MECHANISMS OF CONVERSATION: AUDIENCE DESIGN AND MEMORY

BY

SI ON YOON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Associate Professor Sarah Brown-Schmidt, Chair
Professor Aaron S. Benjamin
Professor Gary S. Dell
Professor Cynthia L. Fisher
Associate Professor Duane G. Watson

Abstract

It is well established that conversational partners jointly establish brief labels for repeatedly mentioned entities. When speaking to a new partner who is unfamiliar with the labels, speakers use longer expressions to facilitate understanding. How this process of audience design scales up to conversations among three or more individuals is unknown. Further, while memory is thought to play an essential role in audience design, the link between memory for language and language use is not well explored. This dissertation consists of two parts that examine these two issues. In Chapter 2, I propose, and test, potential hypotheses regarding how speakers design referring expressions in multiparty conversation. The results of four experiments help to elucidate the mechanisms that support audience design in multiparty conversation. In Chapter 3, I explore the memory contributions to the referential phenomenon of lexical differentiation, aiming to understand the relationship between memory for discourse referents in dialogue and referential form. The results of three experiments provide insights into how memory for the discourse history guides language use during conversation. Taken together, these findings allow us to better understand the mechanisms of audience design and the interplay between language use and memory in conversation.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

It is well-established in studies of two-party conversation that interlocutors jointly establish brief labels when repeatedly referring to a series of entities (Brennan & Clark, 1996; Schober & Clark, 1989; Clark & Wilkes-Gibbs, 1986; Isaacs & Clark, 1987; Krauss & Weinheimer, 1964, 1966; Wilkes-Gibbs & Clark, 1992; Van der Wege, 2009; Yoon & Brown-Schmidt, 2014). This process, sometimes referred to as *lexical entrainment* (Brennan & Clark, 1996), is thought to reflect the establishment of partner-specific, joint knowledge for these labels.

Speakers are sensitive to their partner's knowledge state and accommodate their utterances accordingly. This process of designing utterances to meet the communicative needs of one's addressee is a process generally known as *audience design* (Clark, 1996, also see Sacks, Schegloff, & Jefferson, 1974). For example, in a situation when conversational partners establish joint labels for a series of entities, and then one of the partners later speaks to a new partner, the original partner tends to produce longer, more elaborated expressions when speaking to the new partner who lacks knowledge of the lexically entrained terms (Wilkes-Gibbs & Clark, 1992; Horton & Spieler, 2007; also see Horton & Gerrig, 2002; 2005; Heller, et al., 2012). Such findings are consistent with a broader set of results that speakers tailor expressions based on the knowledge state of the current addressee (Isaacs & Clark, 1987; Brennan & Clark, 1996; Bromme, Jucks, & Wagner, 2005), rather than their own egocentric knowledge (cf. Horton & Keysar, 1996). One question that has been relatively less explored is how audience design in two-party conversation scales up to multiparty conversation—conversation among three or more individuals. This is the question that I will address in Chapter 2.

The second question is how interlocutors formulate memory representations to support successful audience design during conversation. Memory for what is and is not jointly known is necessary in order for speakers to tailor their utterances based on the knowledge and perspective of the addressee. Successful audience design may depend on the degree to which these representations are accessible in memory (Horton & Gerrig, 2005). If a speaker fails to remember what they have said in the past, they might repeat this information to the same addressee multiple times, thus making conversation inefficient and redundant. In spite of obvious ties between memory and language use, little research has examined the link between memory for language and language use. Thus, in Chapter 3, I present the findings of three experiments that examine the link between memory for discourse referents in dialogue, and referential form.

The current series of experiments form two streams of research that aim to address the two questions mentioned above. In Chapter 2 (Experiments 1-4), I test potential hypotheses regarding how speakers design referring expressions in multiparty conversation. In Chapter 3 (Experiments 5-7), I explore the nature of the memory representations that support language use in dialogue and examine potential asymmetries between speakers and listeners in the memorial representation of the discourse record. In what follows, I review existing research on multiparty conversation and raise questions about whether speakers maintain distinct representations of common ground for each individual, or whether these representations are blended during multiparty conversation. I then review the literature in the language and memory traditions, to set the stage for exploration of the link between memory for language and language use.

Multiparty Conversation

The literature on two-party conversation demonstrates that speakers adjust language use based on the knowledge of their conversational partner. However, it is largely unknown how these processes scale up to communication in multiparty conversation. Multiparty conversation, in which more than two individuals converse, presents complexities for the speaker due to the fact that the joint knowledge held between any two pairs of individuals within the larger group necessarily differs. After all, each individual brings to a conversation their own perspective and beliefs. As a result, the pairings of individuals within the larger group differ in the knowledge that is mutually known[1]. In such situations, a speaker may address multiple addressees at once, or may direct utterances to only one of several potential addressees. Whether speakers can maintain separate representations of joint knowledge for multiple potential addressees in such situations is relatively unexplored. Likewise, little is known about the ways in which speakers design utterances for multiple individuals. Multiparty conversation is an important test case because multiparty communication is common in day-to-day life. Moreover, the ways in which speakers approach the problem of audience design in multiparty conversation can provide insights into the underlying mechanisms of audience design, including those mechanisms at work in two-party conversation.

In one study that examined audience design in three-party conversation, a speaker simultaneously designed utterances for two addressees who shared distinct degrees of common ground with the speaker (Yoon & Brown-Schmidt, 2014). In the first phase of

---

[1] Critically, mutual knowledge differs from knowledge that two individuals happen to share in that they know each other knows (see Clark & Marshall, 1978; 1981). In the current experiments, the relevant mutual, or joint knowledge is the knowledge of the object labels that is acquired during the task-based conversation.

the task, the speaker first established brief labels for a series of abstract images with one partner while the second partner waited in another room. In the second phase, the speaker referred to the same images again, addressing either the same (knowledgeable) partner alone, or addressing both the knowledgeable partner and the naïve partner at the same time. During this second phase, speakers continued to use the brief labels when speaking to the knowledgeable partner. By contrast, when simultaneously addressing the knowledgeable and the naïve partners, speakers produced longer descriptions. These findings show that speakers are sensitive to the presence of a naïve addressee in three-party conversation. What these findings do not specify, however, is *how* speakers approach the task of audience design in multiparty settings. This is the focus of the first part of the present research (Chapter 2: Experiments 1-4).

Distinctiveness and Flexibility

In order to tailor language use to the knowledge of a particular individual in a multiparty setting, the speaker must maintain distinct representations of the joint knowledge shared with each partner. Moreover, speakers must be able to draw on these representations in a partner-specific manner, such that utterances designed for one partner are based on the joint knowledge held with that person, and not influenced by joint knowledge held with others. The open question, then, is whether speakers are capable of maintaining and using distinct, partner-specific representations in such a way.

Initial evidence from studies in which a speaker address a naïve partner (Wilkes-Gibbs & Clark, 1992; Yoon & Brown-Schmidt, 2014), or discuss a new piece of information with a familiar partner (Heller, et al., 2012; Horton & Gerrig, 2002; 2005) show that speakers can detect one addressee's lack of knowledge, and adjust accordingly.

Less clear, however, is whether speakers can maintain distinct representations of the task-relevant joint knowledge for different individuals in the same setting, particularly when neither partner is naïve.

Evidence consistent with the idea that speakers may face difficulty in maintaining distinct representations comes from studies in which speakers share some knowledge with one partner and different knowledge with another partner. Horton and Gerrig (2005) examined dialogues in which a speaker named a series of images with partner A, and then a different series of images with partner B. Speakers then described all the images for partner A alone. When describing images that were new to partner A, speakers provided extra details, demonstrating audience design. However, speakers were less successful when the images shared with the two partners were from the same category (e.g., frogs vs. different frogs) than when the images associated with the two partners were from distinct categories (e.g., fish vs. frogs). These findings highlight the role that memory processes play in constraining the process of audience design: If the appropriate partner-specific knowledge is difficult to access, audience design may suffer.

Thus, in a more taxing context, such as when a speaker must alternate addressing one of two co-present partners, each of whom holds partial knowledge of the topic at hand, appropriately designing utterances for each partner may prove too difficult. While in one study, young adults were able to alternate designing utterances for a familiar partner, and a new, naive partner (Horton & Spieler, 2007), that study used a non-interactive test where partners were represented by photographs. It is not clear if this ability to switch between partner representations would extend to live conversation. A contributing factor is that it may take time and effort to access the relevant common ground representations associated with the current addressee. Switching costs are well-established in a variety of cognitive domains (Rogers & Monsell, 1995), including spatial perspective-taking (Ryskin, et al., 2014). Moreover, recall of items in memory may often be clustered (Herrmann & Pearle, 1981) such that participants may prefer to stick with one type of recall for stretches at a time, rather than switching (see Young, 2004). Given these findings, having to switch between accessing the relevant memorial representations of common ground held with the multiple different matchers may be too taxing for audience design to proceed appropriately.

There are two potential ways in which speakers might maintain representations of joint knowledge in cases where each of two partners has some knowledge (Figure 1.1). For example, consider a situation in which a speaker discusses an unusual image once with the first partner, but discusses the same image four times with a second partner, establishing a lexically entrained term (e.g., *the 3d chair*). The first partner would have a low degree of common ground (CG) for the entrained term, whereas the second partner would have high-CG for the same term. If speakers maintain distinct representations of joint knowledge, speakers should describe the image in more detail for the partner with low-CG, whereas they should produce the short entrained term with the high-CG partner. Alternatively, speakers may have difficulty maintaining distinct representations of the two partners. If so, non-distinct representations of common ground would result in expressions that are not tailored for either individual, and possibly reflect the overall knowledge of both partners.

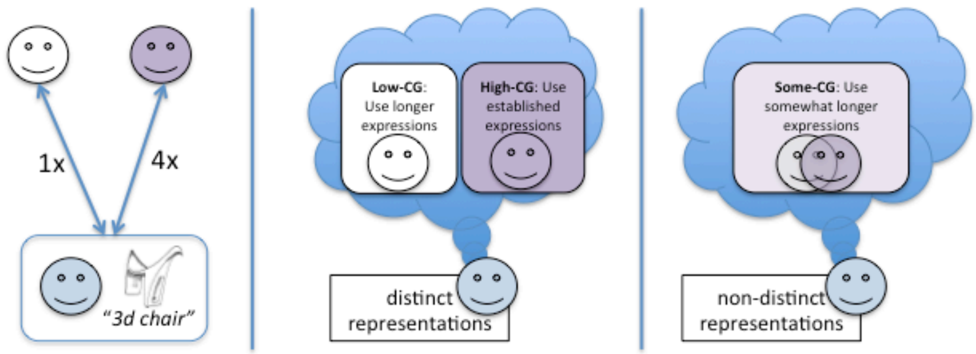


Figure 1.1. Left panel: In multiparty conversation, the speaker (blue face) jointly discusses a novel image 4x with one partner (purple face), but only 1x with another partner (white face). Center panel: The speaker forms distinct representations of the joint knowledge held with each partner, and designs utterances accordingly. Right panel: The speaker forms non-distinct representations for the set of addressees.

