model where participants who were more likely to be guessers have less influence on the estimated relations between RT and age/vocabulary. Note that we use this approach only in the analysis of RT, since "guessing behavior" is integral to our measure of children's mean accuracy in the VLP task, but not to our measure of mean RT. The Supplemental Material available online provides more details about the analysis model, as well two additional sensitivity analyses, which provide evidence that our results are robust to different specifications of prior distributions and to different analysis windows. We also provide a parallel set of analyses using a non-Bayesian approach, which resulted in comparable findings.

To provide evidence of developmental change, we report the strength of evidence for a linear model with an intercept and slope, compared to an intercept-only model in the form of a Bayes Factor (BF) computed via the Savage-Dickey method (Wagenmakers et al., 2010). To estimate the uncertainty around our estimates of the linear associations, we report the 95% Highest Density Interval (HDI) of the posterior distribution of the intercept and slope. The HDI provides a range of plausible values and gives information about the uncertainty of our point estimate of the linear association. Models with categorical predictors were implemented in STAN (Stan Development Team, 2016), and models with continuous predictors were implemented in JAGS (Plummer, 2003). Finally, we chose the linear model because it a simple model of developmental change with only two

[3] The assumption that first shifts reflects speed of incremental word recognition depends on the visual display containing candidate objects with minimal initial phonological overlap. If there are phonological competitors present (e.g., candy vs. candle), then participants' early shifting behavior could reflect consideration of alternative lexical hypotheses for the incoming linguistic information.

parameters to estimate, and the outcome measures – mean RT and Accuracy for each participant – were normally distributed. All of the linear regressions include only children's data and take the form: *processing measure* *age* and *processing measure* *vocabulary*.

2.2.3 Results

The results are presented in five sections addressing the following central questions in this research. First, where do ASL users look while processing sign language in real-time? Here we provide an overview of the time course of looking behavior in our task for both adults and children. Second, would young ASL-learners and adult signers show evidence of rapid gaze shifts that reflect lexical processing, despite the apparent competition for visual attention between the language source and the nonlinguistic visual world? In this section, we estimate the degree to which children and adults tended to initiate eye-movements prior to target sign offset, providing evidence that these gaze shifts occur prior to sign offset and index speed of incremental ASL comprehension. Third, do deaf and hearing native signers show a similar time course of eye movements, despite having differential access to auditory information in their daily lives? Or would deaf children's daily experience relying on vision to monitor both the linguistic signal and the potential referents in the visual world result in a qualitatively different pattern of performance, e.g., their waiting longer to disengage from the signer to seek the named object? Fourth, do young ASL-learners show age-related increases in processing efficiency that parallel those found in spoken languages? Here we compare ASL-learners' processing skills to those of adult signers and exploring relations to age among the children. Finally, is individual variation in children's ASL processing efficiency related to the size of their productive ASL vocabularies?

Overview of looking behavior during real-time ASL comprehension

The first question of interest was where do ASL users look while processing sign language in real-time? Figure 2 presents an overview of adults (2A) and children's (2B) looking behavior in the VLP task. This plot shows changes in the mean proportion of trials on which participants fixated the signer, the target image, or the distracter image at every 33-ms interval of the stimulus sentence. At target-sign onset, all participants were looking at the signer on all trials. As the target sign unfolded, the mean proportion looking to the signer decreased rapidly as participants shifted their gaze to the

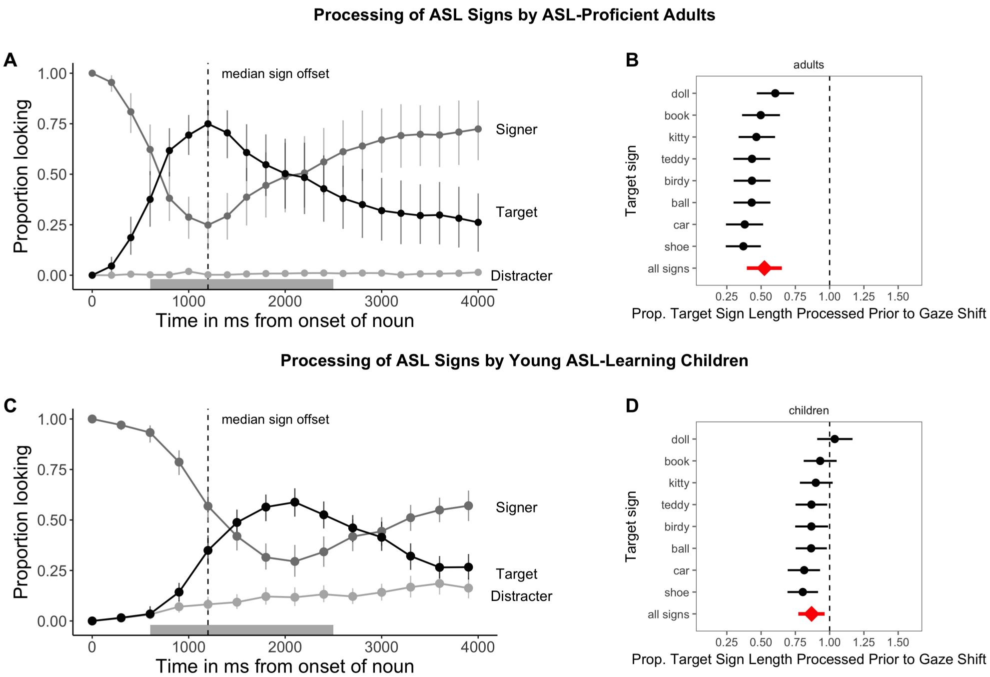


Figure 2.3: The time course of looking behavior for ASL-proficient adults (2A) and young ASL-learners (2C). The curves show mean proportion looking to the signer (dark grey), the target image (black), and the distracter image (light grey). The grey shaded region marks the analysis window (600-2500ms); error bars represent +/- 95% CI computed by non-parametric bootstrap. The mean proportion of each target sign length (see the Methods section for details on how sign length was defined) 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. Error bars represent 95% Highest Density Intervals.

target or the distracter image. Proportion looking to the target increased sooner and reached a higher asymptote, compared to proportion looking to the distracter, for both adults and children. After looking to the target image, participants tended to shift their gaze rapidly back to the signer, shown by the increase in proportion looking to the signer around 2000 ms after target-noun onset. Adults tended to shift to the target picture sooner in the sentence than did children, and well before the average offset of the target sign. Moreover, adults rarely looked to the distractor image at any point in the trial. This systematic pattern of behavior – participants reliably shifting attention from the signer to the named object and back to the signer – provides qualitative evidence that the VLP task is able to capture interpretable eye movement behavior during ASL comprehension.

Evidence that eye movements during ASL processing index incremental sign comprehension

One of the behavioral signatures of proficient spoken language processing is the rapid influence of language on visual attention, with eye movements occurring soon after listeners have enough information to identify the named object. Our second question of interest was whether young ASL-learners and adult signers would also show evidence of rapid gaze shifts in response to signed language, despite the apparent competition for visual attention between the language source and the nonlinguistic visual world. Or would signers delay their shifts until the very end of the target sign, or even until the end of the utterance, perhaps because they did not want to miss subsequent linguistic information?

To answer these questions, we conducted an exploratory analysis, computing the proportion of each target sign that participants processed before generating an eye movement to the named object. Figure 2 shows this measure for each target sign for both adults (2B) and children (2D). Adults shifted prior to the offset of the target sign for all items and processed on average 51% of the target sign before generating a response (M = 0.51, 95% HDI [0.35, 0.66]). Children processed 88% of the target sign on average, requiring more information before shifting their gaze compared to adults. Children reliably initiated saccades prior to the offset of the target sign overall (M = 0.88, 95% HDI [0.79, 0.98]) and for five out of the eight signed stimuli.

These results suggest that young signers as well as adults process signs incrementally as they unfold in time (for converging evidence see Lieberman et al., 2015, 2017). It is important to point out that we would not interpret signers waiting until the end of the sign or the end of the sentence as evidence against an incremental processing account since there could be other explanations for that pattern of results such as social norms of looking at a person until they finish speaking. However, this result provides positive evidence that eye movements in the VLP task provide an index of speed of incremental ASL comprehension, allowing us to perform the subsequent analyses that estimate (a) group differences in looking behavior and (b) links between individual variation in speed and accuracy of eye movements during ASL processing and variation in productive vocabulary.

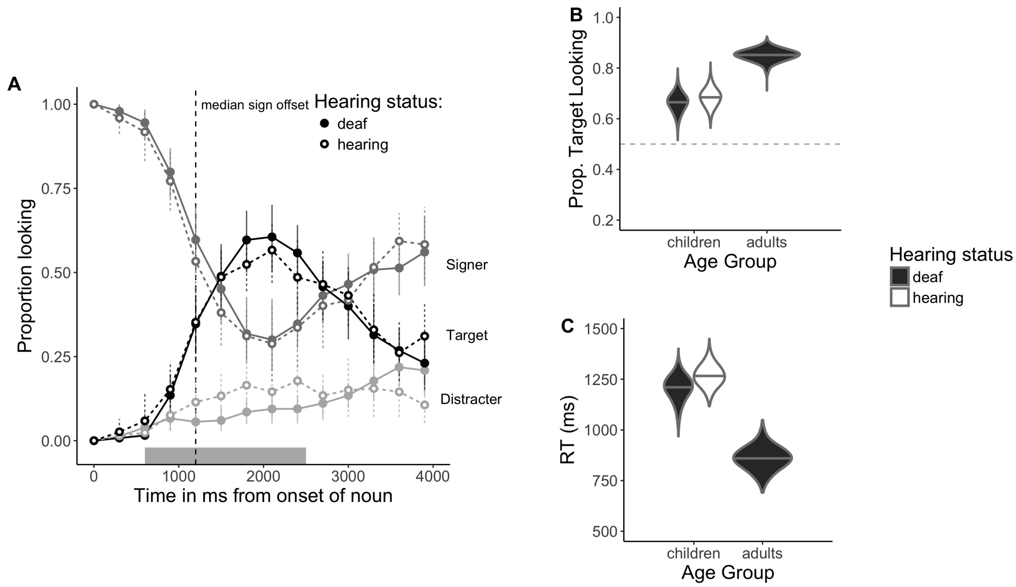


Figure 2.4: The time course of looking behavior for young deaf and hearing ASL-learners (3A). Filled circles represent deaf signers, while open circles represent hearing signers; All other plotting conventions are the same as in Figure 2. Panels B and C show full posterior distributions over model estimates for mean Accuracy (3B) and Reaction Time (3C) for children and adults. Fill (white/black) represents children's hearing status. (Note that there were no hearing adult signers in our sample).

Real-time ASL comprehension in deaf and hearing children and deaf adults

The third question of interest was whether deaf and hearing native signers show a similar time course of lexical processing, driven by their similar language experiences and the in-the-moment constraints of interpreting a sign language in real time? Or would deaf children's daily experience relying on vision to monitor both the linguistic signal and the potential referents in the visual world result in a qualitatively different pattern of performance, e.g., their waiting longer to disengage from the signer to seek the named object?

Figure 3A presents the overview of looking behavior for deaf and hearing children. At target-sign onset, all children were looking at the signer on all trials. Overall, deaf and hearing children showed a remarkably similar time course of looking behavior: shifting away from the signer, increasing looks to the target, and shifting back to the signer at similar time points as the sign unfolded. To quantify any differences, we compared the posterior distributions for mean accuracy (Figure 3B) and mean RT (Figure 3C) across the deaf and hearing groups. We did not find evidence for a difference in mean accuracy ($M_{hearing} = 0.68$, $M_{deaf} = 0.65$; $\beta_{diff} = 0.03$, 95% HDI $[-0.07, 0.13]$) or RT ($M_{hearing} = 1265.62$ ms, $M_{deaf} = 1185.05$ ms; $\beta_{diff} = 78.32$ ms, 95% HDI $[-86.01ms, 247.04ms]$), with the 95% HDI including zero for both models. These parallel results provide evidence that same-aged hearing and deaf native ASL-learners showed qualitatively similar looking behavior during real-time sentence processing, suggesting that decisions about where to allocate visual attention are not modulated by differential access to auditory information, but rather are shaped by learning ASL as a first language (see Bavelier et al., 2006 for a review of the differential effects of deafness compared to learning a visual language on perception and higher-order cognitive skills). Moreover, these results provide additional justification (over and above children's highly similar language background experience) for analyzing all the native ASL-learning children together, regardless of hearing status, in the subsequent analyses. Next, we compared real-time processing efficiency in ASL-learners and adult signers. Returning to the overview of looking behavior shown in Figure 2, we see that adults tended to shift to the target picture sooner in the sentence than did children, and well before the average offset of the target sign. Moreover, adults rarely looked to the distractor image at any point in the trial. To quantify these differences we computed the full posterior distribution for children and adults' mean Accuracy (Figure 3B) and RT (Figure 3C). Overall, adults were more accurate ($M_{adults} = 0.85$, $M_{children} = 0.68$, $\beta_{diff} = 0.17$, 95% HDI for the difference in means [0.11, 0.24]) and faster to shift to the target image compared to children ($M_{adults} = 861.98$ ms, $M_{children} = 1229.95$ ms; $\beta_{diff} = -367.76$ ms, 95% HDI for the difference in means [-503.42 ms, -223.85 ms]). This age-related difference parallels findings in spoken language (Fernald et. al., 2006) and shows that young ASL learners are still making progress towards adult-levels of ASL processing efficiency.

Links between children's age and efficiency in incremental sign comprehension

The fourth question of interest was whether young ASL-learners show age-related increases in processing efficiency that parallel those found in spoken languages. To answer this question, we estimated relations between young ASL learners' age-related increases in the speed and accuracy with which they interpreted familiar signs (see Table 3 for point and interval estimates). Mean accuracy

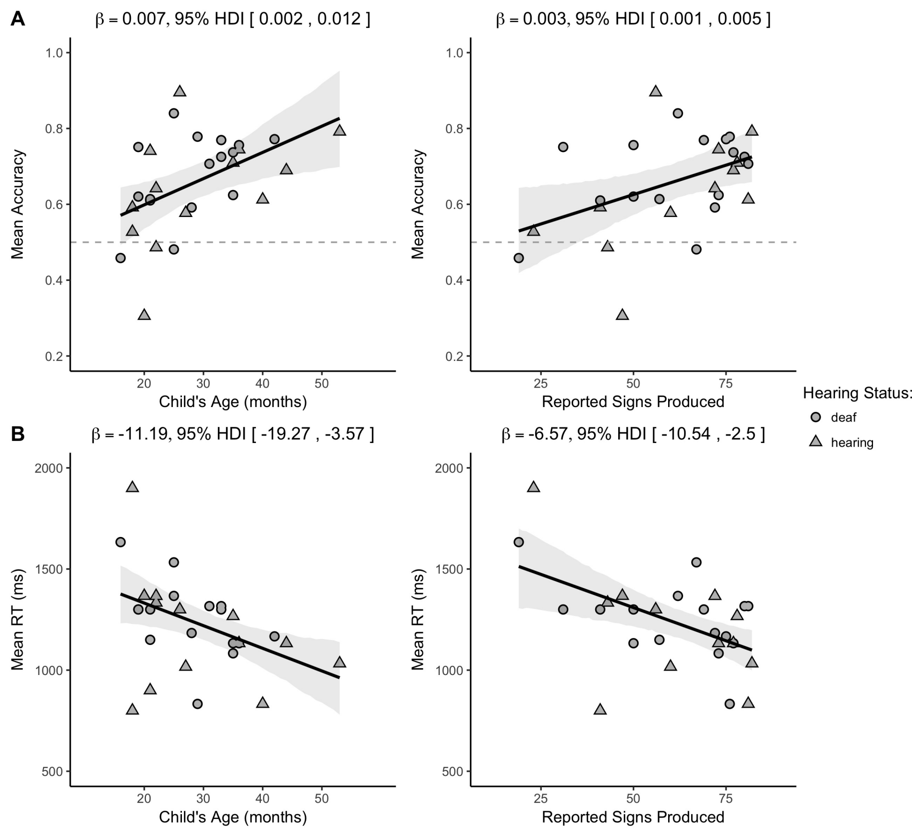


Figure 2.5: Scatterplots of relations between children's age and vocabulary and measures of their mean accuracy (4A) and mean RT (4B) in the VLP procedure. Shape represents children's hearing status. The solid black line is the maximum a posteriori model estimate for the mean accuracy at each age point. The shaded gray regions represent the 95% Highest Density Interval (range of plausible values) around the regression line.

was positively associated with age (Figure 4A), indicating that older ASL learners were more accurate than younger children in fixating the target picture. The Bayes Factor (BF) indicated that a model including a linear association was 12.8 times more likely than an intercept-only model, providing strong evidence for developmental change. The β estimate indicates that, for each month of age, children increased their accuracy score by 0.007, i.e., an increase of ~1% point, meaning that over the course of one year the model estimates a ~12% point gain in accuracy when establishing reference in the VLP task. Mean RTs were negatively associated with age (Figure 4A), indicating that older children shifted to the target picture more quickly than did younger children. The BF was ~14, providing strong evidence for a linear association. The model estimates a ~11 ms gain in

RT for each month, leading to a ~132 ms gain in speed of incremental ASL comprehension over one year of development.

Together, the accuracy and RT analyses showed that young ASL learners reliably looked away from the central signer to shift to the named target image in the VLP task. Importantly, children varied in their response times and accuracy, and this variation was meaningfully linked to age. Thus, like children learning spoken language, ASL learners improve their real-time language processing skills over the second and third years of life as they make progress towards adult levels of language fluency.

Table 2.3: Summary of the four linear models using children's age and vocabulary size to predict accuracy (proportion looking to target) and reaction time (latency to first shift in ms). BF is the Bayes Factor comparing the evidence in favor of linear model to an intercept-only (null) model; Mean Beta is the mean of the posterior distribution for the slope parameter for each model (i.e., the linear association); and the Highest Density Interval (HDI) shows the interval containing 95% of the plausible slope values given the model and the data.