Moreover, even if distinct representations are maintained, it is unclear whether speakers can flexibly alternate between these representations, as in the case where a speaker alternates addressing one partner and then the other partner in a multiparty setting. According to one proposal (Horton, 2007), people serve as contextual memory cues for retrieval of information that is associated with them, even in situations that are not overtly communicative. If so, in cases where multiple partners are present in the context, it may be difficult to access the representations of joint knowledge associated with the intended addressee, as information associated with the non-intended addressee may become automatically activated as well.

Audience design in multiparty conversation

How do speakers design utterances for more than one individual? Three potential approaches that speakers can pursue in multiparty conversation were examined for this question in Experiment 2. One approach to audience design in these settings is for speakers to Aim low (Figure 1.2., top left), such that in situations with multiple addressees, the speaker designs utterances to match the knowledge state of the least knowledgeable person. Aiming low would have the benefit of ensuring that all addressees would understand what was said, and would be generally consistent with findings that speakers design longer expressions when speaking to a naïve individual in dialogue (Wilkes-Gibbs & Clark, 1992; Horton & Spieler, 2007).

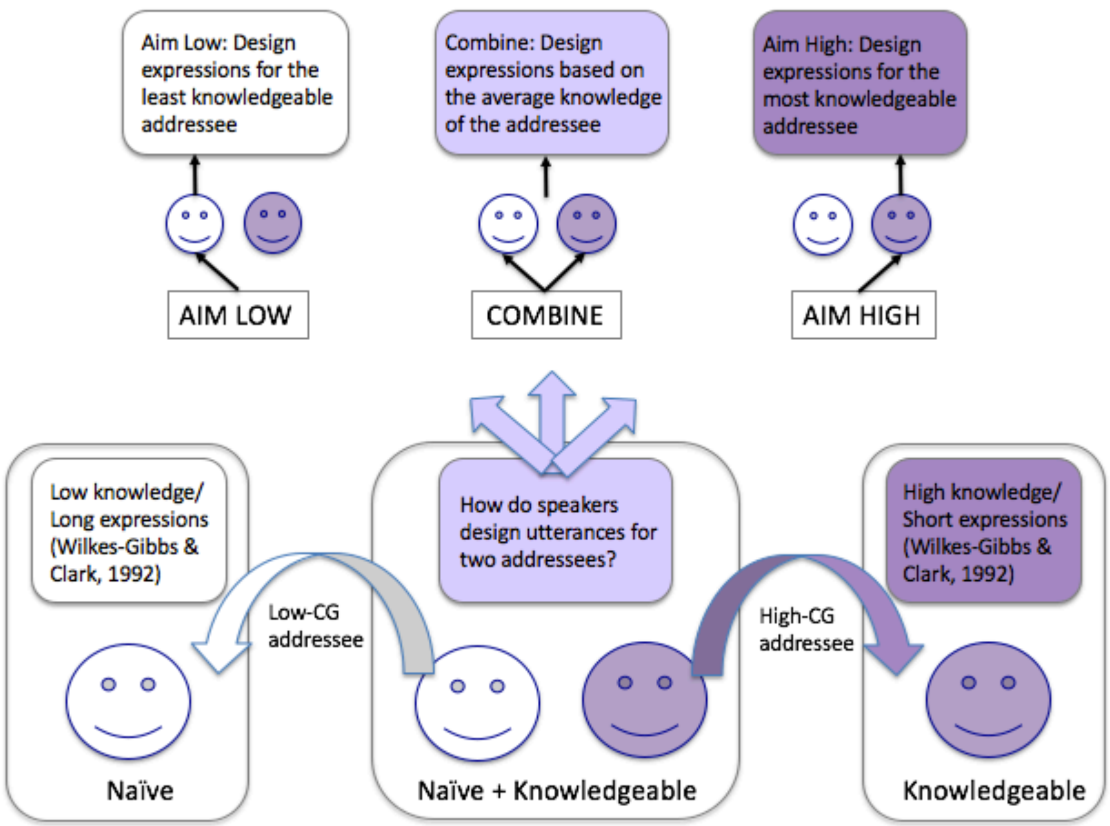


Figure 1.2. Candidate approaches to audience design in multiparty conversation.

Alternatively, speakers may <u>Combine</u> knowledge state of the addressees (Figure 1.2., top middle), and plan utterances with respect to a combination of the knowledge states. This approach is similar to the proposal that when engaging in audience design, speakers bring to mind a generic addressee (see Dell & Brown, 1991). The advantage of an combining strategy would be that it may balance the needs of the two addressees, such that the less-knowledgeable addressee would be able to follow along (with some effort), without distracting the more knowledgeable addressee with unnecessary details (see Engelhardt, Bailey, & Ferreira, 2006). Note that there are multiple possibilities for how knowledge states might be combined; we return to this issue in the general discussion.

A final possibility is that speakers Aim high (Figure 1.2., top right), and design utterances for the most knowledgeable addressee. One reason to think that speakers may adopt this approach is evidence that while speakers do engage in audience design, they also overattribute knowledge to the addressee when the speaker themselves is more knowledgeable (Fussell & Krauss, 1991; 1992).

So far, I have proposed three candidate approaches for how speakers design utterances in multiparty conversation. However, it is less clear whether speakers always pursue one approach or flexibly switch strategies based on contextual pressures such as the number of addressees in a group or their knowledge states. In multiparty conversation where each addressee holds a different amount of common ground, addressees may have distinct needs depending on their background knowledge (knowledgeable vs. naïve). Successful communication with knowledgeable addressees might require conversational efficiency—a brief, established label (e.g., *the 3d chair*) might be the quickest way to achieve reference with a partner who shares common ground for the object labels. In contrast, successful communication with a naïve addressee might demand an emphasis on conversational informativity –the use of long and descriptive labels that provide all of the critical information needed to understand a reference to an object label for the first time (e.g., *it looks like a 3-dimensional chair… with long legs…*). If a speaker were to emphasize informativity, this may take extra time and effort, but the speaker could be assured that every member of the conversation would be able to understand the intended message. For instance, imagine a situation where two research assistants working in the same laboratory talk about coding data. In such a situation they would not need to discuss the basics of the coding process, since they both know this information already. In this case, they are likely to skip discussion of this background information, and focus on the new details. This is an example where conversational partners use their common ground to have an efficient conversation. In contrast, imagine a situation where a new research assistant joins the lab. In this situation the existing (knowledgeable) RA should give detailed instructions to the new person when talking about the basic aspects of the data coding process. This is an example where the speaker (the knowledgeable RA) emphasizes informativity at the expense of efficiency.

In multiparty conversation, speakers must strike a balance between two distinct communicative pressures: informativity and efficiency. When speaking to a group of addressees who vary in their background knowledge, emphasizing only informativity or only efficiency may cause some of the addressees to become bored or alternatively, confused. As a result, speakers in multiparty conversation may need to estimate and combine the addressees' background knowledge to design their utterances with respect to the demands of multiple individuals. How speakers deal with addressees who have conflicting needs is not well understood. Thus, in Experiments 3 and 4, I examine how speakers cope with the needs of different addressees in multiparty conversation, whether they continue to use one approach or flexibly switch to different approaches depending on the conversational context.

Memory for the discourse history

Multiple types of memorial representations are likely to be relevant to the audience design process, including memory for the content of what was said, or *item memory* (e.g. the fact that "*I got the job"* was said), as well as memory for *contextual* information. The relevant context for a conversation likely includes information about whom one's conversational partner is (see Brown-Schmidt, Yoon, & Ryskin, 2015), the relevant information that was discussed in the past (sometimes called the *historical discourse context*, Brennan & Clark, 1996), as well as information in the immediate context. While it is broadly understood that successful audience design depends on access to the relevant memory representations (e.g., Horton & Gerrig, 2005a, b; Knutsen & Le Bigot, 2012), the link between memory for language and language use has been less explored.

Research in the language tradition

Traditional approaches to language processing separately examine the processes of language production and language comprehension (e.g., Levelt, 1993; Townsend & Bever, 2001). Motivation for the separate investigation of production and comprehension processes include claims that language production and comprehension are modular (Fodor, 1983) such that the two processing systems are independent and do not themselves interact. A traditional neurolinguistic model, such as the Lichtheim-Broca-Wernicke model, also supported distinct processing associated with production and comprehension, based on the idea that production and comprehension deficits in aphasia were linked to damage to different brain areas (Lichtheim, 1885; Ben Shalom & Poeppel, 2008).

More recently, the idea that language production and comprehension are interwoven has gained interest, as have methodologies that support the simultaneous investigation of language production and comprehension in conversation (e.g., Trueswell & Tanenhaus, 2005; Richardson & Dale, 2005; Richardson, Dale, & Kirkham, 2007; Award, Warren, Scott, Turkheimer, & Wise, 2007). For example, Pickering and Garrod (2004; 2013) proposed that production and comprehension are tightly interwoven actions that draw on the same core representations at all levels of language processing including phonology, semantics, and syntax. According to their model, production and comprehension are not two separate systems, but inextricably linked together. Evidence in support of this claim includes findings that interlocutors are highly likely to take turns without delays during a daily conversation. If production and comprehension involved two separate systems, it would be difficult to maintain a stream of conversation without temporal delays in turn-taking. Recent neurobiological data supports this claim, showing that brain areas activated during production and comprehension are highly correlated (Silbert et al. 2014).

However, other work suggests that production and comprehension draw on distinct representations, for example, of the discourse context. Yoon and Brown-Schmidt (2013) examined the use of discourse context by speakers and listeners during a natural conversation. We measured a phenomenon termed *lexical differentiation,* where speakers differentiate two sequentially presented objects from the same category (Van der Wege, 2009). For example, if a speaker described one shirt in the context of several unrelated objects, she might refer to it as "*the shirt*". However, if she later referred to a second, distinct shirt in the context of unrelated objects, she might differentiate the second shirt from the first by using a modifier, as in "*the striped shirt*", even though the modifier is not necessary in the local context. According to Van Der Wege (2009), speakers lexically differentiate in order to avoid giving the same label to two different entities, a process she terms "pre-emption by similar form" (also see Clark & Clark 1979). While speakers tended to differentiate, listeners did not appear to learn this tendency; there was no evidence that listeners expected speakers to differentiate the two objects. One interpretation of these findings is that listeners are more sensitive to the immediate local context, whereas speakers are more sensitive to the historical discourse context (Brennan & Clark, 1996). However, the speakers' overall differentiation rate was not high (~7.5%), suggesting that in many cases neither interlocutor remembers the previous referent, thus reducing the lexical differentiation rate.

Research in the memory tradition

A well-known finding in the memory literature is that speaking, writing or otherwise producing items promotes better item memory compared to hearing or reading those items, a phenomenon known as *the generation effect* (Slamecka & Graf, 1978; also see MacLeod, et al., 2010). The meta-analytic effect size of the generation effect is large (d = .40; Bertsch, Pesta, Wiscott, & McDaniel, 2007), and observed in a variety of testing paradigms, including recognition, cued recall and free recall (Donaldson & Bass, 1980; Graf, 1980; Graf, 1982; Slamecka & Graf, 1978; Johnson et al., 1981). The effect obtains across a variety of different experimental paradigms, including various encoding rules (e.g., associate, rhyme, category, opposite, synonym), diverse materials (single word, word pair, sentences, and even pictures and arithmetic problems), different presentation rates (timed or self-paced), in both intentional and incidental learning paradigms, and in both between and within subjects designs (Slamecka & Graf, 1978; McNamara & Healy, 2000; deWinstanley & Bjork, 2004; McDaniel, Waddill, & Einstein, 1988).

While it is clear that generation promotes item memory, less clear is how it affects memory for contextual information, including memory for one's conversational partner. For example, Gopie and MacLeod (2009, Experiment 1), examined memory in a paradigm in which 60 different facts were paired with 60 different faces. In one condition participants saw a face and then read a sentence silently, imagining that the person in the photo told them the information (similar to playing the role of listener). In the other condition, participants read a sentence silently and then saw a face, and then said aloud the sentence to that face (similar to playing the role of speaker). Whereas memory for items (the sentences) was equivalent across these two conditions, memory for contexts (the faces) and for the pairing of item and context (face-fact pairings) was significantly better when participants received information than when they told that information to someone else (also see Jurica & Shimamura, 1999). One explanation is that when producing information, a speaker puts more focus on the to-be-communicated information, whereas when receiving information attention is more likely to spread to the context (see Koriat et al., 1991). Note, however, that in other paradigms in which context is operationalized in different ways (e.g., the color a word is presented in), generation does sometimes improve context memory (Marsh, Edelman, & Bower, 2001; Marsh & Hicks, 2002).

How might these findings extend to true conversational settings? In free-recall of unscripted conversation, participants tend to report more of what was *heard* than what was *said* (Stafford et al., 1987; Stafford & Daly, 1984; cf. Ross & Sicoly, 1979), though this may be due to an egocentric bias to recall what seemed new to you (what someone said to you). Indeed, in an analysis of what information tends to get repeated over the course of a conversation, Knutsen and Le Bigot (2014) report that referring expressions like "*the market"* are more likely to be repeated in a conversation by the person who first introduced that topic into the conversation, consistent with a generation effect. McKinley, Brown-Schmidt, and Benjamin (2015) similarly report a generation benefit for item recognition in a natural conversation paradigm where the "items" were pictures that participants discussed with one of two conversational partners. This generation benefit for item recognition suggests that the tendency to report previously heard information in free-recall (Stafford et al., 1987; Stafford & Daly, 1984) may reflect a response bias. Further, inconsistent with the findings by Gopie and MacLeod (2009), McKinley et al. (2015) found no difference in context recognition between speakers and listeners, where context was defined as the partner with whom a picture was studied (though better item memory did boost performance overall for speakers). Taken together, these findings suggest that during the course of a natural conversation, speakers and listeners may arrive at different memory representations of the discourse history.

CHAPTER 2: MULTIPARTY CONVERSATION

In Experiments 1-4, I address a set of questions regarding the processes by which speakers design utterances in multiparty conversation. Across four experiments, I manipulate the identity of addressee or addressees and their knowledge states, and examine how speakers design utterances with respect to the intended addressees' combined knowledge states. The first line of inquiry concerns the issue of whether speakers can maintain and flexibly use distinct representations of common ground shared with different individuals. To answer this question, in Experiment 1, speakers alternate between addressing a knowledgeable and a naïve partner, and I evaluate whether the form of the speaker's expressions reflect the current addressee's knowledge state. A positive answer to that question sets the stage for the second line of inquiry, which examines the way in which speakers design utterances in multiparty conversation. In Experiment 2, I test three potential hypotheses regarding how speakers produce referential expressions in three-party conversation. In Experiments 3 and 4, I create more taxing situations to examine whether speakers always pursue the same audience design strategy, or whether they develop different strategies depending on the situation. In Experiment 3, I create different combinations of knowledge states among the addressees in 4- and 5-party conversations in order to examine how the audience design processes scales up to even larger groups. In Experiment 4, I manipulate the local visual context in order to examine how speakers take into account both the immediate context and the addressees' background knowledge at the same time. The results of Experiments 1-4 will contribute to our understanding of the mechanisms of audience design in multiparty conversation.

Experiment 1

Experiment 1 examined how Directors design referring expressions in a conversation with two Matchers who hold different degrees of common ground for a set of abstract image labels. To establish common ground for the labels, in a series of Entrainment trials, the Director completed a referential communication task with one Matcher once, and the other Matcher four times. To test for distinct representations of common ground, all three partners then completed a series of test trials together where the Director uniquely identified the images for one Matcher at a time.

Method

Participants

Sixty undergraduates at the University of Illinois at Urbana-Champaign participated in the experiment in return for either partial course credit or cash payment ($12). Three participants took part in the study at the same time and were randomly assigned to the roles of Director, Matcher 1, and Matcher 2. All participants were native speakers of North American English, and had no reported hearing or uncorrected visual impairment.

Materials and Procedure

The experiment was conducted in two phases: Entrainment and Test. During the entrainment phase, pairs of participants (Director with the first Matcher; then the Director with the second Matcher) played a sorting game either 4 times (high-CG condition) or 1 time (low-CG condition), entraining names for game-pieces in dyads. Following entrainment, all three participants sat in the same room to complete the test trials. The two phases of the experiment, entrainment and testing, were completed four times with four separate sets of stimuli.

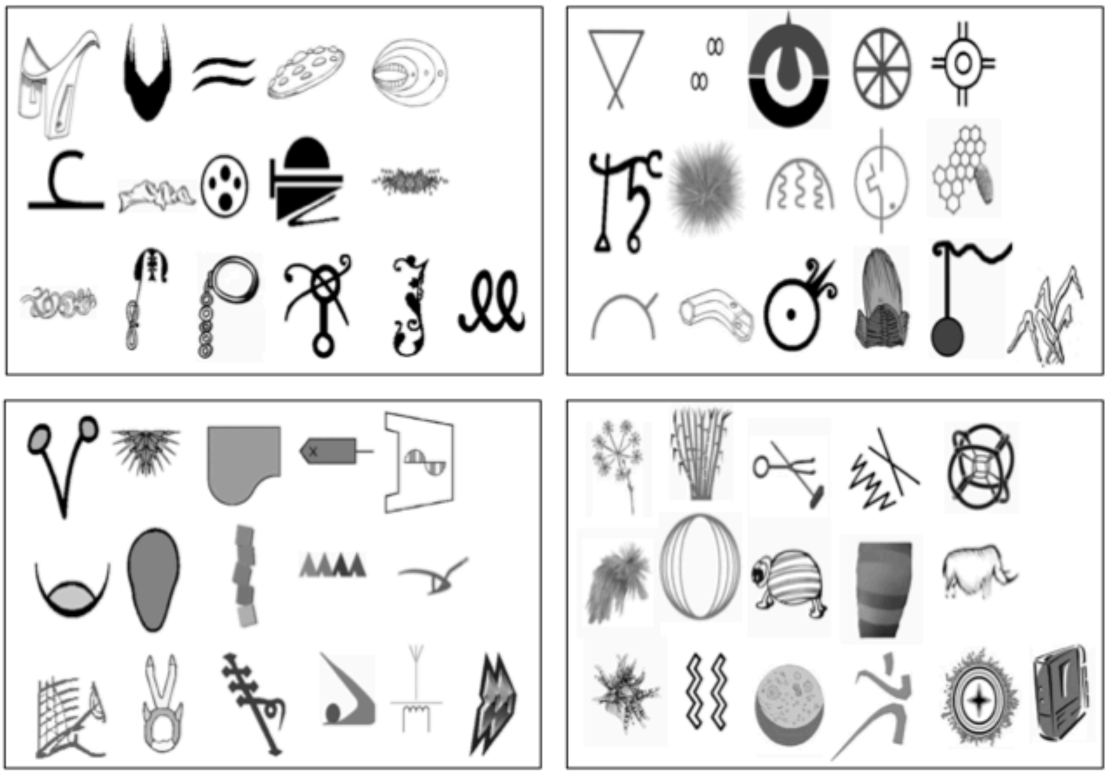


Figure 2.1. Example stimuli during the first (Entrainment) phase of the task in Experiments 1, 2, and 3. The four different sets of figures were used once each across the four rounds of play.

In the first (entrainment) phase of the task, the Director and one of Matchers played a version of a referential communication task (Krauss & Weinheimer, 1966; Wilkes-Gibbs & Clark, 1992) while sitting at separate computers in the same room. Each partner viewed a computer screen with 16 abstract images, arranged in a different order on each screen (Figure 2.1). Across the entire experiment, there were a total of four sets of 16 images (64 total), which were adapted from previous experiments (Arnold, Hudson Kam, & Tanenhaus, 2007; Brown-Schmidt, 2009a), and similar pictures.

During each entrainment trial, the 16 images were in set locations on the Director's screen, whereas the Matcher was able to use the mouse to move the images.

The Director's task was to instruct the Matcher on how to sort the images into the same order as the Director's images. Typically, the dialogue involved the Director describing the 16 images, one at a time, to the Matcher, who re-arranged them into a new order. The Director and Matcher repeated the same task either 4 times (high-CG condition) or 1 time (low-CG condition) to establish either a high or a low degree of common ground for descriptions of these images. Both voices were recorded directly to disk. Participants were allowed to freely describe the images and there was no restriction on what they could say to complete the task. The task was interactive and participants were encouraged to ask questions as needed. While the two participants played the sorting game, the other participant performed a series of unrelated filler tasks (e.g., verbal and math tasks) in a separate room behind a closed door. These filler tasks were not scored; they were simply used to keep the third person occupied while the other two participants completed the sorting task. When the Director and the first Matcher were done sorting, two Matchers switched places, and the Director proceeded to sort the same 16 pictures with the second Matcher either 1 time (low-CG condition) or 4 times (high-CG condition).

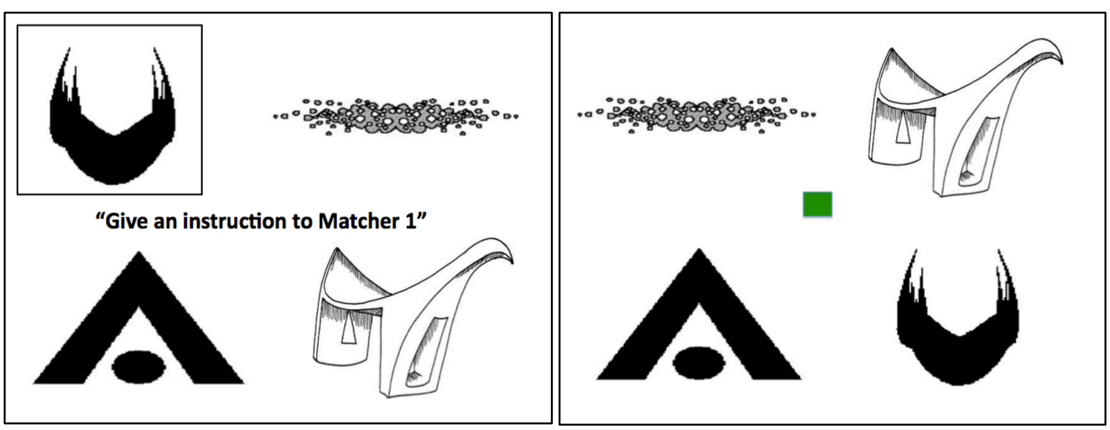


Figure 2.2. Example test display for Director (left) and addressee-Matcher (right) in Experiment 1. The Matcher who was not the addressee saw the same image as the addressee-Matcher, except his or her screen had a red square. Test displays always consisted of one new image (e.g., the bottom left one) and three old images.