Model specification	**Bayes Factor**	**Mean Beta**	**95% HDI**
Accuracy ~ Age	12.8	0.007	[0.002, 0.012]
Accuracy ~ Vocab	6.8	0.003	[0.001, 0.005]
RT ~ Age	14.4	-11.2 ms	[-19.3 ms, -3.6 ms]
RT ~ Vocab	18.7	-6.6 ms	[-10.5 ms, -2.5 ms]

Links between children's incremental sign comprehension and productive vocabulary

The final question of interest was whether individual differences in processing skills were related to the size of children's ASL vocabularies. As shown in Figure 4B, children with higher accuracy scores also had larger productive vocabularies (BF = 6.8), with the model estimating a 0.003 increase for each additional sign known. Moreover, children who were faster to recognize ASL signs were those with larger sign vocabularies (BF = 18.7), with each additional sign resulting in a ~7 ms decrease in estimated RT. Taken together, older children and children with larger expressive vocabularies were more accurate and efficient in identifying the referents of familiar signs. It is important to point out that the independent effect of vocabulary size on ASL processing could not be assessed here given the correlation between age and vocabulary (r = 0.76) in our sample of children ages one to four years. However, these findings parallel results in the substantial body of previous research with monolingual children learning spoken languages, such as English (Fernald et al., 2006) and Spanish (Hurtado, Marchman, & Fernald, 2007).

2.3 Discussion

Efficiency in establishing reference in real-time lexical processing is a fundamental component of language learning. Here, we developed the first measures of young ASL learners' real-time language comprehension skills. There are five main findings from this research.

First, both adults and children showed a similar qualitative pattern of looking behavior as signs unfolded in time. They began by looking at the signer to gather information about the signed sentence, before shifting gaze to the named object, followed by a return in looking to the signer. All signers allocated very few fixations to the distractor image at any point during the signed sentence.

Second, children and adults tended to shift their gaze away from the signer and to the named referent prior to sign offset, providing evidence of incremental ASL processing. This rapid influence of language on visual attention in ASL is perhaps even more striking since premature gaze shifts could result in a degraded the linguistic signal processed in the periphery or in missing subsequent linguistic information altogether. Furthermore, evidence of incremental gaze shifts suggests that eye movements during ASL processing index efficiency of lexical comprehension, as previously shown in spoken languages, which is important for future work on the psycholinguistics of early sign language acquisition.

Third, deaf and hearing native signers, despite having differential access to auditory information, showed remarkably similar looking behavior during real-time ASL comprehension. Even though the deaf and hearing children had differential access to auditory information in their daily lives, this experience did not change their overall looking behavior or the timing of their gaze shifts during ASL comprehension. Instead, both groups showed parallel sensitivity to the in-the-moment constraints of processing ASL in real time. That is, both deaf and hearing children allocated similar amounts of visual attention to the signer, presumably because this was the only fixation point in the visual scene that also provided information with respect to their goal of language comprehension. This is in stark contrast to what hearing children could potentially do in a similar grounded language comprehension task where a speaker was a potential visual target. In that case, the hearing listener could choose to look at the speaker or to look elsewhere, without losing access to the incoming language via the auditory channel. Thus, they can look while they listen.

Fourth, like children learning spoken language, young ASL-learners were less efficient than adults in their real-time language processing, but they showed significant improvement with age over the first four years. Moreover, although all target signs were familiar to children, older children identified the named referents more quickly and accurately than younger children. This result suggests that the real-time comprehension skills of children who are learning ASL in native contexts follow a similar developmental path to that of spoken language learners, as has been shown in previous work on ASL production (Lillo-Martin, 1999; Mayberry & Squires, 2006). By developing precise measures of real-time ASL comprehension, we were able to study children's language skills earlier in development as compared to other methods.

Fifth, we found a link between ASL processing skills and children's productive vocabularies. ASL-learning children who knew more signs were also faster and more accurate to identify the correct referent than those who were lexically less advanced. These results are consistent with studies of English- and Spanish-learning children, which find strong relations between efficiency in online language comprehension and measures of linguistic achievement (Fernald et al., 2006; Marchman & Fernald, 2008).

2.3.1 Limitations and open questions

This study has several limitations. First, while the sample size is larger than in most previous studies of ASL development, it is still relatively small compared to many studies of spoken language acquisition - an unsurprising limitation, given that native ASL-learners are a rare population. Thus more data are needed to characterize more precisely the developmental trajectories of sign language processing skills. Second, testing children within a narrower age range might have revealed independent effects of vocabulary size on ASL processing, which could not be assessed here given the correlation between age and vocabulary size in our broad sample of children from one to four years. To facilitate replication and extension of our results, we have made all of our stimuli, data, and analysis code publicly available (`https://github.com/kemacdonald/SOL`).

Third, we did not collect measures of age-related gains in children's general cognitive abilities. Thus, it is possible that our estimates of age-related changes in lexical processing are influenced by children's developing efficiency in other aspects of cognition, e.g., increased control of visual attention. Work on the development of visual attention from adolescence to early adulthood shows that different components of visual attention (the ability to distribute attention across the visual field, attentional recovery from distraction, and multiple object processing) develop at different rates (Dye and Bavelier, 2009). Moreover, work by Elsabbagh et. al., (2013) shows that infants become more efficient in their ability to disengage from a central stimulus to attend to a stimulus in the periphery between the ages 7 months and 14 months. However, there is a large body of work showing that features of language use and structure (e.g., the frequency of a word, a word's neighborhood density, and the amount of language input a child experiences) affect the speed and accuracy of eye movements in the Looking-While-Listening style tasks (see Tanenhaus et al., 2000 for a review). Thus, while it possible that age-related improvements in general cognitive abilities are a factor in our results, we think that the strength of the prior evidence suggests that more efficient gaze shifts in the VLP task are indexing improvements in the efficiency of incremental ASL comprehension.

A fourth limitation is that characteristics of our task make it difficult to directly compare our findings with previous work on ASL processing by adults. For example, in contrast to prior gating studies (e.g., Emmorey & Corina, 1990; Morford & Carlsen, 2011), our stimuli consisted of full sentences in a child-directed register, not isolated signs, and we used a temporal response measure rather than an open-ended untimed response. However, it is interesting to note that the mean reaction time of the adults in our task (M = 862 ms) is strikingly close to the average performance of native adult signers in Lieberman et al.'s (2015) "unrelated" condition (M = 844 ms). In addition, we did not select stimuli that parametrically varied features of signs that may influence speed of incremental ASL comprehension, including iconicity and degree of phonological overlap. However, we were able to use a recently created database of lexical and phonological properties of 1000 signs (Caselli et. al., 2017) to explore this possibility. We did not see evidence that iconicity or degree of phonological overlap influenced speed or accuracy of eye movements in children or adults in our sample of eight target signs (see Figures S4 and S5 in the online supplement).

We also cannot yet make strong claims about processing in signed vs. spoken languages in absolute terms because the VLP task included the signer as a central fixation, resulting in different task demands compared to the two-alternative procedure used to study children's spoken language processing (e.g., Fernald et al. 1998). However, a direct comparison of the timecourse of eye movements during signed and spoken language processing is a focus of our ongoing work (MacDonald et al., 2017). Nevertheless, the current results reveal parallels with previous findings showing incremental processing during real-time spoken language comprehension (see Tanenhaus et al., 2000) and sign language comprehension in adults (Lieberman et al., 2015). Moreover, we established links between early processing efficiency and measures of vocabulary in young ASL-learners, suggesting that parallel mechanisms drive language development, regardless of the language modality.

Finally, our sample is not representative of most children learning ASL in the United States. Since most deaf children are born to hearing parents unfamiliar with ASL, many are exposed quite inconsistently to sign language, if at all. We took care to include only children exposed to ASL from birth. The development of real-time ASL processing may look different in children who have inconsistent or late exposure to ASL (Mayberry, 2007). An important step is to explore how variation in ASL processing is influenced by early experience with signed languages. Since children's efficiency in interpreting spoken language is linked to the quantity and quality of the speech that they hear (Hurtado, Marchman, & Fernald, 2008; Weisleder & Fernald, 2013), we would expect similar relations between language input and outcomes in ASL-learners. We hope that the VLP task will provide a useful method to track precisely the developmental trajectories of a variety of ASL-learners.

2.4 Conclusion

This study provides evidence that both child and adult signers rapidly shift visual attention as signs unfold in time and prior to sign offset during real-time sign comprehension. In addition, individual variation in speed of lexical processing in child signers is meaningfully linked to age and vocabulary. These results contribute to a growing literature that highlights parallels between signed and spoken language development when children are exposed to native sign input, suggesting that it is the quality of children's input and not features of modality (auditory vs. visual) that facilitate language development. Moreover, similar results for deaf and hearing ASL-learners suggest that both groups, despite large differences in their access to auditory information in their daily lives, allocated attention in similar ways while processing sign language from moment to moment. Finally, these findings indicate that eye movements during ASL comprehension are linked to efficiency of incremental sign recognition, suggesting that increased efficiency in real-time language processing is a language-general phenomenon that develops rapidly in early childhood, regardless of language modality.

Chapter 3

Children flexibly seek visual information during signed and spoken language comprehension[1]

In this chapter, we present two studies of eye movements during real-time familiar language processing. Within our broader active-social framework, these studies explore whether children adapt their eye movements to query locations that are more useful for the goal of rapid language comprehension. Moreover, these studies investigate some factors that influence children's decisions of when to stop fixating on a social partner and seek a named referent, synthesizing work on threshold models of decision making with research on language-driven visual attention.

Real-time language comprehension involves linking the incoming linguistic signal to the visual world. Information gathered via visual fixations can facilitate the comprehension process. But do listeners seek language-relevant visual information? Here, we propose that children flexibly adapt their eye movements to seek information from social partners that supports language understanding. We present evidence for our explanation using two case studies of eye movements during real-time language processing: children learning spoken English vs. children learning American Sign Language

[1]Parts of this chapter are published as MacDonald, Blonder, Marchman, Fernald, & Frank (2017) An information-seeking account of eye movements during spoken and signed language comprehension and as MacDonald, Marchman, Fernald, & Frank (2018) Adults and preschoolers seek visual information to support language comprehension in noisy environments. Proceedings of the 39th and 40th Annual Meetings of the Cognitive Science Society.

and children processing English in noisy vs. clear auditory environments. Across both studies, we found that listeners adapted their gaze to fixate longer on a social partner when it was useful for language comprehension. Fixating longer on their social partner led to a higher proportion of gaze shifts landing on the named objects, and more language-driven, as opposed to random, shifts. These results suggest that children can increase their information gathering thresholds to seek additional visual information from social partners that supports real-time language comprehension.

Chapter 4

Social cues modulate attention and memory during cross-situational word learning[1]

In this chapter, we present a series of studies exploring adults' word learning in the presence of social cues that disambiguate reference. Within our broader active-social framework, these experiments investigate how social information changes statistical word learning by (1) providing stronger information (i.e., answers) about the target word-object link and (2) constraining the number of potential word-object links (i.e., hypotheses) that learners track over time. Overall, this line of work brings together ideas from social-pragmatic and statistical accounts of language acquisition to explore how social cues can shape the representations that support cross-situational word learning.

Because learners hear language in environments that contain many things to talk about, figuring out 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

[1] This chapter is published in MacDonald, Yurovsky, & Frank (2017) Social cues modulate the representations underlying cross-situational learning. *Cognitive Psychology* , 94, 67-84.

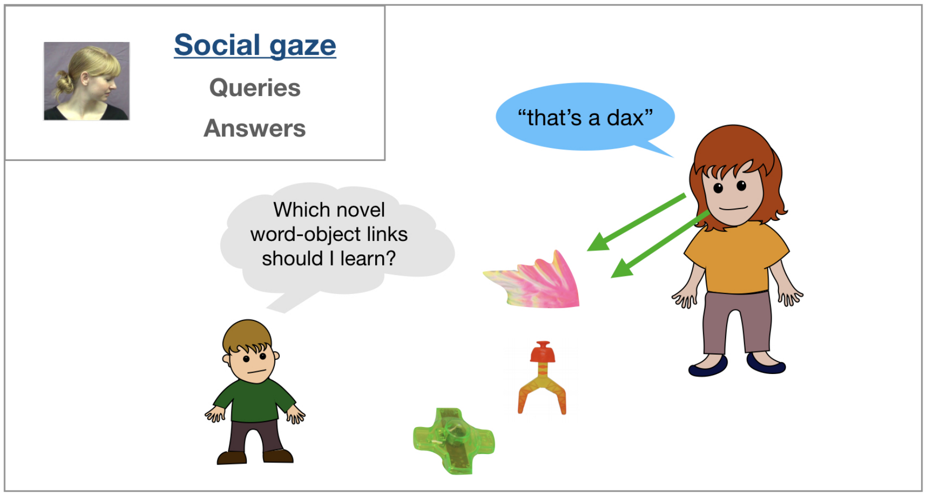


Figure 4.1: A schematic showing the components of the active-social learning framework addressed by the case studies in Chapter 4.

cues. Social cues shifted learners away from tracking multiple hypotheses and towards storing only a single hypothesis (Experiments 1 and 2). Also, 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). 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). These results suggest that the representations underlying cross-situational word learning of concrete object labels flexibly respond to uncertainty in the input. And when ambiguity is high, learners tend to store a broader range of information.

4.1 Introduction

Learning the meaning of a new word should be hard. Consider that even concrete nouns are often used 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.[2] Remarkably, word

[2]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,"

learning proceeds despite this uncertainty, with estimates of adult vocabularies ranging between 50,000 to 100,000 distinct words (P. 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 (Siskind, 1996; Yu & Smith, 2007). Recent 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 & 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 (K. Smith, Smith, & Blythe, 2011).

Although 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 how closely learners approximate the input distribution (for review, see Smith, Suanda, & Yu 2014). One approach has been to model learning as a process of updating connection strengths between multiple word-object links (McMurray, Horst, & Samuelson, 2012), while other approaches have argued that learners store only a single word-object hypothesis (Trueswell, Medina, Hafri, & Gleitman, 2013). In recent experimental and modeling work Yurovsky and Frank (2015) suggest an integrative explanation: learners allocate a fixed amount of attention to a single hypothesis and distribute the rest evenly among the remaining alternatives. As the set of alternatives grows, the amount of attention allocated to each object approaches zero.

In addition to the debate about representation, researchers have disagreed about how to characterize the ambiguity of the input to cross-situational learning mechanisms. One way to quantify the uncertainty in a naming event is to show adults video clips of caregiver-child interactions and measure their accuracy at guessing the meaning of an intended referent (Human Simulation Paradigm: HSP [Gillette, Gleitman, Gleitman, and Lederer, 1999]). Using the HSP, Medina, Snedeker, Trueswell, and Gleitman (2011) found that approximately 90% of learning episodes were ambiguous (< 33% accuracy) and only 7% were relatively unambiguous (> 50% accuracy). In contrast, Yurovsky,

"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.

Smith, and Yu (2013) found a higher proportion of clear naming events, with approximately 30% being unambiguous ($> 90\%$ accuracy). Consistent with this finding, Cartmill, Armstrong, Gleitman, Goldin-Meadow, Medina, and Trueswell (2013) showed that the proportion of unambiguous naming episodes varies across parent-child dyads, with some parents rarely providing highly informative contexts and others' doing so relatively more often.[3]

Thus, representations in 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. These results raise an interesting question: could learners be sensitive to the ambiguity of the input and use this information to alter the representations they store in memory? In the current line of work, we investigated how the presence of referential cues in the social context might alter the ambiguity of the input to statistical word learning mechanisms.

Social-pragmatic theories of language acquisition emphasize the importance of social cues for word learning (P. Bloom, 2002; E. V. Clark, 2009; Hollich et al., 2000). 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 strong links between early intention-reading skills (e.g., gaze following) and later vocabulary growth (Brooks & Meltzoff, 2005, 2008; Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998). Moreover, studies outside the domain of language acquisition have shown that the presence of social cues: (a) produce better spatial learning of audiovisual events (Wu, Gopnik, Richardson, & Kirkham, 2011), (b) boost recognition of a cued object (Cleveland, Schug, & Striano, 2007), and (c) lead to preferential encoding of an object's featural information (J. M. Yoon, Johnson, & Csibra, 2008). Together, the evidence suggests that social cues could alter the representations stored during cross-situational word learning by modulating how people allocate attention to the relevant statistics in the input.

The goal of our current investigation was to ask whether the presence of a valid social cue – a speaker's gaze – could change the representations underlying cross-situational word learning. We

[3]The differences in the estimates of referential uncertainty in these studies could be driven by the different sampling procedures used to select naming events for the HSP. Yurovsky, Smith, and Yu (2013) sampled utterances for which the parent labeled a co-present object, whereas Medina, Snedeker, Trueswell, et al. (2011) randomly sampled any utterances containing concrete nouns. Regardless of these differences, the key point here is that variability in referential uncertainty across naming events exists and thus could alter the representations underlying cross-situational learning.

used a modified version of Yurovsky and Frank (2015)'s paradigm to provide a direct measure of memory for alternative word-object links during cross-situational learning. In Experiment 1, we manipulated the presence of a referential cue at different levels of attention and memory demands. At all levels of difficulty, learners tracked a strong single hypothesis but were less likely to track multiple word-object links when a social cue was present. In Experiment 2, we replicated the findings from Experiment 1 using a more ecologically valid social cue. In Experiment 3, we moved to a parametric manipulation of referential uncertainty by varying the reliability of the speaker's gaze. Learners were sensitive to graded changes in reliability and retained more word-object links as uncertainty in the input increased. Finally, in Experiment 4, we equated the length of the initial naming events with and without the referential cue. Learners stored less information in the presence of gaze even when they had visually inspected the objects for the same amount of time. In sum, our data suggest that cross-situational word learners are quite flexible, storing representations with different levels of fidelity depending on the amount of ambiguity present during learning.

4.2 Experiment 1

We set out to test the effect of a referential cue on the representations underlying cross-situational word learning. We used a version of Yurovsky and Frank (2015)'s paradigm where we manipulated the ambiguity of the learning context by including a gaze cue from a schematic, female interlocutor. Participants saw a series of ambiguous exposure trials where they heard one novel word that was either paired with a gaze cue or not and selected the object they thought went with each word. In subsequent test trials, participants heard the novel word again, this time paired with a new set of novel objects. One of the objects in this set was either the participant's initial guess (Same test trials) or one of the objects was *not* their initial guess (Switch test trials). Performance on Switch trials provided a direct measure of whether referential cues influenced the number of alternative word-object links that learners stored in memory. If learners performed worse on Switch trials after an exposure trial with gaze, this would suggest that they stored fewer additional objects from the initial learning context.