In the second phase of the task, all three participants completed the test session together in the same room (Figure 2.2). Each participant sat at a separate computer that showed three of the images that had previously been discussed during the entrainment phase, and one new image that had never been seen before by any of the three participants. The participants completed 28 test trials, including 16 target trials and 12 filler trials. Target trials referred to the old images, whereas filler trials referred to the new images. Across the four different sets, stimuli on test trials included these 64 critical images as well as 112 additional images that served as novel fillers.

On the Matchers' screens, each test trial began with a fixation cross for 1000ms, then the cross disappeared and 4 images appeared (Figure 2.3). The Director's screen showed the same four images in a different arrangement, one of which (the target) was indicated to the Director with a black box. At the center of the Director's screen was a sentence that instructed the Director on whom they would be directing on that trial (either Matcher 1 or Matcher 2, e.g., "Give an instruction to Matcher 1"). The Director was prompted to tell either Matcher 1 or Matcher 2 (randomly alternating) to click on one of the 4 images. At the center of the Matchers' screens, there was a square that indicated whether he or she was the current addressee or not. If the square was green, he or she was the current addressee, and was to follow the Director's instructions to click the target. If the square was red, he or she was not the current addressee, and was instead asked to press enter after the other Matcher clicked the target. The order of trials was randomized, but overall Matcher 1 and Matcher 2 were each the intended addressee half of the time. The Director was allowed to describe the target in any way they saw fit, with the exception of locative phrases, such as "*the top left one*"; Matchers were free to converse as they saw fit. The trial ended when the intended addressee clicked the target and the other Matcher hit enter to advance (Figure 2.3). The location of the 4 images on the Matchers' and Director's screen was random.

Matcher 2' eye movements were monitored during the test session with an Eyelink 1000 desktop mounted eye tracker, however these data are not the focus of the present investigation so we do not discuss them further[2].

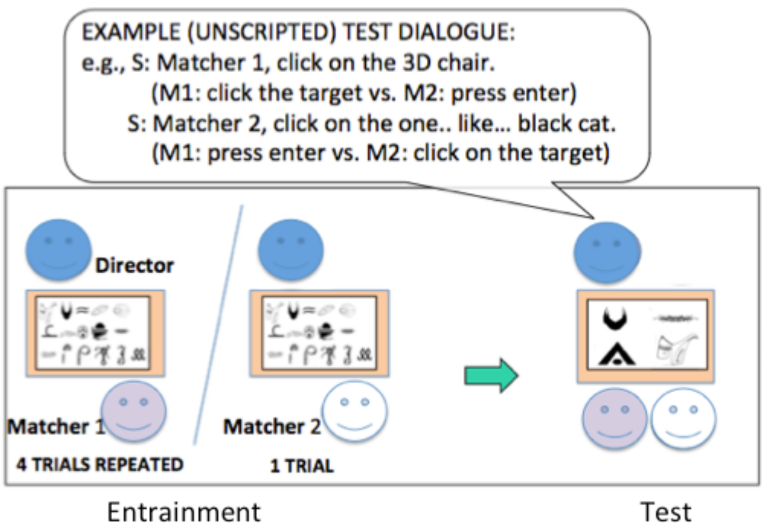


Figure 2.3. Experiment 1: Example procedure and hypothesized utterances at test. Matcher 1 is in the non-recent/ high CG condition; Matcher 2 is in the recent/ low CG condition.

The two phases of the experiment (Phase 1: sorting; Phase 2: testing) were repeated four times for each triplet of participants (Round 1: Phase 1-sort, Phase 2-test; Round 2: Phase1-sort, Phase 2-test, … Round 4: Phase 1- sort, Phase 2- test). Four different sets of images were used for the four different rounds of play. The order of the stimulus sets across blocks was counterbalanced across groups of participants. The

---

2 Eye movements were originally recorded in order to examine a supplemental question-- whether listeners could adjust their processing of the Director's expression based on the perspective of the other Matcher in cases where the eye-tracked Matcher was not the addressee. However, the pattern of eye movements when the eye-tracked Matcher was not the addressee indicated that in such circumstances they were largely not paying attention to the instructions. As a result, we do not focus on the eye-gaze data in the present analyses.

Director, Matcher 1, and Matcher 2 maintained the same roles throughout the experiment (e.g., Matcher 1 never became Matcher 2, etc.). The entire experiment took approximately 90 minutes.

Across the four rounds of game play, we counterbalanced whether Matcher 1 or Matcher 2 held a high degree of common ground (high and low CG) with the Director, and whether Matcher 1 or Matcher 2 completed the entrainment trials first. I refer to the condition in which the Director and the first Matcher established a high degree of common ground and the second Matcher a low degree of common ground as the "High CG/ Low CG" condition. The condition in which the first Matcher had a low degree of common ground, and the second a high degree of common ground was labeled the "Low CG/ High CG" condition. The data analysis focused on the length of the Director's referential descriptions as a function of the degree of common ground held between the Director and the current Matcher.

I also included the recency of the experience that established common ground (recent and non-recent) as a control variable in the statistical analysis. This control variable was necessary because entrainment trials with Matcher 1 and Matcher 2 were completed in sequence, and was possible that very recently established common ground would be less sensitive to the degree of common ground manipulation. At test, when the Director spoke to the Matcher that went second during the entrainment phase, common ground was coded as "recent", and "non-recent" when the addressee at test had gone first in the entrainment phase.

Predictions

Previous research has established that in referential communication tasks like the one examined here, partners develop brief, lexically-entrained terms as they repeatedly describe the same images (Schober & Clark, 1989; Clark & Wilkes-Gibbs, 1986; Isaacs & Clark, 1987; Krauss & Weinheimer, 1964, 1966; Wilkes-Gibbs & Clark, 1992; Van der Wege, 2009). Moreover, these terms are partner-specific such that when speaking with a new, naive partner who is not familiar with the way the images had been described, speakers use longer, more elaborated expressions (Wilkes-Gibbs & Clark, 1992; Holler & Wilkin, 2009). Based on these findings, during the entrainment trials, I expected Directors to use shorter expressions over successive entrainment trials with the same Matcher, but longer expressions when switching to the second, naive Matcher. Such a finding would set the stage for our investigation of the flexibility of the use of common ground representations in test trials.

During test trials, if Directors can simultaneously maintain and flexibly alternate between distinct representations of the knowledge states of the two co-present Matchers, they should use short expressions when addressing the Matcher who has higher common ground. By contrast, Directors should be more likely to use longer expressions and produce more disfluencies when addressing the addressee with less knowledge. However, if Directors are unable to maintain distinct representations, or unable to flexibly switch back and forth between these representations, Directors may produce similar expressions for both Matchers, regardless of the degree of common ground they share.

Results

The referential descriptions produced by the Directors during both the entrainment and test trials were transcribed. The primary analyses focus on the length of the Director's referential expressions during entrainment and test. Following Yoon & Brown-Schmidt (2014), I coded whether the referential expression at test was fluent as a supplemental measure, based on the assumption that disfluency reflects speakers' planning difficulty (Clark & Wasow, 1998; Ferreira, 1991). I analyzed the data using mixed-effects models; full model details are presented in Appendix A.

Entrainment Trials

The length of referring expressions during entrainment trials was calculated in terms of the number of all words used to describe the image. Word counts included all function (e.g., "the") and content words (e.g., "circle") in the referring expression, but not disfluencies, such as ("uh" and "um"). The average number of words used to describe each object for each trial is presented in Table 1. Across repeated trials with the same partner and stimuli, the length of each referring expression decreased gradually (Figure 2.4), consistent with the formation of common ground. However, when switching to the second Matcher, Directors used longer expressions, demonstrating sensitivity to the new partner's lack of knowledge. A typical transcript is illustrated below:

> <u>Example transcript for the description of one of the 16 objects across rounds in the High CG/ Low CG condition</u>:
> 1st: *This is a weird thing with a triangle hole on the left and like a rectangular hole on the right. Like pointy at the end… looks like a swan 3D chair maybe*?
> 2nd: *It is that weird, swan-shaped thing with the triangle and other thing. Saddle? 3D chair*?
> 3rd: *the weird swan 3D chair shape*.
> 4th: *the swan 3D chair*.
> (Partner switches from Matcher 1 to Matcher 2)
> 5th: *It looks like a chair…. Has a triangle cut out of the bottom of it. It's three dimensions, has a pointed end at the top, looks like a swan 3D chair*.

The analysis of expression length during entrainment included condition (High/Low CG and Low/High CG) and trial order (1-5) as fixed effects. A significant effect of trial order was due to a gradual decrease in the number of words used to describe each item across trials (z=-13.004, p<.05). An interaction between trial order and condition (z= 7.135, p<.05) was explored in planned comparisons between the conditions at each trial separately (Table 2.1)[3]. These comparisons revealed a significant effect of condition starting at trial 2 such that Directors produced significantly longer expressions when they switched to the second Matcher (the Low CG/ High CG condition) compared to when Directors continued to address the same partner (the High CG/ Low CG condition) (see Figure 2.4). This partner-switch effect persisted during trials 3 and 4 as well, with Directors producing significantly longer expressions in the Low CG/ High CG condition. Finally, at trial 5, in the High CG/ Low CG condition, the Director began addressing the naïve second Matcher. This switch is reflected in the Director's use of significantly longer expressions in the High CG/ Low CG condition, compared to the Low CG/ High CG condition.

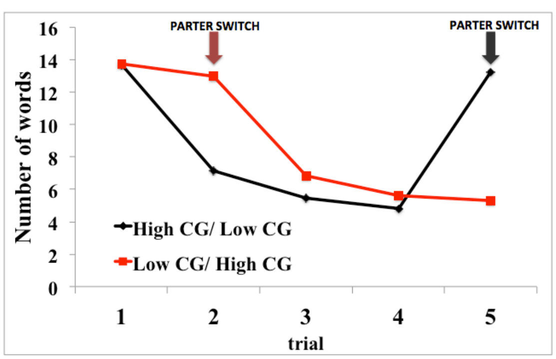


Figure 2.4. Experiment 1: The average number of words used by the Director to describe each object during the entrainment phase.

[3] A supplementary analysis tested whether the Director's egocentric knowledge of the labels influenced referring by comparing conditions where addressee naiveté was held constant, but Director knowledge changed. A comparison of expression length between the 1st and 5th entrainment trials in High CG/Low CG condition was not significant (t=-0.532, p>.05). Likewise, expression length between the 1st and 2nd trials in Low CG/High CG condition did not differ (t=-1.521, p>.05). Thus, Directors designed expressions with respect to addressee knowledge; there was no evidence of an egocentric knowledge influences on utterance length (cf., Gann & Barr, 2012).

Table 2.1. Experiment 1: The average number of words (and standard deviation) used to describe each image during the entrainment phase of Experiment 1 and the results of significance tests, comparing two conditions for each trial separately. The * indicates trials where the partner switched.

Trial	High/Low condition	Low/High condition	Significance test (*t*-value)	*p*-value
1	13.67 (9.98)	13.73 (10.93)	-0.57	0.69
2	7.13 (5.57)	13.00* (9.82)	**-8.56**	**<.001**
3	5.46 (3.87)	6.84 (4.99)	**-3.63**	**<.001**
4	4.82 (3.50)	5.60 (4.05)	**-3.43**	**<.001**
5	13.22* (11.48)	5.31 (4.13)	**7.34**	**<.001**

Test Trials – Expression Length

Our primary analysis focused on the Director's referential descriptions to evaluate whether descriptions were tailored to the common ground shared with the current addressee (Figure 2.5). When common ground had been established recently, Directors used short referring expressions, and the length did not differ depending on the amount of shared knowledge. By contrast, when common ground had not been established recently, Directors produced shorter expressions for high-CG partners than low-CG partners.

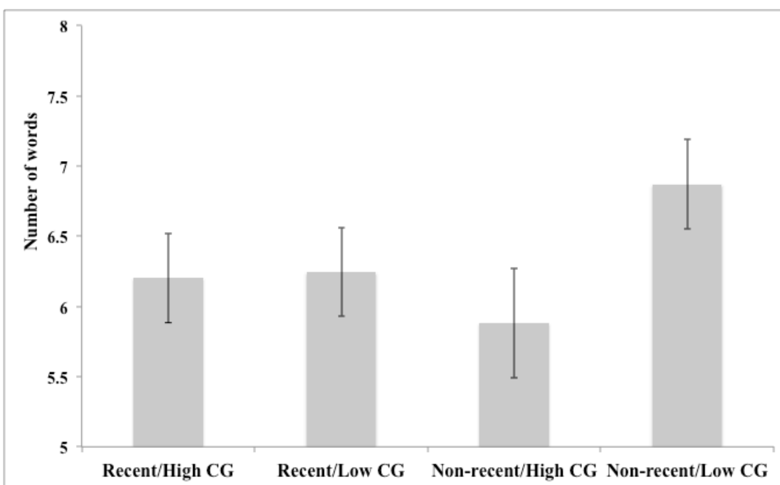


Figure 2.5. Experiment 1: The average number of words per item on test trials. Error bars represent the standard error of the mean by participants.

The statistical analysis revealed a significant interaction between the degree of common ground and recency (z=3.13, p<.05). Planned comparisons revealed that in the non-recent conditions, referring expressions were significantly longer when Directors addressed low-CG partners compared to high-CG partners (z=-2.059, p<.05). By contrast, there was no effect of partner knowledge when common ground had been established recently (z=-0.017, p>.05)[4].

Test Trials – Disfluency

As a supplemental measure, I analyzed the frequency with which speakers produced a disfluency when describing target objects during the test trials (Table 2.2). If Directors have difficulty designing expressions for the low common ground Matcher, disfluency should be more likely. Expressions were coded as disfluent if they contained an utterance-initial disfluency (e.g., *um/uh*… Click on the…"), lengthening of the definite article (e.g., "Click on *thee*…"), or if the expressions included additional descriptive phrases (e.g., "Click on the one… *that looks like*…"). While the rate of disfluency was highest in Non-recent/Low CG condition (38.28%), disfluency rates were low overall. Consistent with a floor effect, the statistical analysis revealed no significant effects of condition (ps>.05).

In my earlier work, speakers produced disfluent expressions over 60% of the time when addressing a naïve matcher (Yoon & Brown-Schmidt, 2014); the uniformly low rate of disfluency in this study may be due to the fact that both the low-CG and the high-CG partners had some experience with the image labels.

Table 2.2. Experiment 1: Percentage of disfluent expressions on test trials.

	Recent/ High CG	Recent/ Low CG	Non-recent/ High CG	Non-recent/ Low CG
Disfluency	32.14%	30.82%	36.96%	38.28%

[4] At test, the speaker randomly alternated addressing the low- and high-CG partners. Whether the speakers addressed the same addressee from trial to trial or switched had no effect on expression length, and there was no interaction with condition (ts<1.0).

Summary and Discussion

In Experiment 1, I examined how interlocutors represent and use common ground in multiparty conversation. I demonstrated that speakers maintain distinct representations of common ground for two different co-present addressees. Despite demonstrations of switching costs in other aspects of perspective-taking (Ryskin, et al., 2014), speakers were able to switch between representations of common ground to accommodate the current addressee. Of note is the fact that the low-CG addressee did, in fact, have some experience with the referential labels. By contrast, most previous work examining the ability to adapt referring expressions to new partners has examined situations with completely naïve addressees (Wilkes-Gibbs & Clark, 1992; Horton & Spieler, 2007). The fact that speakers distinguished between addressees with 1 vs. 4 trials of experience is consistent with previous findings that interlocutors gradually build common ground (Brennan & Clark, 1996), such that the degree to which a given referent is considered common ground varies in a gradient fashion (Brown-Schmidt, 2012).

I also found that when common ground had been established immediately prior to test, that speakers used uniformly short expressions, regardless of the degree of common ground. This effect of recency may be closely tied to the lack of an intervening experience with the other addressee in the recent condition, rather than the minimal time-delay between establishment of entrained terms and their use in this condition: Even individuals with dense amnesia are able to recall lexically-entrained terms at accuracy of 80% six months following entrainment (Duff, Hengst, Tranel, & Cohen, 2006). Similarly, in work on perceptual learning, a short (25min) delay between training and test did not diminish learning of a novel pronunciation (Kraljic & Samuel, 2005).

Of note is the fact that the low-CG addressee did, in fact, have some experience with the referential labels. By contrast, most previous work examining the ability to adapt referring expressions to new partners has examined situations with completely naïve addressees (Wilkes-Gibbs & Clark, 1992; Horton & Spieler, 2007). The fact that speakers distinguished between the low-CG and the high-CG Matcher in the non-recent conditions is consistent with the view that the degree to which a given referent is considered common ground varies in a gradient fashion (Brown-Schmidt, 2012). By contrast, this effect cannot be easily accounted for by a model in which common ground is encoded in a one-bit fashion, as in, my partner has vs. has not discussed this image with me (Galati & Brennan, 2010). Both the low-CG and the high-CG matcher had experience with the labels, so a one-bit partner model would not support the design of longer expressions for the low-CG Matcher.

These findings allow for two important conclusions regarding the process of audience design. The first is that speakers seem to have some degree of control over the representations of common ground that they use when designing referring expressions. Even if the common ground shared with the non-addressee was automatically activated (i.e., Horton, 2007), utterance form was governed by the common ground held with the person the speaker intended to address. Second, this process of accessing the right representation of common ground for the intended addressee is remarkably fast and flexible. The fact that speakers were successful despite frequent partner switches emphasizes the importance of the relevant memory mechanisms in play (see discussion in Horton & Gerrig, 2005). Repeatedly accessing the representations of common ground for each of the two Matchers may have facilitated this process (Mayr, 2006).

Experiment 2

The results of Experiment 1 show that speakers are able to maintain distinct representations of the perspective of each of two conversational partners. Thus, speakers readily adjust referring expressions to be appropriate for a less knowledgeable addressee (Figure 1.2, bottom left), or a more knowledgeable addressee (Figure 1.2, bottom right). Building on this finding, in Experiment 2, I address the question of how speakers design utterances when simultaneously addressing two individuals. Using a method similar to Experiment 1, the Director and one Matcher completed a referential communication task, establishing labels for abstract images. At test, I then examined situations in which the Director must simultaneously address two addressees, only one of whom is familiar with the entrained object labels. How do speakers approach the problem of audience design in situations like this, where the knowledge shared with the addressees is heterogeneous?

Method

Participants

Sixty undergraduates at the University of Illinois at Urbana-Champaign participated in the experiment in return for either partial course credit or cash payment ($12). Three participants took part in the study at the same time and were randomly assigned to the roles of Director, Matcher 1, and Matcher 2. Participants maintained the same roles throughout the experiment. All participants were native speakers of North American English with normal hearing and normal or corrected-to-normal vision. None had participated in Experiment 1.

Materials and Procedure

After being assigned to their respective roles, the participants as a group were given an overview of the experiment and each of their tasks. Like Experiment 1, the experiment was conducted in two phases. During the first phase, the Director established terms with Matcher 1 across 5 repeated trials. The procedure and materials were identical to the entrainment phase of Experiment 1, with the exception that Matcher 2 never participated in the sorting task. This change was instituted in order to simplify the experimental design. (For consistency, I will continue to refer to Matcher 2 as the "low-CG" matcher despite the fact that this partner did not have knowledge of the object labels going into the test phase)

In the second phase of the task, the participants completed 40 test trials, including 32 target and 8 filler trials. On each of the 32 target trials, the Director referred to one of the old images that had previously been described in Phase 1. Each old image was referred to twice during the test trials, such that 16 of the target trials were first descriptions, and 16 were second descriptions. Old images were referenced twice in order to examine whether individual participants would describe the same image with more words when speaking to the low-CG matcher than the high-CG matcher. In addition, on 8 filler trials, the Director referred to the new image (new images were only referenced once).

The test display was identical to the test phase of Experiment 1 (Figure 2.3). Unlike Experiment 1, I did not track eye-gaze. The primary manipulation was the identity of the addressee(s) during test trials. Four different addressee conditions were manipulated in a blocked, within-subjects design (Figure 2.7): High CG (Matcher 1 only), Low CG (Matcher 2 only), High+Low CG (Matcher 1 and 2 together), and High/Low CG

Switching (Matcher 1 and 2 alternating). In all conditions, following entrainment, Matcher 1 left the room, and then the Matcher(s) for the test trials entered the room.

In the High CG and Low CG conditions, the Director gave instructions to one partner, either the knowledgeable Matcher 1 or the naive Matcher 2, while the other Matcher performed unrelated tasks (e.g., verbal and math tasks) in a separate room.

In the High+Low CG conditions, both Matcher 1 and Matcher 2 were present in the room during test trials, and the Director gave instructions that both Matcher 1 and Matcher 2 followed at the same time. The experimental instructions made it clear to all participants that Matcher 1 and 2 were to both follow the Director's instructions.

Lastly, in High/Low CG Switching condition, as in Experiment 1, both Matcher 1 and Matcher 2 were present during test, and the Director was prompted to tell either Matcher 1 or 2 (randomly alternating) to click on one of the 4 objects.