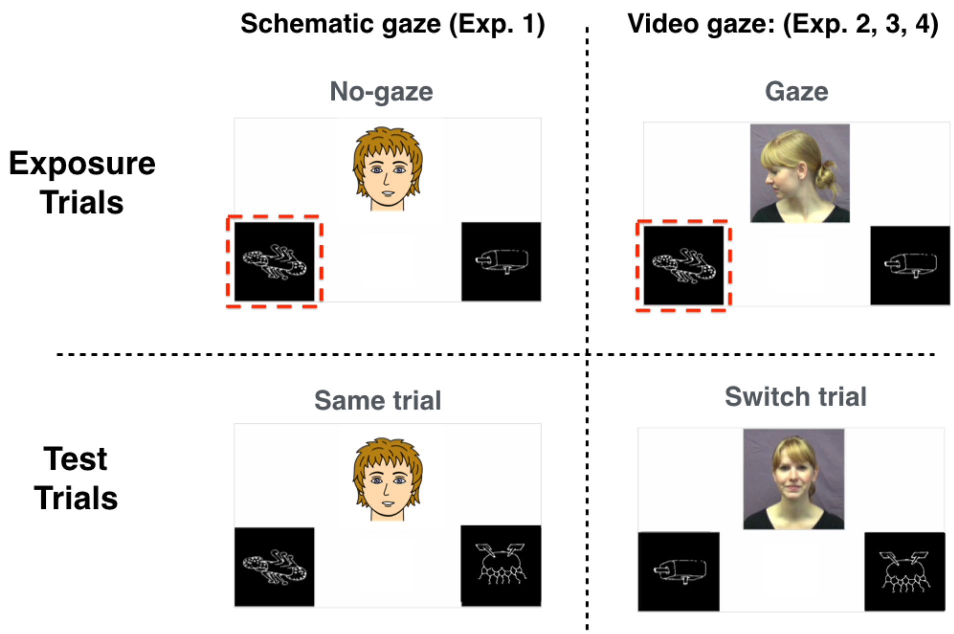


Figure 4.2: Screenshots of exposure and test trials from Experiments 1-4. The top left panel shows an exposure trial in the No-gaze condition using the schematic gaze cue (Experiment 4.1). The top right panel shows an exposure trial in the Gaze condition using the video gaze cue (Experiments 4.2-4.4). 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. Participants either saw 2, 4, 6, or 8 referents on the screen depending on condition assignment.

4.2.1 Method

Participants

We posted a set of Human Intelligence Tasks (HITs) to Amazon Mechanical Turk. Only participants with US IP addresses and a task approval rate above 95% were allowed to participate, and each HIT paid 30 cents. 50-100 HITs were posted for each of the 32 between-subjects conditions. Data were excluded if participants completed the task more than once or if participants did not respond correctly on familiar object trials (131 HITs). The final sample consisted of 1438 participants.

Stimuli

Figure 1 shows screenshots taken from Experiment 1. Visual stimuli were black and white pictures of familiar and novel objects taken from Kanwisher, Woods, Iacoboni, and Mazziotta (1997). Auditory stimuli were recordings of familiar and novel words by an AT&T Natural Voices ^{TM}(voice: Crystal) speech synthesizer. Novel words were 1-3 syllable pseudowords that obeyed all rules of English phonotactics. A schematic drawing of a human speaker was chosen for ease of manipulating the direction of gaze, the referential cue of interest in this study. All experiments can be viewed and downloaded at the project page: `https://kemacdonald.github.io/soc_xsit/`.

Design and Procedure

Participants 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. 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 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.

Participants were randomly assigned to one of the 32 between-subjects conditions (4 Referents X 4 Intervals X 2 Gaze conditions). Participants either saw 2, 4, 6, or 8 referents on the screen and test trials occurred at different intervals after exposure trials: either 0, 1, 3, or 7 trials from the initial exposure to a word. For example, in the 0-interval condition, the test trial for that word would occur immediately following the exposure trial, but in the 3-interval condition, participants would see three additional exposure trials for other novel words before seeing the test trial for the initial word. The interval conditions modulated the time delay and the number of intervening trials between learning and test, and the number of referents conditions modulated the attention demands present during learning.

Participants were assigned to either the Gaze or No-Gaze condition. In the Gaze condition, gaze was directed towards one of the objects on exposure trials; in the No-Gaze condition, gaze was always directed straight ahead (see Figure 1 for examples). At test, gaze was always directed straight ahead. To show participants that their response had been recorded, a red box appeared around the selected object for one second. This box always appeared around the selected object, even if participants' selections were incorrect.

4.2.2 Results and Discussion

Analysis plan

The structure of our analysis plan is parallel across all four experiments. First, we examined accuracy on exposure trials in the Gaze condition and then we compared response times on exposure trials across the Gaze and No-Gaze conditions. These analyses tested whether learners were (a) sensitive to our experimental manipulation and (b) altered their allocation of attention in response to the presence of a social cue. Accuracy on exposure trials was defined as selecting the referent that was the target of gaze in the Gaze condition. (Note that there was no "correct" behavior for exposure trials in the No-Gaze condition.) Next, we examined accuracy on test trials to test whether learners' memory for alternative word-object links changed depending on the ambiguity of the learning context. Accuracy on test trials (both Same and Switch) was defined as selecting the referent that was present during the exposure trial for that word.

The key behavioral prediction of our hypothesis was that the presence of gaze would result in reduced memory for multiple word-object links, operationalized as a decrease in accuracy on Switch test trials after seeing exposure trials with a gaze cue. To quantify participants' behavior, we used mixed-effects regression models with the maximal random effects structure justified by our experimental design: by-subject intercepts and slopes for each trial type (Barr, Levy, Scheepers, & Tily, 2013). We limited all models to include only two-way interactions because the critical test of our hypothesis was the interaction between gaze condition and trial type, and we did not have theoretical predictions for any possible three-way or four-way interactions.

In the main text, we only report effects that achieved statistical significance at the $\alpha = .05$

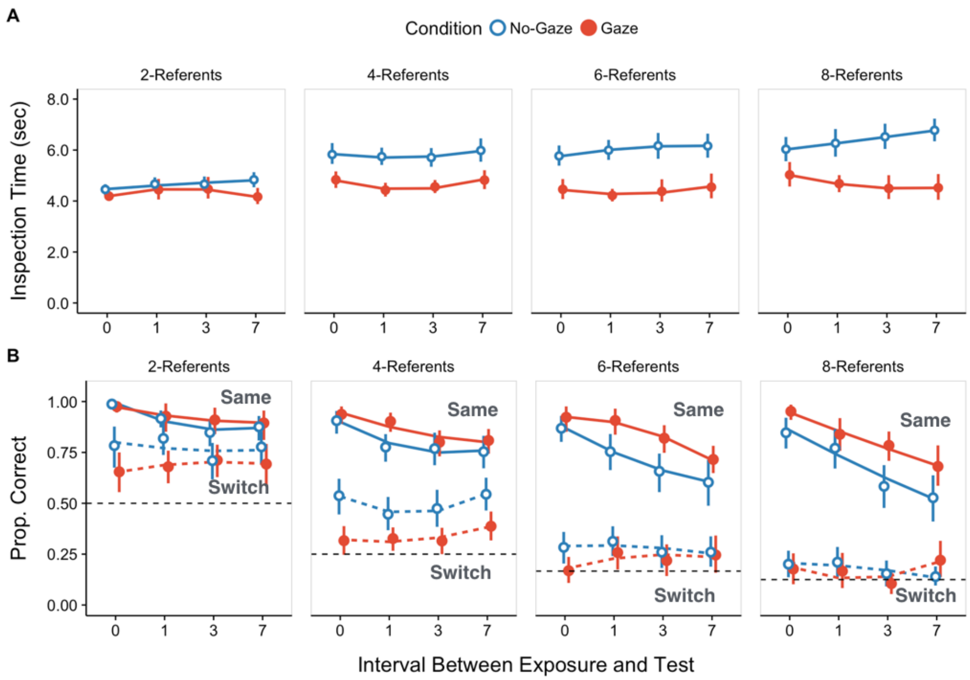


Figure 4.3: Experiment 4.1 results. The top row shows average inspection times on exposure trials for all experimental conditions as a function of the number of trials that occurred between exposure and test. Each panel represents a different number of referents, and line color represents the Gaze and No-Gaze conditions. The bottom row shows accuracy on test trials for all conditions as a function of the number of intervening trials. The horizontal dashed lines represent chance performance for each number of referents, and the type of line (solid vs. dashed) represents the different test trial types (Same vs. Switch). Error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

threshold. In the Appendix, we report the full model specification and output for each of the models in the paper. All models were fit using the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2013), and all of our data and our processing/analysis code can be viewed in the version control repository for this paper at `https://github.com/kemacdonald/soc_xsit`.

Exposure trials

To ensure that our referential cue manipulation was effective, we compared participants' accuracies on exposure trials in the Gaze condition to a model of random behavior defined as a Binomial distribution with a probability of success $\frac{1}{NumReferents}$. Correct performance was defined as selecting the object that was the target of the speaker's gaze. Following Yurovsky and Frank (2015), we fit logistic regressions for each gaze, referent, and interval combination specified as `Gaze Target ∼ 1 + offset(logit(1/Referents))`. The offset encoded the chance probability of success given the number of referents, and the coefficient for the intercept term shows on a log-odds scale how much more likely participants were to select the gaze target than would be expected if participants were selecting randomly. In all conditions, participants used gaze to select referents on exposure trials more often than expected by chance (smallest $\beta = 1.4$, z = 9.38, $p < .001$). However, the mean proportion of gaze following varied across conditions (overall $M = 0.84$, range: 0.77–0.93).

We were also interested in differences in participants' response times across the experimental conditions. Since these trials were self-paced, participants could choose how much time to spend inspecting the referents on the screen, thus providing an index of participants' attention. To quantify the effects of gaze, interval, and number of referents, we fit a linear mixed-effects model that predicted participants' inspection times as follows: `Log(Inspection time) ∼ (Gaze * Log(Interval) + Log(Referents))`^{2} `+ (1 | subject)`. We found a significant main effect of the number of referents ($\beta = 0.34$, p $<$ .001) with longer inspection times as the number of referents increased, a significant interaction between gaze condition and the number of referents ($\beta =$ -0.27, p $<$ .001) with longer inspection times in the No-Gaze condition, especially as the number of referents increased, and a significant interaction between gaze condition and interval ($\beta =$ -0.08, $p = 0.004$) with longer inspection times in the No-Gaze condition, especially as the number of intervening trials increased (see the top row of Figure 2). Shorter inspection times on exposure trials with gaze provide evidence that the presence of a referential cue focused participants' attention on a single referent and away from alternative word-object links.

Test trials

Next, we explored participants' accuracy in identifying the referent for each word in all conditions for both kinds of test trials (see the bottom row of Figure 2). We first compared the distribution of correct responses made by each participant to the distribution expected if participants were selecting randomly defined as a Binomial distribution with a probability of success $\frac{1}{NumReferents}$. Correct performance was defined as selecting the object that was present on the exposure trial for that word. We fit the same logistic regressions as we did for exposure trials: `Correct ∼ 1 +`

Predictor	Estimate	Std. Error	z value	p value	
Intercept	3.01	0.29	10.35	$< .001$	***
Switch Trial	-1.36	0.24	-5.63	$< .001$	***
Gaze Condition	0.12	0.26	0.47	0.64	
Log(Interval)	-0.45	0.11	-4.08	$< .001$	***
Log(Referents)	0.23	0.11	2.02	0.04	*
Switch Trial*Gaze Condition	-1.09	0.12	-9.07	$< .001$	***
Switch Trial*Log(Interval)	0.52	0.05	9.50	$< .001$	***
Switch Trial*Log(Referent)	-0.59	0.09	-6.49	$< .001$	***
Gaze Condition*Log(Interval)	0.06	0.06	1.00	0.32	
Gaze Condition*Log(Referent)	0.20	0.09	2.15	0.03	*
Log(Interval)*Log(Referent)	-0.04	0.04	-1.02	0.31	

Table 4.1: Predictor estimates with standard errors and significance information for a logistic mixed-effects model predicting word learning in Experiment 4.1.

`offset(logit(1/Referents))`. In 31 out of the 32 conditions for both Same and Switch trials, participants chose the correct object more often than would be expected by chance (smallest $\beta =$ 0.36, $z = 2.44$, $p = 0.01$). On Switch trials in the 8-referent, 3-interval condition, participants' responses were not significantly different from chance ($\beta = 0.06$, z = 0.33, $p = 0.74$). Participants' success on Switch trials replicates the findings from Yurovsky and Frank (2015) and provides direct evidence that learners encoded more than a single hypothesis in ambiguous word learning situations even under high attentional and memory demands and in the presence of a referential cue. To quantify the effects of gaze, interval, and number of referents on the probability of a correct response, we fit the following mixed-effects logistic regression model to a filtered dataset where we removed participants who did not reliably select the object that was the target of gaze on exposure trials:[4] `Correct` $\sim$ `(Trial Type + Gaze + Log(Interval) + Log(Referents))`^{2} `+ offset(logit(`$^1/_{Referents}$`)) + (TrialType | subject)`. We coded interval and number of referents as continuous predictors and transformed these variables to the log scale.[5]

Table 1 shows the output of the logistic regression. We found significant main effects of the number of referents ($\beta = 0.23$, $p < .001$) and interval ($\beta = -0.45$, $p < .001$), such that as each of these factors increased, accuracy on test trials decreased. We also found a significant main effect of trial type ($\beta = -1.36$, $p < .001$), with worse performance on Switch trials. There were

[4] We did not predict that there would be a subset of participants who would not follow the gaze cue, thus this filtering criterion was developed posthoc. However, we think that the filter is theoretically motivated because we would only expect to see an effect of gaze if participants actually used the gaze cue. The filter removed 94 participants (6% of the sample). The key inferences from the data do not depend on this filtering criterion.

[5] If we allowed for three-way interactions in the model, the key interaction between gaze condition and trial type remained significant (β = -1.3, p = 0.006).

significant interactions between trial type and interval ($\beta = 0.52$, $p < .001$), trial type and referents ($\beta = -0.59$, $p < .001$), and gaze condition and referents ($\beta = 0.2$, $p < .05$). These interactions can be interpreted as meaning: (a) the interval between exposure and test affected Same trials more than Switch trials, (b) the number of referents affected Switch trials more than Same trials, and (c) participants performed slightly better at the higher number of referents in the Gaze condition. The interactions between gaze condition and referents and between referents and interval were not significant. Importantly, we found the predicted interaction between trial type and gaze condition ($\beta = -1.09$, $p < .001$), with participants in the Gaze condition performing worse on Switch trials. This interaction provides direct evidence that the presence of a referential cue reduces participants' memory for alternative word-object links.

We were also interested in how the length of inspection times on exposure trials would affect participants' accuracy at test. So we fit an additional model where participants' inspection times were included as a predictor. We found a significant interaction between inspection time and gaze condition ($\beta = -0.17$, $p = 0.01$) such that longer inspection times provided a larger boost to accuracy in the No-Gaze condition. Importantly, the key test of our hypothesis, the interaction between gaze condition and trial type, remained significant in this alternative version of the model (β = -1.02, p = p < .001).

Taken together, the inspection time and accuracy analyses provide evidence that the presence of a referential cue modulated learners' attention during learning, and in turn made them less likely to track multiple word-object links. We saw some evidence for a boost to performance on Same trials in the Gaze condition at the higher number of referent and interval conditions, but reduced tracking of alternatives did not always result in better memory for learners' candidate hypothesis. This finding suggests that the limitations on Same trials may be different than those regulating the distribution of attention on Switch trials.

There was relatively large variation in performance across conditions in the group-level accuracy scores and in participants' tendency to *use* the referential cue on exposure trials. Moreover, we found a subset of participants who did not reliably use the gaze cue at all. It is possible that the effect of gaze was reduced because the referential cue that we used – a static schematic drawing of a speaker – was relatively weak compared to the cues present in real-world learning environments. Thus we do not yet know how learners' memory for alternatives during cross-situational learning would change in the presence of a stronger and more ecologically valid referential cue. We designed Experiment 2 to address this question.

4.3 Experiment 2

In Experiment 2, we set out to replicate the findings from Experiment 1 using a more ecologically valid stimulus set. We replaced the static, schematic drawing with a video of an actress. While these stimuli were still far from actual learning contexts, they included a real person who provided both a gaze cue and a head turn towards the target object. To reduce the across-conditions variability that we found in Experiment 1, we introduced a within-subjects design where each participant saw both Gaze and No-Gaze exposure trials in a blocked design. We selected a subset of the conditions from Experiment 1 and tested only the 4-referent display with 0 and 3 intervening trials as between-subjects manipulations. Our goals were to replicate the reduction in learners' tracking of alternative word-object links in the presence of a referential cue and to test whether increasing the ecological validity of the cue would result in a boost to the strength of learners' recall of their candidate hypothesis.

4.3.1 Method

Participants

Participant recruitment and inclusion/exclusion criteria were identical to those of Experiment 1. 100 HITs were posted for each condition (1 Referent X 2 Intervals X 2 Gaze conditions) for a total of 400 paid HITs (33 HITs excluded).

Stimuli

Audio and picture stimuli were identical to Experiment 1. The referential cue in the Gaze condition was a video (see Figure 1). On each exposure trial, the actress looked out at the participant with a neutral expression, smiled, and then turned to look at one of the four images on the screen. She maintained her gaze for 3 seconds before returning to the center. On test trials, she looked straight ahead for the duration of the trial.

Design and Procedure

Procedures were identical to those of Experiment 1. The major design change was a within-subjects manipulation of the gaze cue where each participant saw exposure trials with and without gaze. The experiment consisted of 32 trials split into 2 blocks of 16 trials. Each block consisted of 8 exposure trials and 8 test trials (4 Same trials and 4 Switch trials) and contained only Gaze or No-gaze exposure trials. The order of block was counterbalanced across participants.

4.3.2 Results and Discussion

We followed the same analysis plan as in Experiment 1. We first analyzed inspection times and accuracy on exposure trials and then analyzed accuracy on test trials.

Exposure trials

Similar to Experiment 1, participants' responses on exposure trials differed from those expected by chance (smallest β = 3.39, z = 31.99, $p < .001$), suggesting that gaze was effective in directing participants' attention. Participants in Experiment 2 were more consistent in their use of gaze with the video stimuli compared to the schematic stimuli used in Experiment 1 ($M_{Exp1} = 0.8$, $M_{Exp2} = 0.91$), suggesting that using a real person increased participants' willingness to follow the gaze cue.

We replicated the findings from Experiment 1. Inspection times were shorter when gaze was present (β = -1.1, $p < .001$) and in the 3-interval condition (β = -0.48, $p < .001$). The interaction between gaze and interval was not significant, meaning that gaze had the same effect on participants' inspection times at both intervals (see Panel A of Figure 3).

Test trials

Across all conditions for both trial types, participants selected the correct referent at rates greater than chance (smallest β = 0.58, z = 9.32, $p < .001$). We replicated the critical finding from Experiment 1: after seeing exposure trials with gaze, participants performed worse on Switch trials, meaning they stored fewer word-object links ($\beta = -0.71$, $p < .001$).[6] Participants were also less accurate as the interval between exposure and test increased (β = -0.93, $p < .001$) and on the Switch trials overall ($\beta = -2.99$, $p < .001$).

[6] As in Experiment 1, we fit this model to a filtered dataset removing participants who did not reliably use the gaze cue.

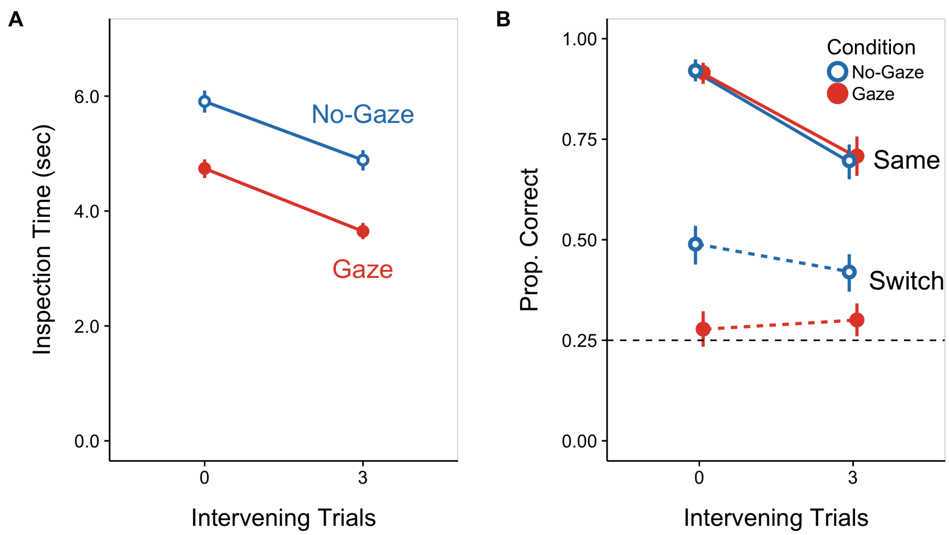


Figure 4.4: Experiment 2 results. Panel A shows inspection times on exposure trials with and without gaze. Panel B shows accuracy on Same and Switch test trials. All plotting conventions are the same as in Figure 2. Error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

In addition, there was a significant interaction between trial type and interval ($\beta = 0.79$, $p < .001$), with worse performance on Switch trials in the 3-interval condition. The interaction between gaze condition and interval was also significant ($\beta = 0.15$, $p = 0.041$), such that participants in the gaze condition were less affected by the increase in interval. Similar to Experiment 1, we did not see evidence of a boost to performance on Same trials in the gaze condition.

Next, we added inspection times on exposure trials to the model. Similar to Experiment 1, the key interaction between gaze and trial type remained significant in this version of the model (β = -0.54, $p < .001$). We also found an interaction between inspection time and trial type ($\beta = 0.21$, $p = 0.05$),

Predictor	Estimate	Std. Error	z value	p value	
Intercept	4.04	0.18	21.97	$< .001$	***
Switch Trial	-2.99	0.19	-16.11	$< .001$	***
Gaze Condition	-0.10	0.16	-0.63	0.53	
Log(Interval)	-0.93	0.10	-9.23	$< .001$	***
Switch Trial*Gaze Condition	-0.71	0.16	-4.49	$< .001$	***
Switch Trial*Log(Interval)	0.79	0.10	8.03	$< .001$	***
Gaze Condition*Log(Interval)	0.15	0.08	2.05	0.04	*

Table 4.2: Predictor estimates with standard errors and significance information for a logistic mixed-effects model predicting word learning in Experiment 4.2.

with longer inspection times providing a larger boost to performance on Switch trials (i.e., stronger memory for alternative word-object links). This result differs slightly from Experiment 1 where we found an interaction between trial type and inspection time, with longer inspection times providing a larger boost to accuracy in the No-Gaze condition. Despite this subtle difference, we speculate that inspection times likely played a similar role in both experiments, with longer inspection times leading to better performance on Switch trials since these trials depended on encoding multiple word-object links. It is also possible that the interaction between gaze condition and inspection time that we found in Experiment 1 was influenced by the different number of referents and interval conditions.

The results of Experiment 2 provide converging evidence for our primary hypothesis that the presence of a referential cue reliably focuses learners' attention away from alternative word-object links and shifts them towards single hypothesis tracking. Moving to the video stimulus led to higher rates of selecting the target of gaze on exposure trials, but did not result in a boost to performance on Same trials. This finding suggests that the level of attention and memory demand present in the learning context might modulate the effect of gaze on the fidelity of learners' single hypothesis.

Thus far we have shown that people store different amounts of information in response to a categorical manipulation of referential uncertainty. In both Experiments 1 and 2, the learning context was either entirely ambiguous (No-Gaze) or entirely unambiguous (Gaze). But not all real-world learning contexts fall at the extremes of this continuum. Could learners be sensitive to more subtle changes in the quality of the input? In our next experiment, we tested a prediction of our account: whether learners would store more word-object links in response to graded changes in referential uncertainty during learning.

4.4 Experiment 3

In Experiment 3, we explored whether learners would allocate attention and memory flexibly in response to *graded* changes in the referential uncertainty that was present during learning. To test this hypothesis, we moved beyond a categorical manipulation of the presence/absence of gaze, and we parametrically varied the reliability of the referential cue. We manipulated cue reliability by adding a block of familiarization trials where we varied the proportion of Same and Switch trials. If participants saw more Switch trials, this provided direct evidence that the speaker's gaze was a less reliable cue to reference because the gaze target on exposure trials would not appear at test. This design was inspired by a growing body of experimental work showing that even young children are sensitive to the prior reliability of speakers and will use this information to decide whom to learn novel words from (e.g., Koenig, Clement, & Harris, 2004).

4.4.1 Method

Participants

Participant recruitment and inclusion/exclusion criteria were identical to those of Experiment 1 and 2 (27 HITs excluded). 100 HITs were posted for each reliability level (0%, 25%, 50%, 75%, and 100%) for total of 500 paid HITs.

Design and Procedure

Procedures were identical to those of Experiments 1 and 2. We modified the design of our 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). Since we were no longer testing the effect of the presence or absence of a referential cue, all exposure trials throughout the experiment included a gaze cue. Finally, at the end of the task, we asked participants to assess the reliability of the speaker on a continuous scale from "completely unreliable" to "completely reliable."

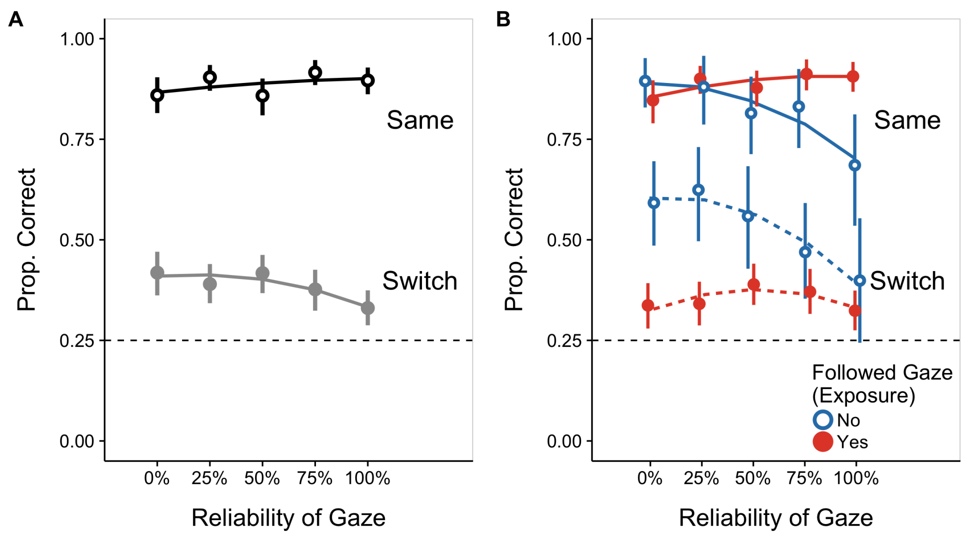


Figure 4.5: Primary analyses of test trial performance in Experiment 3. Panel A shows performance as a function of reliability condition. Panel B shows performance as a function of reliability condition and whether participants chose to follow gaze on exposure trials. The horizontal dashed lines represent chance performance, and error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

4.4.2 Results and Discussion

Exposure trials

Participants reliably chose the referent that was the target of gaze at rates greater than chance (smallest $\beta = 2.62$, z $= 31.99$, $p < .001$). We fit a mixed effects logistic regression model predicting the probability of selecting the gaze target as follows: `Correct-Exposure` $\sim$ `Reliability Condition * Subjective Reliability + (1 | subject)`. We found an effect of reliability condition ($\beta =$ 3.28, $p = 0.03$) such that when the gaze cue was more reliable, participants were more likely to use it ($M_{0\%} = 0.83$, $M_{25\%} = 0.82$, $M_{50\%} = 0.87$, $M_{75\%} = 0.9$, $M_{100\%} = 0.94$). We also found an effect of subjective reliability ($\beta = 7.26$, $p < .001$) such that when participants thought the gaze cue was reliable, they were more likely to use it. This analysis provides evidence that participants were sensitive to the reliability manipulation both in how often they used the gaze cue and in how they rated the reliability of the speaker at the end of the task.

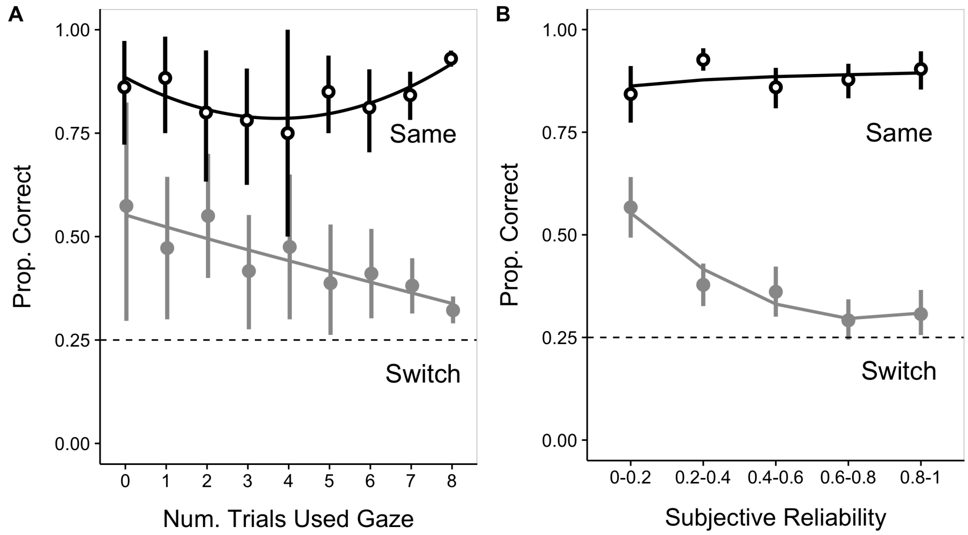


Figure 4.6: Secondary analyses of test trial performance in Experiment 3. Panel A shows accuracy as a function of the number of exposure trials on which participants chose to use the gaze cue. Panel B shows accuracy as a function of participants' subjective reliability judgments. The horizontal dashed lines represent chance performance, and error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

Test trials

Next, we tested whether the reliability manipulation altered the strength of participants' memory for alternative word-object links in the second block of test trials that followed the initial familiarization phase. Across all conditions, participants selected the correct referent at rates greater than chance (smallest $\beta = 0.42$, z = 3.69, $p < .001$). Our primary prediction was an interaction between reliability and test trial type, with higher levels of reliability leading to worse performance on Switch trials (i.e., less memory allocated to alternative word-object links). To explore this prediction, we performed four complementary analyses: our primary analysis, which tested the effect of the reliability manipulation, and three secondary analyses, which explored the effects of participants' (a) use of the gaze cue, (b) subjective reliability assessments, and (c) inspection time on exposure trials.

Reliability condition analysis

To test the effect of reliability, we fit a model predicting accuracy at test using reliability condition and test trial type as predictors. We found a significant main effect of trial type ($\beta = -3.95$, $p <$

.001), with lower accuracy on Switch trials. We also found the key interaction between reliability condition and trial type (β = -0.76, p = 0.044), such that when gaze was more reliable, participants performed worse on Switch trials (see Panel A of Figure 4). This interaction suggests that people store more word-object links as the learning context becomes more ambiguous. However, the interaction between reliability and trial type was not particularly strong, and – similar to Experiment 1 – performance varied across conditions (see the 50% reliable condition in Panel A of Figure 4). So to provide additional support for our hypothesis, we conducted three follow-up analyses.

Gaze use analyses

We would only expect to see a strong interaction between reliability and trial type if learners chose to use the gaze cue during exposure trials. To test this hypothesis, we fit two additional models that included two different measures of participants' use of the gaze cue. First, we added the number of exposure trials on which participants chose to use the gaze cue as a predictor in our model. We found a significant interaction between use of the gaze cue on exposure trials and trial type ($\beta = -1.43$, $p < .001$) with worse performance on Switch test trials when participants used gaze on exposure trials (see Panel B of Figure 4). We also found an interaction between gaze use and reliability ($\beta = 0.97$, $p = 0.004$) such that when gaze was more reliable, participants were more likely to use it. The β value for the interaction between trial type and reliability changed from -0.76 to -0.62, (p = 0.086). This reduction suggests that participants' tendency to use the gaze cue is a stronger predictor of learners' memory for alternative word-object links compared to our reliability manipulation.[7]

We also hypothesized that the reliability manipulation might change how often individual participants chose to use the gaze cue throughout the task. To explore this possibility, we fit a model with the same specifications, but we included a predictor that we created by binning participants based on the number of exposure trials on which they chose to follow gaze (i.e., a gaze following score). We found a significant interaction between how often participants chose to follow gaze on exposure trials and trial type ($\beta = -0.26$, $p < .001$), such that participants who were more likely to use the gaze cue performed worse on Switch trials, but not Same trials (see Panel A of Figure 5).[8] Taken together, the two analyses of participants' use of the gaze cue provide converging evidence

[7]We are grateful to an anonymous reviewer for suggesting this analysis, but we would like to note that it is exploratory.

[8]We found this interaction while performing exploratory data analyses on a previous version of this study with an independent sample (N = 250, $\beta = -0.24$, $p < .001$). The results reported here are from a follow-up study where testing this interaction was a planned analysis.

that when the speaker's gaze was reliable participants were more likely to use the cue, and when they followed gaze, they tended to store less information from the initial naming event.

Subjective reliability analysis

The strong interaction between use of the gaze cue and memory for alternative word-object links suggests that participants' subjective experience of reliability in the experiment mattered. Thus, we fit the same model but substituted subjective reliability for the frequency of gaze use as a predictor of test trial performance. We found a significant interaction between trial type and participants' subjective reliability assessments ($\beta = -1.63$, $p = 0.01$): when participants thought the speaker was more reliable, they performed worse on Switch trials, but not Same trials (see Panel B of Figure 5).

Inspection time analyses

Finally, we analyzed the effect of inspection times on exposure trials, fitting a model using inspection time, trial type, and reliability condition to predict accuracy at test. We found a main effect of inspection time ($\beta = 0.31$, $p = 0.001$), with longer inspection times leading to better performance for both Same and Switch trials. The interaction between inspection time and reliability condition was not significant. The key interaction between reliability condition and trial type remained significant in this version of the model ($\beta =$ -0.58, $p =$ 0.048).

Next, we explored the factors that influenced inspection time on exposure trials by fitting a model to predict inspection times as a function of reliability condition and participants' use of the gaze cue. We found a main effect of participants' use of the gaze cue (-0.32, $p <$ .001) with shorter inspection times when participants followed gaze. The main effect of reliability condition and the interaction between reliability and use of gaze were not significant. These analyses provide evidence that inspection times were similar across the different reliability conditions and that use of the gaze cue was the primary factor affecting how long participants explored the objects during learning.

Together, these four analyses show that when the speaker's gaze was more reliable, participants were more likely to: (a) use the gaze cue, (b) rate the speaker as more reliable, and (c) store fewer word-object links, showing behavior more consistent with single hypothesis tracking. These findings support and extend the results of Experiments 1 and 2 in several important ways. First, similar to Experiment 2, participants' performance on Same trials was relatively unaffected by changes in performance on Switch trials. The selective effect of gaze on Switch trials provides converging evidence that the limitations on Same trials may be different than those regulating the distribution of attention on Switch trials. Second, learners' use of a referential cue was a stronger predictor of reduced memory for alternative word-object links compared to our reliability manipulation. Although we found a significant effect of reliability on participants' use of the gaze cue, participants' tendency to use the cue remained high. Consider that even in the 0% reliability condition the mean proportion of gaze following was still 0.82. It is reasonable that participants would continue to use the gaze cue in our experiment since it was the only cue available and participants did not have a strong reason to think that the speaker would be deceptive.