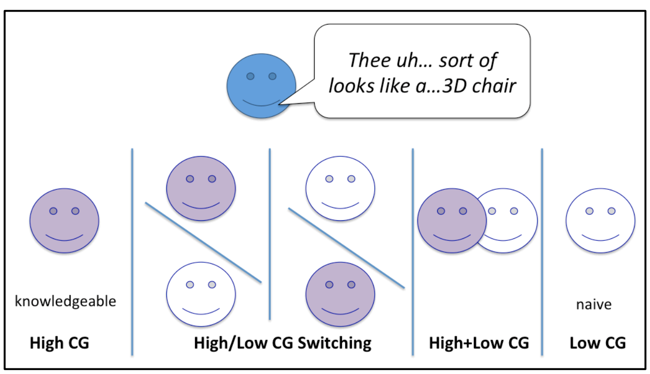


Figure 2.6. Experiment 2: Illustration of the four different addressee conditions that were manipulated in a blocked, within-subjects design during test trials.

Like Experiment 1, the two phases of the experiment (Phase-1: sorting; Phase-2: testing) were repeated four times for each triplet of participants with different materials. The order of the rounds was counterbalanced across groups of participants. The four rounds of play were otherwise identical with Director and Matcher 1 completing the image sorting task 5 times, followed by test trials with four different conditions. The entire experiment lasted approximately 90 minutes.

Data Analysis

As in Experiment 1, I first analyzed the change in the length of referring expressions during entrainment to verify that partners converged on brief labels for each image. Our primary analysis examined the length of the referring expressions at test. The disfluency rate on test trials was used as a measure of planning difficulty. Finally, I examined the frequency with which Directors reconceptualized their descriptions of the referents at test (Horton & Gerrig, 2002), as a secondary measure of audience design. When describing a referent for a naive addressee (Horton & Gerrig, 2002; Yoon & Brown-Schmidt, 2014), speakers tend to change previously established referential descriptions by adding new or different content words. Such reconceptualizations may help addressees identify the intended referent in situations where an expression previously established with a different partner would be opaque to the new partner[5].

Predictions

I evaluated three candidate hypotheses regarding the way in which speakers may approach audience design in multiparty settings. The predictions are as follows:

Hypothesis 1: AIM LOW. Speakers design utterances for the person with the least common ground at the expense of efficient communication with knowledgeable addressees. Directors should use longer expressions when describing test images in the

[5] Note Experiment 1 did not include a reconceptualization analysis because partial common ground was established with both high-CG and low-CG matchers at entrainment. Thus higher reconceptualization rates with the low-CG matcher would not unambiguously point to a re-design for the low-CG matcher; instead it might reflect distinct entrainment with the two matchers.

Low CG and High+Low CG conditions, compared to High CG condition, with no difference between the Low CG and High+Low CG conditions.

Hypothesis 2: Combine. Speakers may estimate access representations of each addressee's knowledge, and combine them in some way. One possibility that we consider is that the speaker calculates the average knowledge state of all addressees and designs expressions with respect to the average. Directors should use longer expressions in Low CG condition than the High CG condition, with expression length in the High+Low CG condition in-between.

Hypothesis 3: AIM HIGH. Speakers may design expressions for the person with the most common ground, at the expense of naïve addressees. Expression length should not differ in the High CG condition and the High+Low condition; both conditions should produce shorter expressions as compared to the Low CG condition.

Finally, the comparison between referring expressions in the high-switching and the low-switching conditions will allow us to replicate the finding of Experiment 1 that speakers readily alternate between distinct common ground representations. Additional comparisons between the switching conditions with the single-addressee conditions (High CG and Low CG) will indicate whether the presence of the non-addressee in the switching conditions affected the audience design process.

In earlier research (Yoon & Brown-Schmidt, 2014), speakers used longer expressions when addressing the knowledgeable and naïve partners simultaneously compared to when speakers addressed the knowledgeable partner alone. While these findings, like the results of Experiment 1, show that speakers are sensitive to the needs of a relatively naïve partner, the design of that experiment did not allow us to distinguish between the *Aim low* and *Combining* hypotheses, as both predict longer references when one of the addressees is naive. The present research adds a critical new comparison condition – one in which the Director addresses the *naïve* partner alone—that allows us to tease apart these two hypotheses. Finally, while Yoon and Brown-Schmidt's (2014) findings are inconsistent with *Aim high*, Experiment 2 provides an opportunity to replicate this finding.

Another outstanding issue is whether in Experiment 1 (and in Yoon & Brown-Schmidt, 2014), the presence of the non-addressee affected the audience design process. While the results of Experiment 1 show that utterance design was guided by the knowledge state of the intended addressee, it is possible that the non-addressee influenced the production process nonetheless. One reason that the non-addressee could have affected referring is the idea that people serve as memory cues that result in automatic activation of information that is associated with them (Horton, 2007). If so, the presence of the more knowledgeable partner may have automatically activated the entrained terms, increasing the likelihood that the speaker would use an entrained term that the naïve addressee was unfamiliar with. In Experiment 2, I address this issue directly by comparing audience design in situations where a non-addressee is vs. is not present in the testing room.

In sum, the aim of Experiment 2 is to evaluate candidate hypotheses regarding the way in which speakers design referring expressions in multiparty conversation. Note that while each of the three hypothetical approaches to audience design that I outline here (*Aim Low*, *Combining*, and *Aim High*) could be accomplished in a strategic fashion, basic attentional and memory mechanisms could also give rise to these processes. The aim is to evaluate which of these approaches best characterizes audience design in the present settings.

Results

As in Experiment 1, data analysis was conducted using mixed-effects models. For full model details, see Appendix A.

Entrainment Trials

Analysis of the Director's referential expressions during entrainment showed a typical pattern of referential shortening across trials (z=12.40, p<.05), consistent with gradual establishment of common ground (Figure 2.).

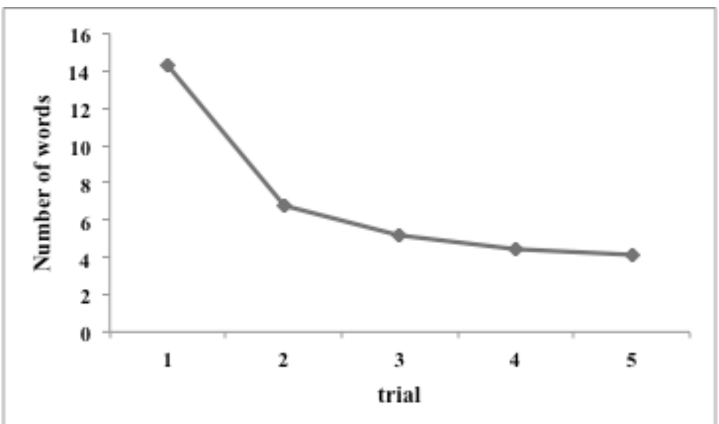


Figure 2.7. Experiment 2: The average number of words used during entrainment trials.

Test Trials – Expression Length

Our primary analyses examine the length of the referring expressions used to describe target images at test (Figure 2.8). Note the High/Low CG Switching condition was separated into two conditions, depending on whether the Director was addressing Matcher 1 (high-switching) or Matcher 2 (low-switching). As in Experiment 1, the dependent measure was word count; fixed effects were condition and repetition (first vs. second mention of the image). Because the critical conditions were not intended to be orthogonal, the data were analyzed in two separate models (see details in Appendix A).

The first set of analyses was designed to examine how Directors approached the problem of audience design when simultaneously addressing both a knowledgeable and a naïve partner. To this end, I compared expression length in the High+Low CG condition, with length in both the High CG and the Low CG conditions. This analysis revealed a significant main effect of repetition (z=4.586, $p<.05$), such that Directors used shorter expressions when describing target objects a second time at test. Directors also used shorter expressions when addressing their partner in the High CG condition than the two conditions where Directors addressed the Low CG partner (Low CG condition and High+Low CG condition; $z=6.751$, $p<.05$). Expressions were of equivalent length in the Low-CG and the High+Low CG conditions ($z=1.158$, $p>.05$). These findings show that Directors used long expressions any time the naïve partner was an addressee. These main effects were qualified by a significant interaction between repetition and partner knowledge ($z=6.072$, $p<.05$). The interaction was due to a significant decrease in the number of words for the second description of each item in the Low CG and High+Low CG conditions ($z=5.424$, $p<.05$), but not in the High CG condition ($z=0.60$, $p>.05$). This is likely because Directors were already using the briefest possible description when addressing the High CG partner. These findings offer support for the Aim Low hypothesis: Directors formulated expressions for the most ignorant addressee.

The second set of analyses compared the high-switching to the low-switching conditions. Consistent with Experiment 1, Directors flexibly designed expressions given the knowledge of the current addressee, using shorter descriptions for the high-CG matcher than the low-CG matcher in High/Low switching condition ($z=7.019$, $p<.05$). Note that the repetition effect was not significant ($z=1.717$, $p>.05$), which is unsurprising because when images were described for a second time, the Director was speaking to a different addressee. An outstanding question that was not addressed by Experiment 1 is whether the presence of the partner that the Director was *not* addressing affected audience design. To answer this question, I compared expression length for the first descriptions of targets in the high-Switching condition with the High CG condition, and the low-Switching condition with the Low CG condition. These analyses revealed that the non-addressee had no effect (high-Switching vs. High CG: z=1.124, p>.05; low-Switching vs. Low CG: z=0.308, p>.05)[6].

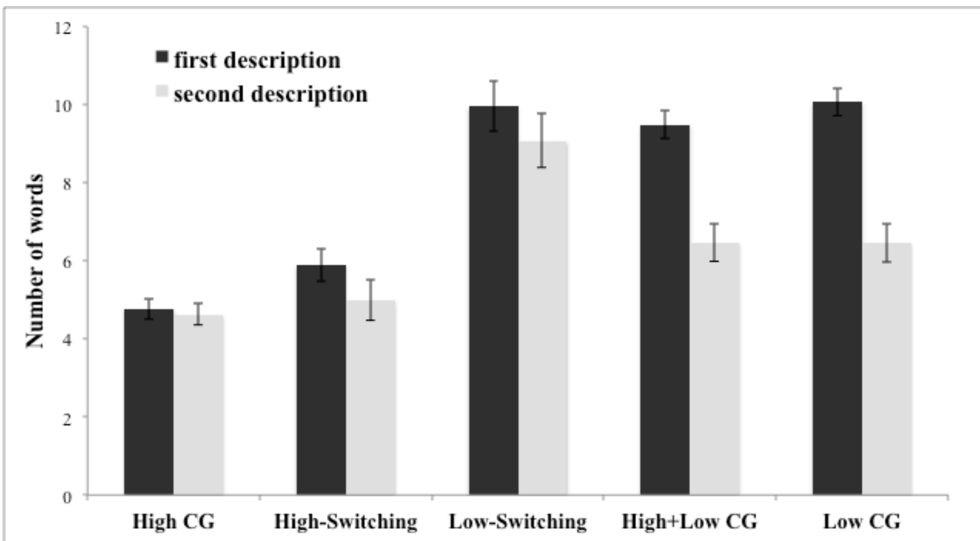


Figure 2.8. Experiment 2: The average number of words used to describe each image on test trials. Error bars represent the standard error of the by-participant mean. In the switching conditions, the first and second descriptions correspond to two different addressee-Matchers; in the other three conditions the Matcher was the same across the first and second descriptions.