The critical contribution of Experiment 3 is to show that learners respond to a graded manipulation of referential uncertainty, with the amount of information stored from the initial exposure tracking with the reliability of the cue. This graded accuracy performance shows that learners stored alternative word-object links with different levels of fidelity depending on the amount of referential uncertainty present during learning.

Across Experiments 1-3, learners tended to store fewer word-object links in unambiguous learning contexts when a clear referential cue was present. However, in all three experiments, participants' responses on exposure trials controlled the length of the trial, meaning that when participants used the gaze cue, they also spent less time visually inspecting the objects. Thus, we do not know whether there is an independent effect of referential cues on the representations underlying cross-situational learning, or if the effects found in Experiments 1-3 are entirely mediated by a reduction in inspection time. In Experiment 4, we addressed this possibility by removing participants' control over the length of exposure trials, which made the inspection times equivalent across the Gaze and No-Gaze conditions.

4.5 Experiment 4

In Experiment 4, we asked whether a reduction in visual inspection time in the gaze condition could completely explain the effect of social cues on learners' reduced memory for alternative word-object links. To answer this question, we modified our paradigm and made the length of exposure trials equivalent across the Gaze and No-Gaze conditions. In this version of the task, participants were shown the objects for a fixed amount of time regardless of whether gaze was present. We also included two different exposure trial lengths in order to test whether gaze would have a differential effect at shorter vs. longer inspection times. If the presence of gaze reduces learners' memory for multiple word-object links, then this provides evidence that referential cues affected the underlying representations over and above a reduction in inspection time.

4.5.1 Method

Participants

Participant recruitment and inclusion/exclusion criteria were identical to those of Experiments 1, 2, and 3. 100 HITs were posted for each condition (1 Referent X 2 Intervals X 2 Inspection Time conditions) for a total of 400 paid HITs (37 HITs excluded).

Stimuli

Audio, picture, and video stimuli were identical to Experiments 2 and 3. Since inspection times were fixed across conditions, we wanted to ensure that participants were aware of the time remaining on each exposure trial. So we included a circular countdown timer located above the center video. The timer remained on the screen during test trials but did not count down since participants could take as much time as they wanted to respond on test trials.

Design and Procedure

Procedures were identical to those of Experiment 1-3. The design was identical to that of Experiment 2 and consisted of 32 trials split into 2 blocks of 16 trials. Each block consisted of 8 exposure trials and 8 test trials (4 Same trials and 4 Switch trials) and contained only Gaze or No-Gaze exposure trials. The order of block was counterbalanced across participants.

The major design change was to make the length of exposure trials equivalent across the Gaze and No-Gaze conditions. We randomly assigned participants to one of two inspection time conditions: Short or Long. Initially, the length of the inspection times was based on participants' self-paced inspection times in the Gaze and No-Gaze conditions in Experiment 2 (Short = 3 seconds; Long = 6 seconds). However, after pilot testing, we added three seconds to each condition to ensure that participants had enough time to respond before the experiment advanced (Short = 6 seconds; Long = 9 seconds). If participants did not respond in the allotted time, an error message appeared

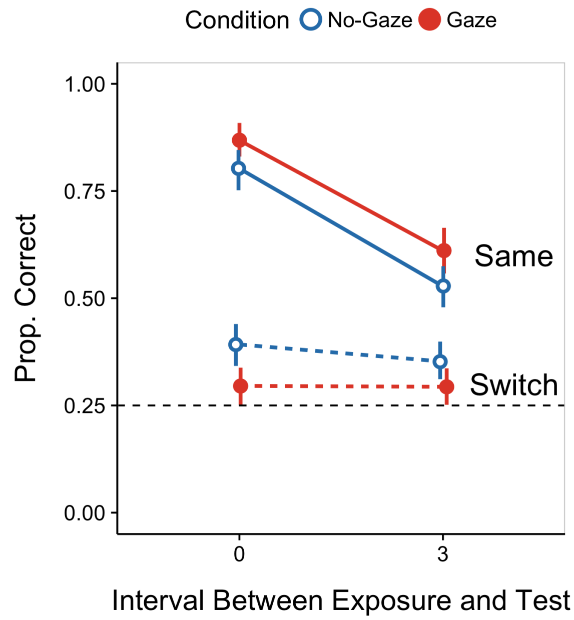


Figure 4.7: Experiment 4.4 results. Accuracy on test trials in Experiment 4 collapsed across the Long and Short inspection time conditions. The dashed line represents chance performance. Color and line type indicate whether there was gaze present on exposure trials. Error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

informing participants that time had run out and encouraged them to respond within the time window on subsequent trials.

4.5.2 Results and Discussion

We did not see strong evidence of an effect of the different inspection times. Thus, all of the results reported here collapse across the short and long inspection time conditions. For all analyses, we removed the trials on which participants did not respond within the fixed inspection time on exposure trials (0.05% of trials).

Exposure Trials

Participants' responses on exposure trials differed from those expected by chance (smallest $\beta = 2.95$, z $= 38.08$, $p < .001$), suggesting that gaze was again effective in directing participants' attention. Similar to Experiment 2, participants were quite likely to use the gaze cue when it was a video of an actress ($M_{0-interval} = 0.93$, $M_{3-interval} = 0.95$).

Test Trials

Figure 6 shows performance on test trials in Experiment 4. In the majority of conditions, participants selected the correct referent at rates greater than chance (smallest β = 0.2, z = 2.2, $p < .05$). However, participants' responses were not different from chance on Switch trials after exposure trials with gaze in the 3-interval condition (β = 0.17, p = 0.06).

We replicate the key finding from Experiments 1-3: after seeing exposure trials with gaze, participants were less accurate on Switch trials (β = 0.9, $p < .001$). Since inspection times were fixed across the Gaze and No-Gaze conditions, this finding provides evidence that the presence of a referential cue did more than just reduce the amount of time participants' spent inspecting the potential word-object links. In contrast to Experiments 2 and 3, visual inspection of Figure 6 suggested that the referential cue provided a boost to accuracy on Same trials. To assess the simple effect of gaze on trial type, we computed pairwise contrasts using the *lsmeans* package in R with a Bonferroni correction for multiple comparisons (Lenth, 2016). Accuracy was higher for Same trials in the Gaze condition (β = 0.49, $p < .001$), but lower for Switch trials (β = -0.41, $p < .001$). The boost in accuracy on Same trials differs from Experiments 2 and 3 and suggests that making inspection times equivalent across conditions allowed the social cue to affect the strength of learners' memory for their candidate hypothesis.

The results of Experiment 4 help to clarify the effect of gaze on memory in our task, providing evidence that the presence of a referential cue did more than just reduce participants' visual inspection time. Instead, gaze reduced memory for alternative word-object links even when people had the same opportunity to visually inspect and encode them. We also found evidence of a boost for learners' memory of their candidate hypothesis in the gaze condition, an effect that we saw at the higher number of referents and the longer intervals in Experiment 1, but that we did not see in Experiments 2 or 3. One explanation for this difference is that in Experiment 4, since participants' use of gaze was independent of the length of exposure trials, inspection times in the gaze condition were longer compared to those in Experiments 1-3. Thus, it could be that the combination of a gaze cue coupled with the opportunity to continue attending to the gaze target led to a boost in performance on Same trials relative to trials without gaze.

4.6 General Discussion

Tracking cross-situational word-object statistics allows word learning to proceed despite the presence of individually ambiguous naming events. But models of cross-situational learning disagree about how much information is actually stored in memory, and the input to statistical learning mechanisms can vary along a continuum of referential uncertainty from unambiguous naming instances to highly ambiguous situations. In the current line of work, we explore the hypothesis that these two factors are fundamentally linked to one another and to the social context in which word learning occurs. Specifically, we ask how cross-situational learning operates over social input that varies the amount of ambiguity in the learning context.

Our results suggest that the representations underlying cross-situational learning are quite flexible. In the absence of a referential cue to word meaning, learners tended to store more alternative word-object links. In contrast, when gaze was present learners stored less information, showing behavior consistent with tracking a single hypothesis (Experiments 1 and 2). Learners were also sensitive to a parametric manipulation of the strength of the referential cue, showing a graded increase in the tendency to use the cue as reliability increased, which in turn resulted in a graded decrease in memory for alternative word-object links (Experiment 3). Finally, learners stored less information in the presence of gaze even when they were shown the objects for the same amount of time (Experiment 4).

In Experiments 2 and 3 reduced memory for alternative hypotheses did not result in a boost to memory for learners' candidate hypothesis. This pattern of data suggests that the presence of a referential cue selectively affected one component of the underlying representation: the number of alternative word-object links, and not the strength of the learners' candidate hypothesis. However, in Experiments 1 and 4, we did see some evidence of stronger memory for learners' initial hypothesis in the presence of gaze: at the higher number of referents and interval conditions (Experiment 1), and when the length of exposure trials was equivalent across the Gaze and No-Gaze conditions (Experiment 4). We speculate that the relationship between the presence of a referential cue and the strength of learners' candidate hypothesis is modulated by how the cue interacts with attention. In Experiment 1, gaze may have provided a boost because, in the absence of gaze, attention would have been distributed across a larger number of alternatives. And, in Experiment 4, gaze may have led to better memory because it was coupled with the opportunity for sustained attention to the gaze target. More work is needed in order to understand precisely when the presence of gaze affects this particular component of the representations underlying cross-situational learning.

In Experiments 1-3, longer inspection times (i.e., more time spent encoding the word-object links during learning) led to better memory at test. We did, however, find slightly different interaction effects across our studies. In Experiment 1, longer inspection times led to higher accuracy in the No-Gaze condition for both Same and Switch trials. In Experiment 2, longer inspection times provided a larger boost to performance on Switch trials compared to Same trials, regardless of gaze condition. Despite these differences, we speculate that inspection time played a similar role across these studies: When a social cue was present, learners' attention was focused and inspection times tended to be shorter, which led to worse performance on Switch trials (i.e., reduced memory for alternative word-object links). Interestingly, in Experiment 4, we found an effect of social cues on memory for alternatives even when participants were given the same opportunity to visually inspect the objects, suggesting that gaze does more than just modulate visual attention during learning.

4.6.1 Relationship to previous work

Why might a decrease in memory for alternatives fail to increase the strength of learners' memory for their candidate hypothesis? One possibility is that participants did not shift their cognitive resources from the set of alternatives to their single hypothesis, but instead chose to use the gaze information to reduce inspection time, thus conserving their resources for future use. Griffiths, Lieder, and Goodman (2015) formalize this behavior by pushing the rationality of computational-level models down to the psychological process level. In their framework, cognitive systems are thought to be adaptive in that they optimize the use of their limited resources, taking the cost of computation (e.g., the opportunity cost of time or mental energy) into account. For example, Vul, Goodman, Griffiths, and Tenenbaum (2014) showed that as time pressure increased in a decision-making task, participants were more likely to show behavior consistent with a less cognitively challenging strategy of matching, rather than with the globally optimal strategy. In the current work, we found that learners showed evidence of altering how they allocated cognitive resources based on the amount of referential uncertainty present during learning, spending less time inspecting alternative word-object links and reducing the number of links stored in memory when uncertainty was low.

Our results fit well with recent experimental work that investigates how attention and memory can constrain infants' statistical word learning. For example, Smith and Yu (2013) used a modified cross-situational learning task to show that only infants who disengaged from a novel object to look at both potential referents were able to learn the correct word–object mappings. Moreover, Vlach and Johnson (2013) showed that 16-month-olds were only able to learn from adjacent cross-situational co-occurrence statistics, and unable to learn from co-occurrences that were separated in time. Both of these findings make the important point that only the information that comes into contact with the learning system can be used for cross-situational word learning, and this information is directly influenced by the attention and memory constraints of the learner. These results also add to a large literature showing the importance of social information for word learning (P. Bloom, 2002; E. V. Clark, 2009) and to recent work exploring the interaction between statistical learning mechanisms and other types of information (Frank, Goodman, & Tenenbaum, 2009; Koehne & Crocker, 2014; Yu & Ballard, 2007). Our findings suggest that referential cues affect statistical learning by modulating the amount of information that learners store in the underlying representations that support learning over time.

Is gaze a privileged cue, or could other, less-social cues (e.g., an arrow) also affect the representations underlying cross-situational learning? On the one hand, previous research has shown that gaze cues lead to more reflexive attentional responses compared to arrows (Friesen, Ristic, & Kingstone, 2004), that gaze-triggered attention results in better learning compared to salience-triggered attention (Wu & Kirkham, 2010), and that even toddlers readily use gaze to infer novel word meanings (Baldwin, 1993). Thus, it could be that gaze is an especially effective cue for constraining word learning since it communicates a speaker's referential intent and is a particularly good way to guide attention. On the other hand, the generative process of the cue – whether it is more or less social in nature – might be less important; instead, the critical factor might be whether the cue effectively reduces uncertainty in the naming event. Under this account, gaze is placed amongst a set of many cues that could produce similar effects as those reported here. Future work could explore a wider range of cues to see if they modulate the representations underlying cross-situational learning in a similar way.

How should we characterize the effect of gaze on attention and memory in our task? One possibility is that the referential cue acts as a filter, only allowing likely referents to contact statistical learning mechanisms (Yu & Ballard, 2007). This 'filtering account' separates the effect of social cues from the underlying computation that aggregates cross-situational information. Another possibility is that referential cues provide evidence about a speaker's communicative intent (Frank et al., 2009). In this model, the learner is reasoning about the speaker and word meanings simultaneously, which places inferences based on social information as part of the underlying computation. A third possibility is that participants thought of the referential cue as pedagogical. In this context, learners assume that the speaker will choose an action that is most likely to increase the learner's belief in the true state of the world (Shafto, Goodman, & Frank, 2012), making it unnecessary to allocate resources to alternative hypotheses. Experiments show that children spend less time exploring an object and are less likely to discover alternative object-functions if a single function is demonstrated in a pedagogical context (Bonawitz et al., 2011). However, because the results from the current study cannot distinguish between these explanations, these questions remain topics for future studies specifically designed to tease apart these possibilities.

4.6.2 Limitations

There are several limitations to the current study that are worth noting. First, the social context that we used was relatively impoverished. Although we moved beyond a simple manipulation of the presence or absence of social information in Experiment 3, we nevertheless isolated just a single cue to reference, gaze. But real-world learning contexts are much more complex, providing learners access to multiple cues such as gaze, pointing, and previous discourse. In fact, Frank, Tenenbaum, and Fernald (2013) analyzed a corpus of parent-child interactions and concluded that learners would do better to aggregate noisy social information from multiple cues, rather than monitor a single cue since no single cue was a consistent predictor of reference. In our data, we did see a more reliable effect of referential cues when we used a video of an actress, which included both gaze and head turn as opposed to the static, schematic stimuli, which only included gaze. It is still an open and interesting question as to how our results would generalize to learning environments that contain a rich combination of social cues.

Second, we do not yet know how variations in referential uncertainty during learning would affect the representations of young word learners, the age at which cross-situational word learning might be particularly important. Recent research using a similar paradigm as our own did not find evidence that 2- or 3-year-olds stored multiple word-object links; instead, children only retained a single candidate hypothesis (Woodard, Gleitman, & Trueswell, 2016). However, performance limitations on children's developing attention and memory systems (Colombo, 2001; Ross-sheehy, Oakes, & Luck, 2003) could make success on these explicit response tasks more difficult. Moreover, our work suggests that different levels of referential uncertainty in naturalistic learning contexts (see Medina, Snedeker, Trueswell, & Gleitman, 2011; Yurovsky & Frank, 2015) might evoke different strategies for information storage, with learners retaining more information as ambiguity in the input increases. Thus, we think that it will be important to test a variety of outcome measures and learning contexts to see if younger learners show evidence of storing multiple word meanings during learning.

In addition, previous work with infants has shown that their attention is often stimulus-driven and sticky (Oakes, 2011), suggesting that very young word learners might not effectively explore the visual scene in order to extract the necessary statistics for storing multiple alternatives. It could be that referential cues play an even more important role for young learners by filtering the input to cross-situational word learning mechanisms and guiding children to the relevant statistics in the input. In fact, recent work has shown that the precise timing of features such as increased parent attention and gesturing towards a named object and away from non-target objects were strong predictors of referential clarity in a naming event (Trueswell et al., 2016). It could be that the statistics available in these particularly unambiguous naming events are the most useful for cross-situational learning.

Finally, the current experiments used a restricted cross-situational word learning scenario, which differs from real-world language learning contexts in several important ways. One, we only tested a single exposure for each novel word-object pairing; whereas, real-world naming events are best characterized by discourse where an object is likely to be named repeatedly in a short amount of time (Frank, Tenenbaum, & Fernald, 2013; Rohde & Frank, 2014). Two, the restricted visual world of 2-8 objects on a screen combined with the forced-choice response format may have biased people to assume that all words in the task must have referred to one of the objects. But, in actual language use, people can refer to things that are not physically co-present (e.g., Gleitman, 1990), creating a scenario where learners would not benefit from storing additional word-object links in the absence of clear referential cues. Finally, we presented novel words in isolation, removing any sentential cues to word meaning (e.g., verb-argument relations). In fact, previous work with adults has shown that cross-situational learning mechanisms only operate in contexts where sentence-level constraints do not completely disambiguate meaning (Koehne & Crocker, 2014). Thus, we need more evidence to understand how the representations underlying cross-situational learning change in response to referential uncertainty at different timescales and in richer language contexts that more accurately reflect real-world learning environments.

4.7 Conclusions

Word learning proceeds despite the potential for high levels of referential uncertainty and despite learners' limited cognitive resources. Our work shows that cross-situational learners flexibly respond to the amount of ambiguity in the input, and as referential uncertainty increases, learners tend to store more word-object links. Overall, these results bring together aspects of social and statistical accounts of word learning to increase our understanding of how statistical learning mechanisms operate over fundamentally social input.

Chapter 5

Integrating statistical and social information during language comprehension and word learning

In this chapter, I present three studies that explore how the presence of a social cue to reference (a speaker's gaze) changes listeners' decisions about visual fixation during language comprehension and word learning. Within our broader active-social framework, these studies ask how the value of information gained from fixating on (i.e., querying) a social partner interacts with learners' developing knowledge of word meanings (i.e., hypotheses) to modulate their information accumulation thresholds (i.e., stopping rules). This work brings together the core elements – active, social, and statistical – of the integrative account described in Chapter 1.

Children process language in complex environments where there are often many things to talk about. How do they understand and learn words despite this noisy input? Statistical learning accounts emphasize that children can aggregate consistent word-object co-occurrences across multiple labeling events to reduce uncertainty over time. Social-pragmatic theories argue that interactions with social partners support learning by reducing ambiguity within individual labeling events. Here, we present three eye-tracking studies that ask how children integrate statistical and social information during real-time language processing. First, children and adults did not delay their gaze shifts to gather a post-nominal social cue to reference (another speaker's eye gaze). Second, when processing novel words, adults fixated more on a speaker who provided a disambiguating gaze and showed stronger recall for word-object links learned via the social cue. Finally, in contrast to the familiar word context, children and adults fixated longer on a speaker who produced a gaze cue when labeling novel objects, which, in turn, led to increased looking to the named object and less looking to the other objects in the scene. Moreover, children, but not adults, increased their looking to the interlocutor throughout the experiment. Together, these results suggest that learners flexibly integrate their knowledge of object labels to decide whether to seek social information, which then shapes the information that comes into contact with statistical learning mechanisms.

Conclusion

In this dissertation, I proposed a framework for understanding children's information-seeking decisions within social contexts. The core of the argument is that the presence of other people can change the *availability* and *usefulness* of information-seeking behaviors by shaping learners' goals, hypotheses, actions, answers, and thresholds for stopping information gathering. Following the theoretical framework, I presented a set of empirical studies that explored whether the dynamics of children's real-time information selection via their eye movements flexibly adapts to gather social information that supports language processing.

Chapter 2 investigated how children learning American Sign Language (ASL) distributed visual attention between the linguistic signal and referents, which both compete for visual attention. Similar to children learning spoken language, ASL learners shifted gaze away from a social partner to seek objects before sign offset, providing evidence that, despite channel competition, language drove rapid shifts in visual attention to named referents. Chapter 3 extended the sign language research by directly comparing ASL learners' gaze dynamics to those of children learning spoken English using parallel real-time language comprehension tasks. Chapter 3 also presented a comparison of English-learning children and adults' eye movements in noisy vs. clear auditory contexts. In both the ASL and noisy speech cases, listeners adapted their gaze to seek additional language-relevant information from social partners before shifting to seek a named referent. Chapters 4 and 5 explored how eye movements change when children and adults processed familiar and novel words accompanied by social cues to reference. Taken together, the social gaze work suggests that children integrate their uncertainty over word-object mappings to decide when to seek social information, which in turn, modulates the input to statistical word learning.

The integrative framework and empirical work described here are limited in several important ways. First, the majority of this research tested binary hypotheses of behavior change – i.e., sign vs. spoken language; noisy vs. clear speech; word learning with vs. without social gaze – to answer the question of whether children would flexibly adapt their information seeking in response to changes in their processing environments. Chapters 2-5 present evidence across a diverse set of case studies that children's real-time information seeking is quite flexible. However, to move the integrative framework forward, we would want to develop a fully-specified model that could make quantitative predictions about how social contexts will change the utility of information-seeking behaviors. This step will require formalizing the notions of value and cost of information-seeking actions in a modeling framework that can incorporate the effects of reasoning about other people's mental states.

We have taken some initial steps towards this goal by developing a model of active-social learning that integrates ideas from Optimal Experiment Design (OED) with formalizations of recursive social reasoning from Bayesian models of pragmatic language interpretation (N. D. Goodman & Frank, 2016b). We found that this integrated model was able to reproduce the qualitative patterns in adults' decisions of whether to forego information seeking in favor of more immediately rewarding actions when their social partner highlighted performance and presentational goals (E. J. Yoon, MacDonald, Asaba, Gweon, & Frank, 2018). The integrative framework described here directly inspired this line of research, and I hope that future versions of the model will be able to generate graded, testable predictions for behavior across a variety of domains – e.g., eye movements, early vocalizations, and verbal question asking.

Second, we used one particular formalization of active inquiry. The OED model focuses on learners' information-seeking decisions given a specific goal to learn and a set of candidate hypotheses. Other computational frameworks have formalized active learning in different ways. For example, foraging models pursue the analogy that human information seeking is similar to animals' decisions about where and how long to look for food if they were trying to maximize caloric intake while minimizing their effort and time (see Pirolli & Card (1999) for a review). Cognitive scientists have successfully modeled a range of behaviors as a form of spatial foraging, such as searching for semantic concepts in memory (Hills, Jones, & Todd, 2012) and decisions about where to direct visual attention in real-time (Manohar & Husain, 2013). In addition to these search models, recent work on curiosity-based learning in developmental robotics has created algorithms that optimize intrinsic estimates of learning progress. This formalization creates systems that focus on seeking activities and stimuli of intermediate complexity where learners' predictions are steadily improving, and uncertainty is steadily decreasing (Oudeyer & Smith, 2016). One of the challenges for researchers trying to integrate active and social learning is that the space of possible connections is quite large. By constraining our framework to active decision making, we were able to make some progress on an important sub-component of a larger set of children's information-seeking behaviors. Future theoretical work, however, should consider possible connections between social learning phenomena and the foraging/curiosity-based learning frameworks.

Third, our ultimate goal for the active-social framework is to incorporate effects at a developmental timescale. The experiments in this thesis, however, often treated children and adults as two endpoints on a continuum, exploring parallels and differences between children's information seeking and our best estimate of the mature state of the language processing system. We did find some clear patterns of developmental change. In Chapter 3, adults were faster to respond to familiar words, generated more language-consistent shifts, and produced fewer early shifts before accumulating enough information. Older children were also faster to respond than younger children but did not generate more language-consistent shifts overall. Older children did, however, produce fewer early, non-language-driven gaze shifts. This pattern of results suggests that what might be developing is an ability to inhibit a behavior – shifting gaze away from the language source – that reduces access to information that is useful for figuring out the identity of a named referent. Prior work also shows that children develop greater flexibility in ignoring irrelevant information to focus on parts of the meaningful parts of a sentence (Zangl & Fernald, 2007). But it is still an open question as to how children's visual information seeking might change as they become more efficient in processing words and develop their skill in focusing on relevant information in their environment.

In addition to change at the developmental timescale, the final experiment in Chapter 5 represents an exception where we measured adaptation of information seeking over the slightly longer timescale of multiple exposures to novel word-object links and in the context of highly familiar words, which children learned through exposure to many prior labeling events in their day-to-day lives. While the study in Chapter 5 is a useful first step, future work should measure change over a longer timescale by densely sampling children at different ages and points of development. For example, it would be useful to know the effect of children's rapidly improving productive language skills, which increases the set of information-seeking actions available by allowing children to ask verbal questions. One prediction based on our framework is that seeking social information via eye movements should become less useful when children can produce the verbal question "What is this thing called?" since it has a higher probability of returning useful information. Another example is children's rapid theory of mind development. Our framework predicts that young children should focus more on learning goals if they are less skilled at reasoning about others' beliefs. But, as their social reasoning abilities mature and their social environments become more complex, children may forego information seeking actions that make them appear incompetent to their social partners.

Fourth, our account is currently underdeveloped concerning individual differences. That is, the model was designed to explore general principles about how qualitative changes to the social environment might shape children's information seeking actions. However, it is possible to use the active-social framework to understand how individual differences in children's input and cognitive abilities might interact to shape how they decide to seek information from social partners. For example, prior research has found that adults vary in the proportion of unambiguous naming episodes they provide, with some parents rarely providing highly informative contexts and others' doing so relatively more often (Cartmill et al., 2013). Within our active-social framework, this differential experience could be instantiated as children learning a model of the probability of getting a high-quality answer when they ask a question. If children do not expect an answer is likely, then this should reduce information seeking even if there is a social partner available. We did find some evidence of this effect in Chapters 4 and 5 where adults were less likely to use an unreliable social cue to reference and where children looked more to a social partner who provided gaze cue than to one who did not.

Individual differences in cognitive abilities could also be included in our model. Prior research shows variability in children's theory of mind and inhibitory control abilities (Carlson & Moses, 2001), in addition to the considerable variability in language processing skill (Marchman & Fernald, 2008). Within our active-social framework, children with a more-developed theory of mind skill might place a higher weight on pursuing social goals over and above informational goals, taking actions that maintain others' beliefs about their abilities. It could also be that stronger perspective-taking skills help children reason about the probability of getting a quality answer from another person, thus modulating whom they choose to ask questions (e.g., seeking information from an expert vs. a novice). Another compelling possibility is that children who have stronger domain-general processing abilities are better able to update their beliefs based on the information they receive, and thus reducing the amount of time they spend seeking information from social partners. These are all interesting, open questions for future research that fall out of the integrative active-social framework proposed in this thesis.

Finally, the empirical research described here aimed to understand how children's information-seeking behaviors adapt to support their language processing. To accomplish this, we measured changes in children's gaze dynamics during language comprehension and word learning in simplified environments. This approach has the benefit of providing a high degree of experimenter control and a relatively well-understood hypothesis linking observable behavior (eye movements) to underlying psychological constructs (e.g., lexical access) (Tanenhaus, Magnuson, Dahan, & Chambers, 2000). The risk, however, is that the responses that we can measure in the lab do not reflect behaviors that support children's learning in their natural environments. That is, children acquire their first language from conversations where social partners produce contingent responses and take actions that control the flow of children's learning experience. This gap suggests two critical next steps for the research described here: (1) measure changes in children's information seeking within free-flowing social interactions with their caregivers (see Franchak, Kretch, & Adolph (2018) for a recent example of this approach using head-mounted eye trackers), and (2) develop more realistic lab-based experiments that incorporate behaviorally-relevant features of children's learning environments such as contingent responding to children's actions (see Benitez & Saffran (2018) for an example of studying word learning using a gaze-contingent eye-tracking paradigm).

In sum, we set out to explore how children's eye movements adapt to a wide range of social contexts during two ecologically-relevant tasks: familiar language comprehension and novel word learning. We found that children could adapt their gaze to seek relevant social information when it was useful for language processing. Moreover, children and adults showed evidence of differential learning of new words when social gaze guided their visual attention. This work highlighted two critical, open challenges for a framework of information-seeking within social contexts: (1) develop a precise quantitative model of how social learning can change the utility of information-seeking behaviors, and (2) move beyond highly-constrained lab experiments to document information seeking behaviors in children's natural learning environments. Despite these challenges, the integrative framework presented in this thesis represents a way forward for understanding how children's information seeking adapts to the wide variety of social environments in which children acquire their first language.

Appendix A

Supplementary materials for Chapter 1

A.1 Mathematical details of Optimal Experiment Design

This supplement contains the mathematical details of the OED approach as described in Coenen, Nelson, & Gureckis (2017). The goal is to provide a concrete foundation for the conceptual analysis of how social learning contexts can influence different components of active learning.

The OED model quantifies the *expected utility* of different information seeking actions. Formally, the set of queries is defined as $Q_1, Q_2, ..., Q_n = \{Q\}$. The expected utility of each query ($EU(Q)$) is a function of two factors: (1) the probability of obtaining a specific answer $P(a)$ weighted by (2) the usefulness of that answer for achieving the learning goal $U(a)$.

$$EU(Q) = \sum_{a \in q} P(a)U(a)$$

There are a variety of ways to define the usefulness function to score each answer. An exhaustive review is beyond the scope of this paper (for a detailed analysis of different approaches, see Nelson (2005)). One standard method is to use *information gain*, which is defined as the change in the learner's overall uncertainty (difference in entoropy) before and after receiving an answer.

$$U(a) = ent(H) - ent(H|a)$$

Where $ent(H)$ is defined using Shannon entropy[1] (MacKay, 2003), which provides a measure of the overall amount of uncertainty in the learner's beliefs about the candidate hypotheses.

$$ent(H) = -\sum_{a \in A} P(h) log_2 P(h)$$

The conditional entropy computation is the same, but takes into account the change in the learner's beliefs after seeing an answer.

$$ent(H|a) = -\sum_{h \in H} P(h|a) log P(h|a)$$

To calculate the change in the learner's belief in a hypothesis $P(h|a)$, we use Bayes rule.

$$P(h|a) = \frac{P(h)P(a|h)}{P(a)}$$

If the researcher defines all these parts of the OED model (hypotheses, questions, answers, and the usefulness function), then selecting the optimal query is straightforward. The learner performs the expected utility computation for each query in the set of possible queries and picks the one that maximizes utility. In practice, the learner considers each possible answer, scores the answer with the usefulness function, and weights the score using the probability of getting that answer.

Before reviewing the behavioral evidence for OED-like reasoning in adults and children, I will present a worked example of how to compute the expected utility of a single query. The goal is to provide simple calculations that illustrate how reasoning about hypotheses, questions, and answers can lead to selecting useful actions. This example is slightly modified from Nelson (2005).

Imagine that you are a biologist, and you come across a new animal that you think belongs to one of two species: "glom" or "fizo." You cannot directly query the category identity, but you can gather information about the presence or absence of two features (eats meat? or is nocturnal?) that you know from prior research are more or less likely for each of the species. The following probabilities

[1] Shannon entropy is a measure of unpredictability or amount of uncertainty in the learner's probability distribution over hypotheses. Intuitively, higher entropy distributions are more uncertain and harder to predict. For example, if the learner believes that all hypotheses are equally likely, then they are in a state of high uncertainty/entropy. In contrast, if the learner firmly believes in one hypothesis, then uncertainty/entropy is low.

summarise this prior knowledge:

- $P(eatsMeat \mid glom) = 0.1$
- $P(eatsMeat \mid fizo) = 0.9$
- $P(nocturnal \mid glom) = 0.3$
- $P(nocturnal \mid fizo) = 0.5$

You also know from previous research that the probability of seeing a glom or a fizo in the wild is:

- $P(glom) = 0.7$
- $P(fizo) = 0.3$

Which feature should you test: eats meat? or sleeps at night? Intuitively, it seems better to test whether the creature eats meat because an answer to this question provides good evidence about whether the animal is a fizo since $P(eatsMeat \mid fizo) = 0.9$. However, the OED computation allows the biologist to go beyond this intuition and compute precisely how much better it is to ask the "eats meat?" question. All the scientist has to do is pass her knowledge about the hypotheses and features through the expected utility computation.

Here are the steps of the OED computation for calculating the utility of the "eats meat?" question. First, we use Bayes rule to calculate how much our beliefs would change if we received a "yes" or a "no" answer.[2]

$$P(glom \mid eatsMeat) = \frac{P(eatsMeat \mid glom) \times P(glom)}{P(eatsMeat)} = \frac{0.1 \times 0.7}{0.34} = 0.21$$

Next, we calculate the uncertainty over the Species hypothesis before doing any experiment. We do this by computing the prior entropy.

[2]Note that the $P(eatsMeat)$ term is computed by taking $P(eatsMeat) = [P(eatsMeat \mid glom)\ timesP(glom)] + [P(eatsMeat \mid fizo) \times P(fizo)] = (0.1 \times 0.7) + (0.9 \times 0.3) = 0.34$

$$
\begin{aligned}
ent(Species) &= -\sum_{h \in H} P(h) \times log_2 P(h) \\
&= [-P(glom) \times log_2 P(glom)] + [-P(fizo) \times log_2 P(fizo)] \\
&= [-(0.7 \times log_2(0.7)] + [-(0.3 \times log_2(0.3)] \\
&= 0.88
\end{aligned}
$$

To calculate information gain, we also need to compute our uncertainty over hypotheses conditional on seeing each answer, or the posterior entropy. First, for the "yes" answer:

$$
\begin{aligned}
ent(Species|eatsMeat = yes) &= -\sum_{a \in A} P(Species \mid eatsMeat = yes) \times log_2 P(species \mid eatsMeat = yes) \\
&= [0.21 \times log_2(0.21)] + [0.79 \times log_2(0.79)] \\
&= 0.73
\end{aligned}
$$

We use the difference between the prior and posterior entropy to compute the utility of the "yes" answer.

$$
\begin{aligned}
U(a = yes) &= ent(Species) - ent(Species \mid eatsMeat = yes) \\
&= 0.88 - 0.73 \\
&= 0.15
\end{aligned}
$$

Next, we do the same process for the "no" answer. First, we calculate the posterior entropy.

$$
\begin{aligned}
ent(Species|eatsMeat = no) &= -\sum_{a \in A} P(Species \mid eatsMeat = no) \times log_2 P(species \mid eatsMeat = no) \\
&= [0.95 \times log_2(0.95)] + [0.05 \times log_2(0.05)] \\
&= 0.27
\end{aligned}
$$

Again, we use the difference between the prior and posterior entropy to compute the utility of the "no" answer.

$$\begin{aligned} U(a = no) &= ent(Species) - ent(Species \mid eatsMeat = no) \\ &= 0.88 - 0.27 \\ &= 0.61 \end{aligned}$$

Note that the $U(a = no) > U(a = yes)$. This captures the intuition that learning that the animal does not eat meat would provide strong evidence against the "fizo" hypothesis since $P(eatsMeat \mid fizo) = 0.9$. Finally, to compute the overall expected information gain for the **"eats meat?" question**, we weight the utility of each answer by its probability:

$$\begin{aligned} EU(Q = eatsMeat) &= \sum_{a \in A} P(a)U(a) \\ &= [P(eatsMeat = yes) \times U(eatsMeat = yes)] + \\ &\qquad [P(eatsMeat = no) \times U(eatsMeat = no)] \\ &= [0.34 \times 0.15] + [0.66 \times 0.61] \\ &= 0.46 \end{aligned}$$

If we performed the same steps to calculate the expected utility of the "sleeps at night?" question, we get $EU(Q = sleepsNight) = 0.026$. So if the biologist wants to maximize the chance of gaining useful information, she should select the "eats meat?" experiment since $EU(Q = eatsMeat) > EU(Q = sleepsNight)$.