Test Trials – Disfluency

Disfluency was analyzed as an indirect measure of planning difficulty (Table 2.3), and was coded in the same way as Experiment 1. The pattern of results was similar to

[6] As in Experiment 1, there was also no evidence that switching between addressees at test affected referring; whether the speaker addressed the same addressee from trial to trial or switched had no effect on expression length, and there was no interaction with condition (*ts*<1.0).

those for expression length. Directors were more likely to produce a disfluent description when speaking to a low-common ground Matcher (Low CG & High+Low CG) than to the knowledgeable Matcher (High CG), (z=6.810, p<.05). Consistent with an *Aim Low* approach to audience design, disfluency rates did not significantly differ between the High+Low CG and Low CG conditions (z=0.644, p>.05). As with the analysis of expression length, an interaction between partner knowledge and repetition (z=4.547, p<.05) was due to decreases in disfluency for repeated descriptions in the High+Low and Low-CG conditions, but no change across repetitions in the High-CG condition.

Analysis of the conditions in which Directors alternated addressing the high- and low-CG partners indicated that Directors were more likely to be disfluent when speaking to a low-CG Matcher than to a high-CG Matcher (z=8.585, p<.05), consistent with more effortful audience design when addressing the naïve Matcher. The presence of an overhearer did not affect disfluency rates (high-Switching vs. High CG: z=0.728, p>.05; low-Switching vs. Low CG: z=-0.657, p>.05).

Table 2.3. Experiment 2: Percentage of disfluent expressions on test trials.

	High CG	High-Switching	Low-Switching	High+Low CG	Low CG
First description	22.61%	25.19%	80.70%	75.89%	74.92%
Second description	21.60%	22.68%	74.54%	26.90%	31.25%

Test Trials – Reconceptualizations

The final set of analyses examined the rate with which Directors reconceptualized referring expressions at test as a supplemental measure of audience design (Table 2.4). To identify reconceptualizations, I compared the final expression that was established with the high-CG Matcher during entrainment, with the first expression used for the same image during test. If the expression on test trials contained different content words compared to the entrained term, this was considered a reconceptualization.

Directors were more likely to reconceptualize when they spoke to a naïve addressee (High+Low and Low-CG) than to the knowledgeable addressee alone ($z=8.039$, $p<.05$), demonstrating sensitivity to the naïve addressee's needs. However, in contrast to the pattern of findings for expression length, Directors were 14% more likely to reconceptualize when addressing the Low CG partner alone, vs. when they addressed both the Low and the High CG partners ($z=2.361$, $p<.05$). This result reveals some sensitivity to the presence of the more knowledgeable addressee in two-addressee settings. Finally, the analysis of the switching conditions revealed that Directors reconceptualized more often when addressing the Low CG partner than the High CG partner ($z=8.793$, $p<.05$). The non-addressee overhearer did not affect referring (high-Switching vs. High CG conditions $z=1.276$, $p>.05$; low-Switching vs. Low CG $z=-1.035$, $p>.05$).

Table 2.4. Experiment 2: Percentage of reconceptualized expressions on test trials.

	High CG	High-Switching	Low-Switching	High+Low CG	Low CG
First description	20%	25.62%	73.08%	59.06%	72.37%
Second description	22.31%	21.39%	70.36%	49.68%	53.12%

Summary and Discussion

Speakers produced longer and more disfluent expressions when simultaneously speaking with a naïve and a knowledgeable addressee in a three-party conversation vs. when they addressed a knowledgeable addressee alone. By contrast, expressions were equally long and disfluent when speakers addressed a naïve partner in two-party conversation, as when they addressed both a naïve and a knowledgeable partner in three-party conversation. These findings show that speakers *Aimed Low* – they designed expressions for the least knowledgeable party. One analysis, did however, offer a hint that speakers were sensitive to the more knowledgeable party in three-person conversations: Speakers were more likely to use an entrained term (and not reconceptualize) when addressing both the naive and knowledgeable matcher vs. the naive matcher alone. This result suggests that the act of addressing a partner with common ground for an entrained term may make that entrained term more accessible in memory.

Finally, the analysis of flexibility replicates the main findings of Experiment 1: speakers maintain distinct representations of common ground for different individuals, and are able to flexibly alternate between these representations in an addressee-appropriate manner. There was no evidence (on any measure) that the non-addressee influenced referring; when speakers were addressing the naive matcher, the mere presence of the knowledgeable partner did not increase the use of entrained terms.

Experiment 3

In Experiment 2, speakers lengthened expressions when simultaneously addressing a naïve and a knowledgeable listener, compared to a knowledgeable listener alone. While these findings are consistent with *Aiming Low*, an open question is whether in other circumstances, speakers might adopt a different strategy and design utterances with respect to some combination of the addressees' knowledge states. Specifically, we ask whether speakers might depart from Aim Low in larger groups where the majority of addressees were knowledgeable of the object labels. I examine this question in both 4-party conversation (Experiment 3a) and 5-party conversation (Experiment 3b).

Experiment 3a

Method

Participants

One hundred twelve undergraduates (twenty-eight groups) at the University of Illinois at Urbana-Champaign participated in the experiment in return for either partial course credit or cash payment ($12). Four participants took part in the study at the same time and were randomly assigned to the roles of Director, Matcher 1, Matcher 2, and Matcher 3. All participants were native speakers of North American English, and had no reported hearing or uncorrected visual impairment. None had participated in Experiments 1 and 2.

Materials and Procedure

The procedure and materials of Experiment 3a was identical to Experiment 2 in that experiment was conducted in two phases, entrainment and test. The primary manipulation was the identity of the Matcher or Matchers and their knowledge during the test phase: one knowledgeable matcher (1K), one naïve matcher (1N), two knowledgeable and one naïve matchers (2K1N), and one knowledgeable and two naïve matchers (1K2N).

During the first phase, the Director and Matcher 1 established image labels as a dyad, while the other Matchers performed an unrelated task (e.g., solving math problems) in another room. Thus, Matcher 1 became knowledgeable about the image labels, while the other Matchers remained naïve. In the 2K1N condition, the Director subsequently repeated the same entrainment phase with Matcher 2. In all cases Matcher 3 remained ignorant of the labels (Figure 2.9).

At test, in each of the four conditions, the Director instructed one addressee (the 1K and 1N conditions) or all three addressees (the 2K1N and 1K2N conditions) to click on one of the four objects. In the 1K and 1N conditions, the other Matchers performed unrelated tasks (e.g., verbal and math tasks) in a separate room. Participants completed 24 target trials, including 16 target and 8 filler trials. The test displays were identical to the displays used in the test phase of Experiment 2, including 3 old images from entrainment and 1 new image. On target trials, the Director referred to one of the old images, while on filler trials, he or she referred to the new image. Unlike in Experiment 2, old images were referred only once during the test phase.

As in Experiments 1-2, the conditions were manipulated in a blocked, within-subjects design. Four different sorting-testing blocks rotated each participant group through the four conditions; the order of blocks was counterbalanced across groups.

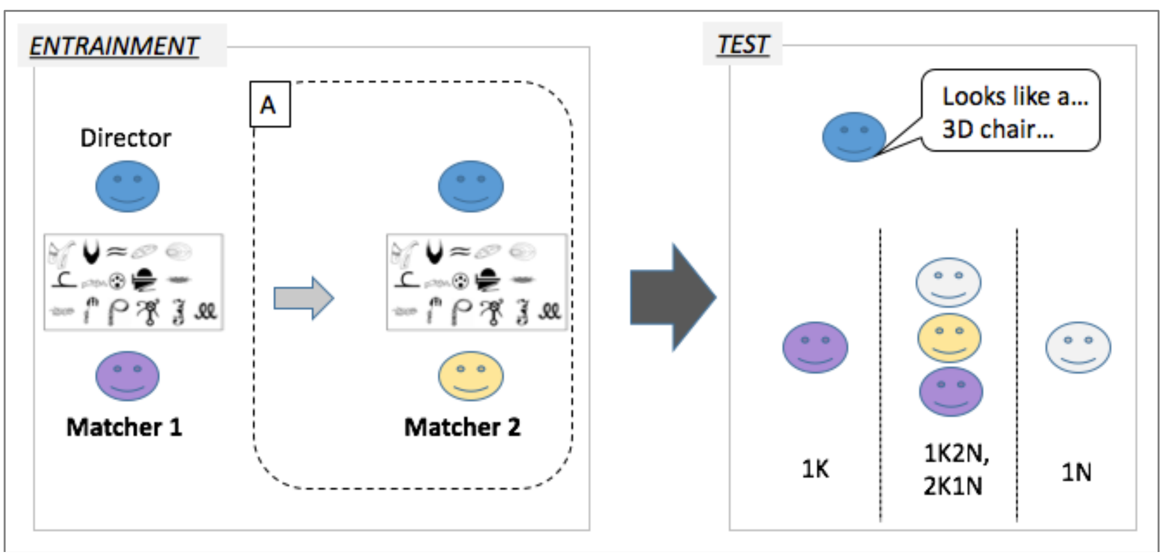


Figure 2.9. Experiment 3a: Example procedure and illustration of the four different addressee conditions that were manipulated in a blocked, within-subjects design during test trials. This is an example in the 2K1N condition. Note that in the 1K, 1K 2N, and 1N conditions, the dotted box A (the second round of entrainment with Matcher 2) was not performed.

Predictions

This experiment was designed to test two candidate hypotheses regarding how speakers design utterances in multiparty conversation where each individual in a group shares different amount of joint knowledge.

Hypothesis 1: *AIM LOW*. As in Experiment 2, Directors' expressions should only be sensitive to the presence of a naïve addressee. They should use longer expressions whenever there is at least one naïve addressee in a group. If so, expression length should be shorter in the 1K condition, and longer in the 2K1N, 1K2N and 1N conditions. Further, the rate of disfluency should be lower in the 1K condition than in the 2K1N, 1K 2N and 1N conditions. However, *Aim Low* predicts that expression length and disfluency rate should not differ between the 2K1N, 1K2N, and 1N conditions.

Hypothesis 2: *COMBINE*. Speakers may be sensitive to the presence of both knowledgeable and naïve partners. Thus speakers must accommodate two distinct needs in conversation: Knowledgeable partners need short, efficient expressions, whereas naïve partners need longer, more informative expressions. If speakers combine the distinct needs of knowledgeable and naïve partners, expressions should be lengthened as combined knowledge state decreases. Thus we should expect an increase in expression length and rate of disfluency across the 1K, 2K1N, 1K2N, and 1N conditions.

Results

As in Experiments 1-2, expression length in both entrainment and test trials and the rate of disfluency during test trials were analyzed as a function of condition. Data analysis was conducted using mixed-effects models (see Appendix A).