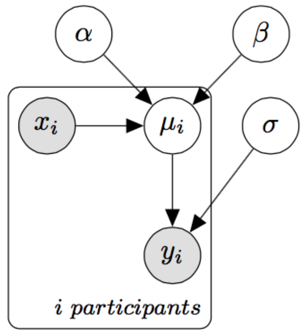
Appendix B

Supplementary materials for Chapter 2

In this appendix, we present four pieces of supplemental information. First, we provide details about the Bayesian models used to analyze the data. Second, we present a sensitivity analysis that provides evidence that the estimates of the associations between age/vocabulary and accuracy/reaction time (RT) are robust to different parameterizations of the prior distribution and different cutoffs for the analysis window. Third, we present the results of a parallel set of analyses using a non-Bayesian approach to show that these results are consistent regardless of choice of analytic framework. And fourth, we present two exploratory analyses measuring the effects of phonological overlap and iconicity on RT and accuracy. In both analyses, we did not see evidence that these factors changed the dynamics of eye movements during ASL processing

B.1 Model Specifications

Our key analyses use Bayesian linear models to test our hypotheses of interest and to estimate the associations between age/vocabulary and RT/accuracy. Figure S1 (Accuracy) and S2 (RT) present graphical models that represent all of the data, parameters, and other variables of interest, and their dependencies. Latent parameters are shown as unshaded nodes while observed parameters and data are shown as shaded nodes. All models were fit using JAGS software (Plummer, 2003) and adapted



$$
\begin{aligned}
\alpha &\sim \text{Gaussian}(0,\, 10)\\
\beta &\sim \text{TruncatedGaussian}_{+}(0,\, 1)\\
\sigma &\sim \text{Uniform}(1/1000,\, 1000)\\
\mu_i &\leftarrow\ \alpha + \beta * x_i\\
y_i &\sim\ \text{Gaussian}(\mu_i,\, \sigma)
\end{aligned}
$$

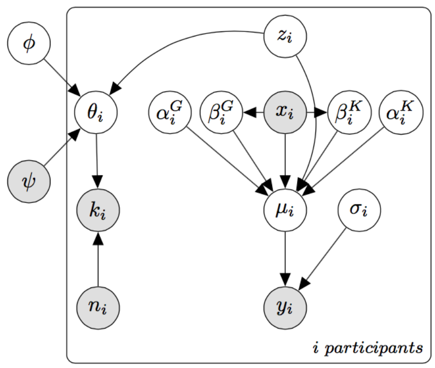
Figure B.1: Graphical model representation of the linear regression used to predict accuracy. The shaded nodes represent observed data (i.e., each participant's age, vocabulary, and mean accuracy). Unshaded nodes represent latent parameters (i.e., the intercept and slope of the linear model).

from models in Kruschke (2014) and Lee and Wagenmakers (2014).

B.1.1 Accuracy

To test the association between age/vocabulary and accuracy we assume each participant's mean accuracy is drawn from a Gaussian distribution with a mean, μ, and a standard deviation, σ. The mean is a linear function of the intercept, α, which encodes the expected value of the outcome variable when the predictor is zero, and the slope, β, which encodes the expected change in the outcome with each unit change in the predictor (i.e., the strength of association).

For α and σ, we use vague priors on a standardized scale, allowing the model to consider a wide range of plausible values. Since the slope parameter β is critical to our hypothesis of a linear association, we chose to use an informed prior: that is, a truncated Gaussian distribution with a mean of zero and a standard deviation of one on a standardized scale. Centering the distribution at zero is conservative and places the highest prior probability on a null association, to reduce the chance that our model overfits the data. Truncating the prior encodes our directional hypothesis that accuracy should increase with age and larger vocabulary size. And using a standard deviation of one constrains the plausible slope values, thus making our alternative hypothesis more precise. We constrained the slope values based on previous research with children learning spoken language showing that the average gain in accuracy for one month of development between 18-24 months to be ˜1.5% (Fernald, Zangl, Portillo, & Marchman, 2008).



$$\alpha \sim \text{Gaussian}(0,\, 10)$$
$$\beta \sim \text{TruncatedGaussian}_{-}(0,\, 1)$$
$$\sigma_i \sim \text{Uniform}(1/1000,\, 1000)$$
$$z_i \sim \text{Bernoulli}(0.5)$$
$$\phi \sim \text{Uniform}(0.5,\, 1)$$
$$\psi \leftarrow 0.5$$
$$\theta_i \leftarrow \begin{cases} \phi & \text{if } z_i = 1 \\ \psi & \text{if } z_i = 0 \end{cases}$$
$$k_i \sim \text{Binomial}(\theta_i,\, n_i)$$
$$\mu_i \leftarrow \begin{cases} \alpha_i^K + \beta_i^K * x_i & \text{if } z_i = 1 \\ \alpha_i^G + \beta_i^G * x_i & \text{if } z_i = 0 \end{cases}$$
$$y_i \sim \text{Gaussian}(\mu_i,\, \sigma_i)$$

Figure B.2: Graphical model representation of the linear regression plus latent mixture model (i.e., guessing model). The model assumes that each individual participant's first shift is either the result of guessing or knowledge. And the latent indicator z_i determines whether that participant is included in the linear regression estimating the association between age/vocabulary and RT.

B.1.2 Reaction Time

The use of RT as a processing measure is based on the assumption that the timing of a child's first shift reflects the speed of their incremental language comprehension. Yet, some children have a first shift that seems to be unassociated with this construct: their first shift behavior appears random. We quantify this possibility for each participant explicitly (i.e., the probability that the participant is a "guesser") and we create an analysis model where participants who were more likely to be guessers have less of an influence on the estimated relations between RT and age/vocabulary.

To quantify each participant's probability of guessing, we computed the proportion of signer-to-target (correct) and signer-to-distracter (incorrect) shifts for each child. We then used a latent mixture model in which we assumed that the observed data, k_i, were generated by two processes (guessing and knowledge) that had different overall probabilities of success, with the "guessing group" having a probability of 50%, ψ, and the "knowledge" group having a probability greater than 50%, ϕ. The group membership of each participant is a latent indicator variable, z_i, inferred based on that participant's proportion of correct signer-to-target shifts relative to the overall proportion of correct shifts across all participants (see Lee & Wagenmakers (2014) for a detailed discussion of this modeling approach). We then used each participant's inferred group membership to determine whether they

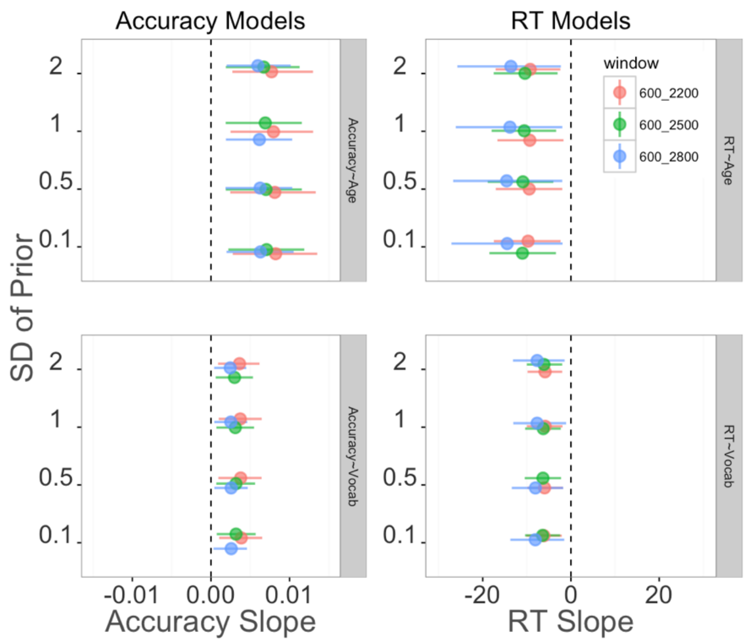


Figure B.3: Coefficient plot for the slope parameter for four different parameterizations of the prior and for three different analysis windows. Each panel shows a different model. Each point represents a coefficient measuring the strength of association between the two variables. Error bars are 95% HDIs around the coefficient. Color represents the three different analysis windows.

were included in the linear regression. In sum, the model allows participants to contribute to the estimated associations between age/vocabulary and RT proportional to our belief that they were guessing.

As in the Accuracy model, we use vague priors for α and σ on a standardized scale. We again use an informed prior for β, making our alternative hypothesis more precise. That is, we constrained the plausible slope values based on previous research with children learning spoken language showing that the average gain in RT for one month of development between 18-24 months to be ˜30 ms (Fernald, Zangl, Portillo, & Marchman, 2008).

Table B.1: Bayes Factors for all four linear models fit to three different analysis windows using four different parameterizations of the prior distribution for the slope parameter.

Analysis window	**SD Slope**	**Acc˜Age**	**Acc˜Vocab**	**RT˜Age**	**RT˜Vocab**
600 âĂŞ 2200 ms	3.2	6.2	3.7	2.4	4.1
NA	1.4	14.1	5.5	3.5	8.6
NA	1.0	19.4	8.9	5.0	9.2
NA	0.7	22.7	11.6	7.8	17.0
600 âĂŞ 2500 ms	3.2	11.0	2.3	5.6	6.1
NA	1.4	9.7	4.0	13.8	10.5
NA	1.0	12.8	6.8	12.5	18.2
NA	0.7	15.6	6.8	17.9	20.7
600 âĂŞ 2800 ms	3.2	6.0	1.1	1.2	1.4
NA	1.4	10.7	2.6	3.5	4.7
NA	1.0	13.5	4.0	3.7	4.0
NA	0.7	15.2	4.6	5.5	5.6

B.2 Sensitivity Analysis: Prior Distribution and Window Selection

We conducted a sensitivity analysis to show that our parameter estimates for the associations between accuracy/RT and age/vocabulary are robust to decisions about (a) the analysis window and (b) the specification of the prior distribution on the slope parameter. Specifically, we varied the parameterization of the standard deviation on the slope, allowing the model to consider a wider or narrower range of values to be plausible a priori. We also fit these different models to two additional analysis windows +/- 300 ms from the final analysis window: 600-2500 ms (the middle 90% of the RT distribution in our experiment). Figure S3 shows the results of the sensitivity analysis, plotting the coefficient for the β parameter in each model for the three different analysis windows for each specification of the prior. All models show similar coefficient values, suggesting that inferences about the parameters are not sensitive to the exact form of the priors. Table S1 shows the Bayes Factors for all models across three analysis windows and fit using four different vales for the slope prior. The Bayes Factor only drops below 3 when the prior distribution is quite broad (standard deviation of 3.2) and only for the longest analysis window (600-2800 ms). In sum, the strength of evidence for a linear association is robust to the choice of analysis window and prior specification.

Table B.2: Results for the four linear models fit using Maxiumum Likelihood Estimation. All p-values are one-sided to reflect our directional hypotheses about the VLP measures improving over development.

Model specification	Mean Beta value	std. error	t-statistic	p-value
logit(accuracy) ˜ age + hearing status	0.003	0.012	2.59	0.008
logit(accuracy) ˜ vocabulary + hearing status	0.002	0.006	2.27	0.015
RT ˜ age + hearing status	-10.050	4.620	-2.17	0.019
RT ˜ vocabulary + hearing status	-6.340	2.180	-2.91	0.003

B.3 Parallel set of non-Bayesian analyses

First, we compare Accuracy and RT of native hearing and deaf signers using a Welch Two Sample t-test and do not find evidence that these groups are different (Accuracy: $t(28) = 0.75$, $p = 0.45$, 95% CI on the difference in means [-0.07, 0.14]; RT: $t(28) = 0.75$, $p = 0.46$, 95% CI on the difference in means [-125.47 ms, 264.99 ms].

Second, we test whether children and adults tend to generate saccades away from the central signer prior to the offset of the target sign. To do this, we use a One Sample t-test with a null hypothesis that the true mean is not equal to 1, and we find evidence against this null (Children: $M = 0.88$, $t(28) = -2.92$, $p = 0.007$, 95% CI [0.79, 0.96]; Adults: $M = 0.51$, $t(15) = -6.87$, $p < 0.001$, 95% CI [0.35, 0.65])

Third, we fit the four linear models using MLE to estimate the relations between the processing measures on the VLP task (Accuracy/RT) and age/vocabulary. We follow recommendations from Barr (2008) and use a logistic transform to convert the proportion accuracy scores to a scale more suitable for the linear model.

B.4 Analyses of phonological overlap and iconicity

First, we analyzed whether phonological overlap of our item-pairs might have influenced adults and children's RTs and accuracy. Signs that are higher in phonological overlap might have been

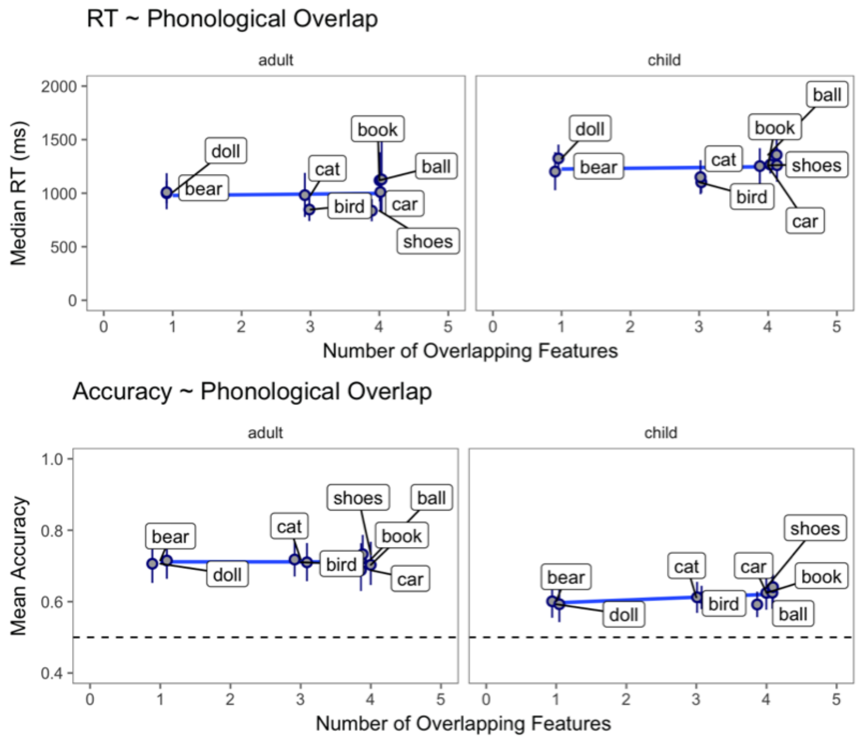


Figure B.4: Scatterplot of the association between degree of phonological overlap and RT (top row) and accuracy (bottom row) for both adults (left column) and children (right column). The blue line represents a linear model fit.

more difficult to process because they are more confusable. Here, phonological overlap is quantified as the number of features (e.g., Selected Fingers, Major Location, Movement, Sign Type) that both signs shared. Values were taken from a recently created database (ASL-LEX) of lexical and phonological properties of nearly 1,000 signs of American Sign Language (Caselli et al., 2017). Our item-pairs varied in degree of overlap from 1-4 features. We did not see evidence that degree of phonological overlap influenced either processing measure in the VLP task. Next, we performed a parallel analysis, exploring whether the iconicity of our signs might have influenced adults and children's RT and accuracy. It is possible that highly iconic signs might be easier to process because of the visual similarity to the target object. Again, we used ASL-LEX to quantify the iconicity of our signs. To generate these values, native signers were asked to explicitly rate the iconicity of each sign on a scale of 1-7, with 1 being not iconic at all and 7 being very iconic. Similar to the phonological overlap analysis, we did see evidence that degree of iconicity influenced either processing measure

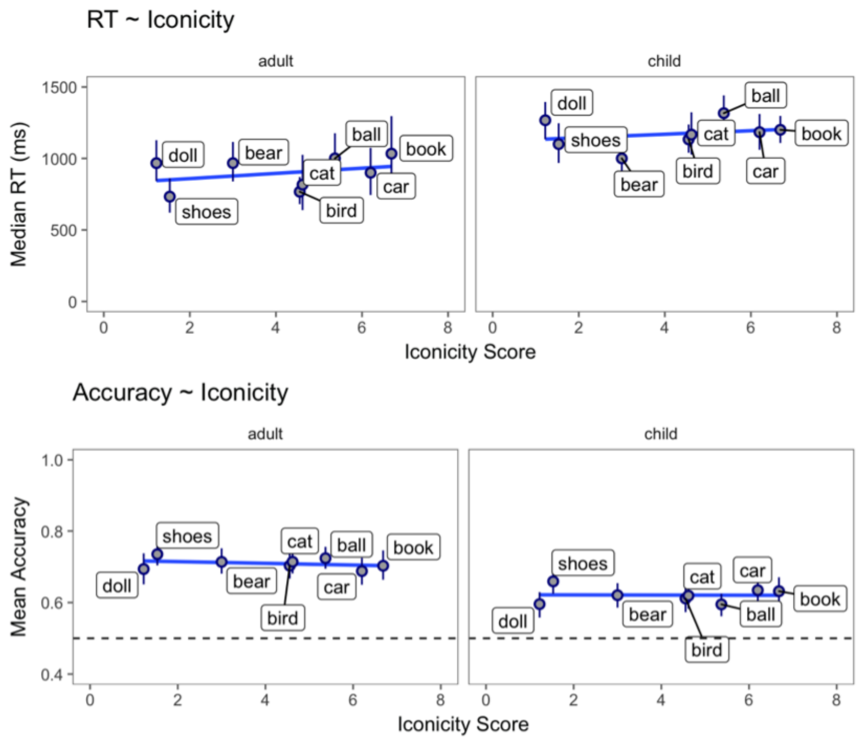


Figure B.5: Scatterplot of the association between degree of iconicity and RT (top row) and accuracy (bottom row) for both adults (left column) and children (right column). The blue line represents a linear model fit.

for either age group in the VLP task.

Appendix C

Supplementary materials for Chapter 3

C.1 Model output for Experiment 3.1

C.2 Model output for Experiment 3.2

Appendix D

Supplementary materials for Chapter 4

D.1 Analytic model specifications and output

D.1.1 Experiment 1

Table A1. Length of inspection times on exposure trials in Experiment 1 as a function of gaze, interval, and number of referents

`Log(Inspection time)` $\sim$ `(Gaze + Log(Interval) + Log(Referents))`^{2} `+ (1 | subject)`

term	estimate	std.error	t.value	p.value	
Intercept	0.83	0.10	8.19	$<$ .001	***
Gaze Condition	0.16	0.11	1.48	0.138	
Log(Interval)	0.06	0.05	1.33	0.184	
Log(Referents)	0.34	0.04	7.91	$<$ .001	***
Gaze Condition*Log(Interval)	-0.08	0.03	-2.86	0.004	**
Gaze Condition*Log(Referent)	-0.27	0.04	-6.01	$<$ .001	***
Log(Interval)*Log(Referent)	-0.00	0.02	-0.19	0.849	

Table A2. Accuracy on test trials in Experiment 1 with inspection times on exposure trials included as a predictor

`Correct` $\sim$ `(Trial Type + Gaze + Log(Interval) + Log(Referents) +`

`Log(Inspection Time))`^{2} `+ offset(logit(`$^1/_{Referents}$`)) + (TrialType | subject)`

term	estimate	std.error	z.value	p.value	
Intercept	2.89	0.34	8.49	$< .001$	***
Switch Trial	-1.45	0.25	-5.76	$< .001$	***
Gaze Condition	0.12	0.27	0.43	0.669	
Log(Interval)	-0.47	0.11	-4.15	$< .001$	***
Log(Referents)	0.05	0.14	0.39	0.693	
Log(Inspection Time)	0.20	0.15	1.38	0.169	
Switch Trial*Gaze Condition	-1.02	0.13	-7.86	$< .001$	***
Switch Trial*Log(Interval)	0.52	0.06	9.39	$< .001$	***
Switch Trial*Log(Referent)	-0.62	0.09	-6.67	$< .001$	***
Switch Trial*Log(Inspection Time)	0.09	0.07	1.36	0.174	
Gaze Condition*Log(Interval)	0.09	0.06	1.61	0.107	
Gaze Condition*Log(Referent)	0.36	0.10	3.68	$< .001$	***
Gaze Condition*Log(Inspection Time)	-0.17	0.07	-2.55	0.011	*
Log(Interval)*Log(Referent)	-0.05	0.04	-1.26	0.207	
Log(Interval)*Log(Inspection Time)	0.02	0.03	0.54	0.589	
Log(Referents)*Log(Inspection Time)	0.05	0.05	0.94	0.345	

D.1.2 Experiment 2

Table A3. Length of inspection times on exposure trials in Experiment 2 as a function of gaze and interval

`Log(Inspection time) ~ Gaze * Log(Interval) + (1 | subject)`

term	estimate	std.error	t.value	p.value	
Intercept	3.90	0.08	50.69	< .001	***
Gaze Condition	-1.10	0.05	-20.90	< .001	***
Log(Interval)	-0.48	0.05	-8.77	< .001	***
Gaze Condition*Log(Interval)	-0.02	0.04	-0.60	0.549	

Table A4. Accuracy on test trials in Experiment 2 with inspection times on exposure trials included as a predictor

$\texttt{Correct} \sim \texttt{(Trial Type + Gaze + Log(Interval) + Log(Inspection Time))}^2 \texttt{ +}$

$\texttt{offset(logit(}^1/_{Referents}\texttt{)) + (TrialType | subject)}$

term	estimate	std.error	z.value	p.value	
Intercept	3.51	0.29	12.13	< .001	***
Gaze Condition	0.13	0.23	0.58	0.559	
Switch Trial	-3.12	0.26	-12.21	< .001	***
Log(Interval)	-0.88	0.14	-6.34	< .001	***
Log(Inspection Time)	0.15	0.13	1.14	0.255	
Switch Trial*Gaze Condition	-0.54	0.17	-3.21	0.001	**
Gaze Condition*Log(Interval)	0.16	0.09	1.85	0.064	.
Gaze Condition*Log(Inspection Time)	-0.14	0.10	-1.37	0.172	
Switch Trial*Log(Interval)	0.77	0.10	8.00	< .001	***
Switch Trial*Log(Inspection Time)	0.21	0.11	1.96	0.05	.
Log(Interval)*Log(Inspection Time)	0.04	0.06	0.77	0.44	

D.1.3 Experiment 3

Table A5. Accuracy on exposure trials in Experiment 3 as a function of reliability condition and participants' subjective reliability judgments

`Correct-Exposure` $\sim$ `Reliability Condition * Subjective Reliability + offset(logit(`$^{1}/_{Referents}$`)) + (1 | subject)`

term	estimate	std.error	z.value	p.value	
Intercept	3.07	0.98	3.13	0.002	**
Reliability Condition	3.28	1.50	2.19	0.029	*
Subjective Reliability	7.26	1.73	4.21	$<$ .001	***
Reliability Condition*Subjective Reliability	-4.58	2.72	-1.68	0.093	.

Table A6. Accuracy on test trials in Experiment 3 as a function of reliability condition

`Correct` $\sim$ `Trial Type * Reliability Condition + offset(logit(`$^{1}/_{Referents}$`)) + (Trial Type | subject)`

term	estimate	std.error	z.value	p.value	
Intercept	4.70	0.36	13.10	$<$ .001	***
Trial Type	-3.95	0.36	-10.92	$<$ .001	***
Reliability Condition	0.38	0.37	1.03	0.302	
Reliability Condition*Trial Type	-0.76	0.38	-2.01	0.044	*

Table A7. Accuracy on test trials in Experiment 3 as a function of reliability condition and participants' use of gaze on exposure trials

`Correct` $\sim$ `(Trial Type + Reliability Condition + Correct-Exposure)`^{2}

`+ offset(logit(`$^1/_{Referents}$`)) + (Trial Type | subject)`

term	estimate	std.error	z.value	p.value	
Intercept	4.50	0.39	11.59	< .001	***
Correct Exposure	0.07	0.29	0.26	0.796	
Trial Type	-2.70	0.38	-7.07	< .001	***
Reliability Condition	-0.43	0.44	-0.98	0.325	
Correct Exposure*Trial Type	-1.43	0.26	-5.41	< .001	***
Correct Exposure*Reliability	0.97	0.33	2.92	0.004	**
Reliability Condition*Trial Type	-0.62	0.36	-1.72	0.086	.

Table A8. Accuracy on test trials in Experiment 3 as a function of each participants' accuracy on exposure trials

`Correct` $\sim$ `Trial Type * Total Correct Exposure + offset(logit(`$^1/_{Referents}$`)) +`

`(Trial Type | subject)`

term	estimate	std.error	z.value	p.value	
Intercept	2.73	0.39	7.01	< .001	***
Total Exposure Correct	0.14	0.06	2.49	0.013	*
Trial Type	-1.39	0.39	-3.55	< .001	***
Total Exposure Correct*Trial Type	-0.26	0.06	-4.66	< .001	***

Table A9. Accuracy on test trials in Experiment 3 as a function of each participants' subjective reliability judgment

`Correct` $\sim$ `Trial Type * Subjective Reliability + offset(logit(`$^1/_{Referents}$`)) + (Trial Type | subject)`

term	estimate	std.error	z.value	p.value	
Intercept	4.54	0.44	10.33	< .001	***
Subjective Reliability	0.40	0.58	0.69	0.493	
Trial Type	-3.44	0.44	-7.81	< .001	***
Subjective Reliability*Trial Type	-1.63	0.59	-2.78	0.005	**

Table A10. Accuracy on test trials in Experiment 3 as a function of reliability condition and inspection time on exposure trials

`Correct` $\sim$ `(Trial Type + Reliability condition + Trial Type + Log(Inspection Time))`^{2} `+ offset(logit(`$^1/_{Referents}$`)) + (Trial Type | subject)`

term	estimate	std.error	z.value	p.value	
Intercept	3.11	0.20	15.94	< .001	***
Log(Inspection Time)	0.31	0.09	3.31	0.001	**
Trial Type	-2.75	0.20	-13.64	< .001	***
Reliability Condition	0.50	0.30	1.66	0.097	.
Log(Inspection Time)*Trial Type	0.03	0.09	0.34	0.736	
Log(Inspection Time)*Reliability Condition	-0.20	0.11	-1.83	0.067	.
Trial Type*Reliability Condition	-0.58	0.29	-1.97	0.048	*

D.1.4 Experiment 4

Table A11. Accuracy on test trials in Experiment 4 as a function of gaze and interval

`Correct` $\sim$ `(Trial Type + Gaze + Log(Interval))`^{2} `+ offset(logit(`$^1/_{Referents}$`)) +`
`(Trial Type | subject)`

term	estimate	std.error	z.value	p.value	
Intercept	3.37	0.16	21.32	$< .001$	***
Trial Type	-3.18	0.16	-19.93	$< .001$	***
Gaze Condition	-0.48	0.14	-3.52	$< .001$	***
Log(Interval)	-0.84	0.10	-8.59	$< .001$	***
Trial Type*Gaze Condition	0.90	0.14	6.63	$< .001$	***
Trial Type*Log(Interval)	0.80	0.09	8.71	$< .001$	***
Gaze Condition*Log(Interval)	-0.01	0.07	-0.10	0.917	

References

Attardo, S. (1997). Locutionary and perlocutionary cooperation: The perlocutionary cooperative principle. *Journal of Pragmatics*, *27*(6), 753–779.

Augustine, S. (1952). Treaties on various issues. Washington, DC: Catholic University of America Press.

Axia, G., & Baroni, M. R. (1985). Linguistic politeness at different age levels. *Child Development*, 918–927.

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, *20*(02), 395–418.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). Lme4: Linear mixed-effects models using eigen and s4. *R Package Version*, *1*(4).

Benitez, V. L., & Saffran, J. R. (2018). Predictable events enhance word learning in toddlers. *Current Biology*, *28*(17), 2787–2793.

Bloom, P. (2002). *How children learn the meaning of words.* The MIT Press.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Bonnefon, J.-F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition*, *112*(2), 249–258.

Boyer, A. (1702). *The english theophrastus: Or, the manners of the age: Being the modern characters of the court, the town, and the city...* W. Turner... R. Basset...; J. Chantry.

Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science*, *8*(6), 535–543.

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language*, *35*(01), 207–220.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

BÃijhler, K. (1934). *Sprachtheorie.* Oxford, England: Fischer.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, *72*(4), 1032–1053.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, i–174.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, *110*(28), 11278–11283.

Clark, E. V. (2009). *First language acquisition.* Cambridge University Press.

Clark, H. H., & Schunk, D. H. (1980). Polite responses to polite requests. *Cognition*, *8*(2), 111–143.

Cleveland, A., Schug, M., & Striano, T. (2007). Joint attention and object learning in 5-and 7-month-old infants. *Infant and Child Development*, *16*(3), 295–306.

Coenen, A., Nelson, J. D., & Gureckis, T. (2017). Asking the right questions about human inquiry.

Colombo, J. (2001). The development of visual attention in infancy. *Annual Review of Psychology*, *52*(1), 337–367.

Ervin-Tripp, S. M. (1967). An issei learns english. *Journal of Social Issues*, *23*(2), 78–90.

Ervin-Tripp, S. M. (1969). Sociolinguistics. *Advances in Experimental Social Psychology*, *4*, 91–165.

Franchak, J. M., Kretch, K. S., & Adolph, K. E. (2018). See and be seen: Infant–caregiver social looking during locomotor free play. *Developmental Science*, *21*(4), e12626.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*(1), 1–24.

Friesen, C. K., Ristic, J., & Kingstone, A. (2004). Attentional effects of counterpredictive gaze and arrow cues. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(2), 319.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

Goffman, E. (1967). *Interaction ritual: Essays on face-to-face interaction*. Aldine.

Goodman, N. D., & Frank, M. C. (2016a). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Goodman, N. D., & Frank, M. C. (2016b). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (Vol. 3, pp. 41–58). Academic Press.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.

Halliday, M. A. K. (1975). *Learning how to mean: Explorations in the development of language.* London: Edward Arnold.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431.

Hirschberg, J. B. (1985). *A theory of scalar implicature.* University of Pennsylvania.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., ... Bloom, L. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, i–135.

Holtgraves, T. (1997). YES, but... positive politeness in conversation arguments. *Journal of Language and Social Psychology*, *16*(2), 222–239.

Ide, S. (1989). Formal forms and discernment: Two neglected aspects of universals of linguistic politeness. *Multilingua-Journal of Cross-Cultural and Interlanguage Communication*, *8*(2-3), 223–248.

Jakobson, R. (1960). Linguistics and poetics. In *Style in language* (pp. 350–377). MA: MIT Press.

Kant, I. (1949). On a supposed right to lie from altruistic motives. *Critical of Practical Reason and Other Writings*, 346–350.

Kanwisher, N., Woods, R. P., Iacoboni, M., & Mazziotta, J. C. (1997). A locus in human extrastriate cortex for visual shape analysis. *Journal of Cognitive Neuroscience*, *9*(1), 133–142.

Koehne, J., & Crocker, M. W. (2014). The interplay of cross-situational word learning and sentence-level constraints. *Cognitive Science*.

Koenig, M. A., Clement, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science*, *15*(10), 694–698.

Lakoff, R. (1973). The logic of politeness; or, minding your p's and q's. In A. W. C. Corum T. Cedric

Smith-Stark (Ed.), *Papers from the ninth regional meeting of the chicago linguistics society* (pp. 292–305). Chicago: Department of Linguistics, University of Chicago.

Leech, G. (1983). *Principles of pragmatics.* London, New York: Longman Group Ltd.

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33. `http://doi.org/10.18637/jss.v069.i01`

MacDonald, K., Blonder, A., Marchman, V., Fernald, A., & Frank, M. C. (2017). An information-seeking account of eye movements during spoken and signed language comprehension. In *Proceedings of the 39th annual conference of the cognitive science society.*

MacDonald, K., LaMarr, T., Corina, D., Marchman, V. A., & Fernald, A. (2018). Real-time lexical comprehension in young children learning american sign language. *Developmental Science*, e12672.

MacDonald, K., Marchman, V., Fernald, A., & Frank, M. C. (2018). Adults and preschoolers seek visual information to support language comprehension in noisy environments. In *Proceedings of the 40th annual conference of the cognitive science society.*

MacDonald, K., Yurovsky, D., & Frank, M. C. (2017). Social cues modulate the representations underlying cross-situational learning. *Cognitive Psychology*, *94*, 67–84.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms.* Cambridge university press.

Manohar, S. G., & Husain, M. (2013). Attention as foraging for information and value. *Frontiers in Human Neuroscience*, *7*, 711.

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, *11*(3), F9–F16.

Matsumoto, Y. (1988). Reexamination of the universality of face: Politeness phenomena in japanese. *Journal of Pragmatics*, *12*(4), 403–426.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019.

Nelson, J. D. (2005). Finding useful questions: On bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4).

Oakes, L. M. (2011). *Infant perception and cognition: Recent advances, emerging theories, and future directions*. Oxford University Press, USA.

Oudeyer, P.-Y., & Smith, L. B. (2016). How evolution may work through curiosity-driven developmental process. *Topics in Cognitive Science*, *8*(2), 492–502.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*(4), 643.

Polite. (2017a). In *OED online*. Oxford University Press. Retrieved from `http://www.oed.com/view/Entry/146878?rskey=4vSu4F&result=1&isAdvanced=false`

Polite. (2017b). In *Cambridge online dictionary*. Cambridge University Press. Retrieved from `http://dictionary.cambridge.org/us/dictionary/english/polite`

Quine, W. V. (1960). 0. word and object. *111e MIT Press*.

Rohde, H., & Frank, M. C. (2014). Markers of topical discourse in child-directed speech. *Cognitive Science*, *38*(8), 1634–1661.

Ross-sheehy, S., Oakes, L. M., & Luck, S. J. (2003). The development of visual short-term memory capacity in infants. *Child Development*, *74*(6), 1807–1822.

Searle, J. (1975). Indirect speech acts. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (Vol. 3, pp. 59–82). Academic Press.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others the consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, *27*, 623–656.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1), 39–91.

Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480–498.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Smith, L. B., & Yu, C. (2013). Visual attention is not enough: Individual differences in statistical word-referent learning in infants. *Language Learning and Development*, *9*(1), 25–49.

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word–referent learning. *Trends in Cognitive Sciences*, *18*(5), 251–258.

Sweetser, E. (1987). The definition of lie. *Cultural Models in Language and Thought*, 43–66.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, *29*(6), 557–580.

Trueswell, J. C., Lin, Y., Armstrong, B., Cartmill, E. A., Goldin-Meadow, S., & Gleitman, L. R. (2016). Perceiving referential intent: Dynamics of reference in natural parent–child interactions. *Cognition*, *148*, 117–135.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*(3), 375–382.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*(2), 729–742.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Watts, R. J. (2003). *Politeness.* Cambridge University Press.

Watts, R. J., Ide, S., & Ehlich, K. (1992). Introduction. In *Politeness in language: Studies in its history, theory and practice (ed. by r.J. watts, s. ide, & k. ehlich)* (pp. 1–17). Mouton de Gruyter.

Woodard, K., Gleitman, L. R., & Trueswell, J. C. (2016). Two-and three-year-olds track a single meaning during word learning: Evidence for propose-but-verify. *Language Learning and Development*, *12*(3), 252–261.

Wu, R., & Kirkham, N. Z. (2010). No two cues are alike: Depth of learning during infancy is dependent on what orients attention. *Journal of Experimental Child Psychology*, *107*(2), 118–136.

Wu, R., Gopnik, A., Richardson, D. C., & Kirkham, N. Z. (2011). Infants learn about objects from statistics and people. *Developmental Psychology*, *47*(5), 1220.

Yoon, E. J., MacDonald, K., Asaba, M., Gweon, H., & Frank, M. C. (2018). Balancing informational and social goals in active learning. In *Proceedings of the 40th annual conference of the cognitive science society.*

Yoon, J. M., Johnson, M. H., & Csibra, G. (2008). Communication-induced memory biases in preverbal infants. *Proceedings of the National Academy of Sciences*, *105*(36), 13690–13695.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13), 2149–2165.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition.*

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition.*

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science*, *16*(6), 959–966.

Zangl, R., & Fernald, A. (2007). Increasing flexibility in children's online processing of grammatical and nonce determiners in fluent speech. *Language Learning and Development*, *3*(3), 199–231.