Entrainment Trials

Directors completed four trials of entrainment with Matcher 1 in the 1K, 1K2N, and 1N conditions. In the 2K2N condition, the Director then completed four more entrainment trials with Matcher 2. During entrainment, Directors produced shorter expressions across trials, consistent with the previous findings in Experiments 1 and 2 (Figure 2.10).

The statistical analysis of the entrainment trials examined how Directors established common ground and whether the entrainment process with Matchers 1 and 2 differed. If the Director's egocentric knowledge of the object labels affects entrainment, we might expect Directors to produce shorter expressions when completing entrainment trials with Matcher 2. The dependent measure was expression length during entrainment trials and fixed effects included trial (1-4) and Matcher (Matcher 1 vs. 2). The model revealed a main effect of trial: the expression length decreased across trials (z=-13.64, p<.05). Neither the main effect of matcher (z=1.59, p>.05) nor the interaction between trials and matcher (z=0.39, p>.05) was significant. When the Directors addressed Matcher 2, they had already described the same images four times with Matcher 1. Although the Director had previous experience describing the same images, we found no evidence that this egocentric knowledge influenced their referential expression (cf. Gann & Barr, 2012). Instead, Directors designed expressions based on their current partners' knowledge state.

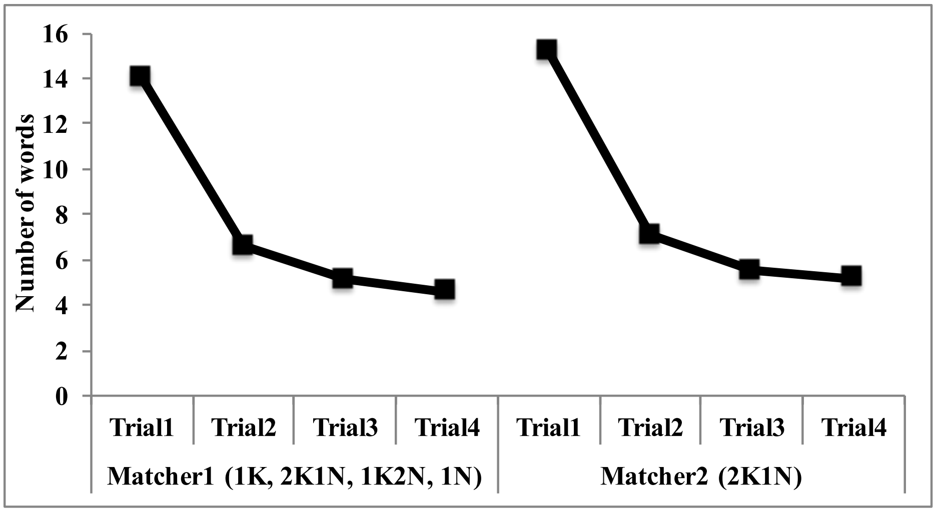


Figure 2.10. Experiment 3a: The average number of words used to describe each image across the four trials of entrainment. Note that Matcher 1 participated in entrainment trials across all four conditions, while Matcher 2 participated in entrainment trials in only the 2K1N condition.

Test Trials – Expression Length

During test trials, Directors produced gradually longer expressions as the combined naiveté of the addressees increased (Figure 2.11). Expression length was shortest in the 1K condition and longest in the 1N condition. Expression lengths in the 2K1N and 1K2N were in-between.

I analyzed the Directors' referential expressions during test using a maximal mixed-effects model. Condition was included as a fixed effect and the dependent measure was the number of words that the Director used to describe each image. Expressions in the 1K condition were shorter than the other four conditions (z=10.261, p<.05), and expressions in the 2K1N condition were shorter than the 1K2N and 1N conditions (z=3.541, p<.05). There was no difference between the 1K2N and 1N conditions (z=0.984, p>.05). Overall, these findings are contrary to the prediction of the *Aim Low* hypothesis, Directors were sensitive to the *combined* knowledge states of the addressees and seemed to design their utterances with respect to the overall knowledge state. The fact that there was no significant difference between the 1K2N and 1N conditions suggests that expression length did not linearly increase as combined knowledge states decreased. The lack of a significant difference between the 1K2N and 1N conditions may reflect a ceiling effect. Alternatively, it may suggest that speakers are not designing utterances based on mathematical average of addressees' knowledge states, and instead weighing the needs of naïve addressees more heavily in determining how to design expressions.

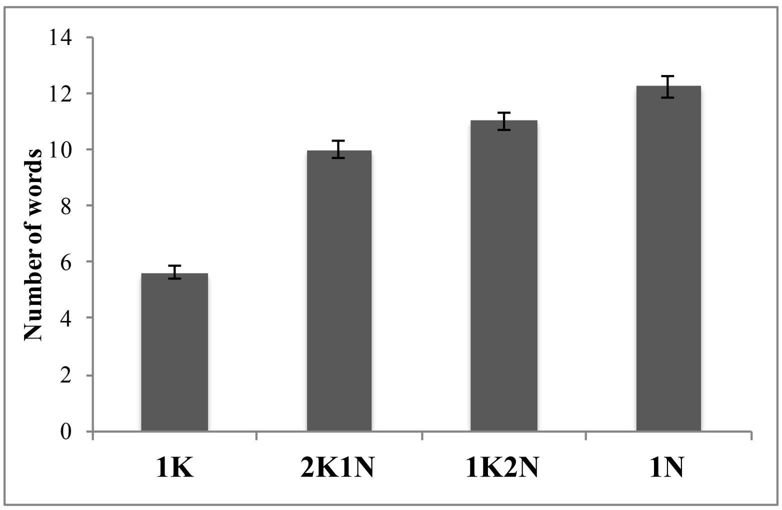


Figure 2.11. Experiment 3a: The average number of words used to describe each image on test trials. Error bars represent the standard error of the by-participant mean.

Test Trials – Disfluency

As in Experiments 1 and 2, I analyzed the rate of disfluency as a measure of the Directors' planning difficulty. Disfluency was coded in the same way as the previous experiments. Similar to the analysis of expression length, Directors were least disfluent in the 1K condition and most disfluent in the 1N condition (Table 2.5).

The analysis of disfluency rate included condition as a fixed effect; the dependent measure was binary, whether the Director was disfluent or not. This analysis revealed that Directors were less disfluent in the 1K condition than in the other three conditions (*z*=9.17, *p*<.05), and also less disfluent in the 2K1N condition than in the 1K2N and 1N conditions (*z*=2.25, *p*<.05). The rate of disfluency did not differ between the 1K2N and 1N conditions (*z*=1.48, *p*>.05).

The results indicated that as combined naiveté increased, speakers were gradually more disfluent. These findings are consistent with the *Combine* hypothesis—speakers designed utterances with respect to a combination of the knowledge of the intended addressees. The fact that the disfluency rate in the 1K2N is significantly higher than in the 2K1N condition (*z*=2.52, *p*<.05) supports the idea that speakers consider the composition of the group's knowledge, rather than group size itself, when they design expressions.

Table 2.5. Experiment 3a: Percentage of disfluent expressions on test trials.

	1K	2K1N	1K2N	1N
Disfluency	31.25%	56.15%	65.10%	70.85%

Experiment 3b

Method

Participants

One hundred sixty-five undergraduates (thirty-five groups) at the University of Illinois at Urbana-Champaign participated in the experiment in return for either partial course credit or cash payment ($16). Five participants took part in the study at the same time and were randomly assigned to the roles of Director, Matcher 1, Matcher 2, Matcher 3, and Matcher 4. All participants were native speakers of North American English, and had no reported hearing or uncorrected visual impairment. None had participated in Experiments 1, 2, or 3a.

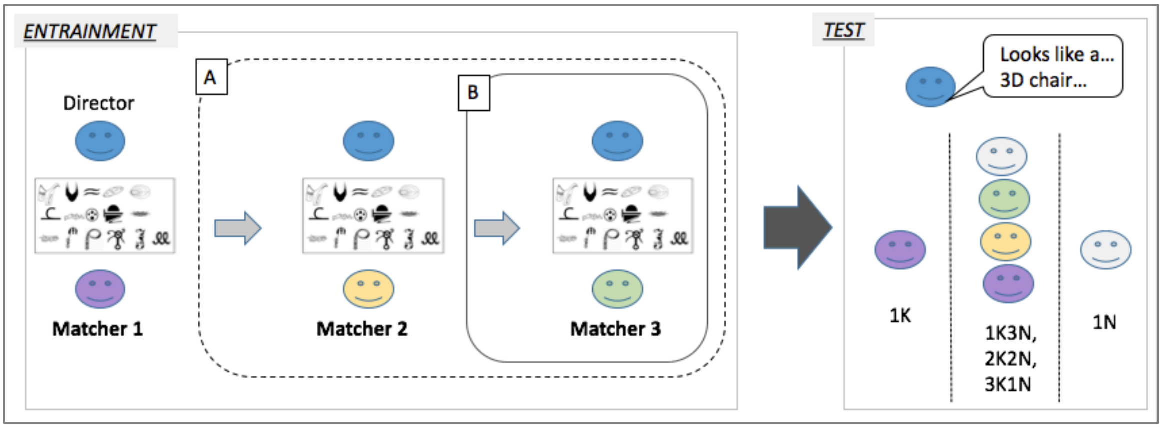


Figure 2.12. Experiment 3b: Example procedure and illustration of the five different addressee conditions that were manipulated in a blocked, within-subjects design during test trials. This is an example in the 3K1N condition. Note that in the 1K, 1K3N, and 1N conditions, dotted box A (entrainment trials with Matcher 2 and 3) was not performed, and in the 2K2N condition, dotted box B (entrainment trials with Matcher 3) was not performed.

Materials and Procedure

The procedure of Experiment 3b was identical to Experiment 3a, including two phases, entrainment and test (Figure 2.12). The key manipulation was the number of Matchers and their knowledge during the test phase: one knowledgeable matcher (1K), one naïve matcher (1N), three knowledgeable and one naïve matchers (3K1N), two knowledgeable and two naïve matchers (2K2N), and one knowledgeable and three naïve matchers (1K3N).

In the 2K2N and 3K1N conditions, the entrainment phases were repeated sequentially with two (2K2N) or three (3K1N) addressees. In the 2K2N condition, Matcher 1 and Matcher 2 sequentially performed entrainment phases, and then all Matchers participated in the test phase. In the 3K1N condition, Matcher 1, 2, and 3 sequentially performed entrainment phases before the test phase. Following entrainment, at test, the Director described the target image for the Matcher or Matchers (depending on the condition) so that the Matcher(s) could click the target on their screen(s). 5 different sets of images were used to rotate each group of participants through the 5 conditions across 5 blocks of trials. The 5 different addressee conditions were manipulated in a blocked, within subject design and the order of blocks was counterbalanced across groups.

Predictions

As in Experiment 3a, I tested the same candidate hypotheses in 5-party conversation and the predictions were similar to those in Experiment 3a. If speakers always *Aim Low* (Hypothesis 1), I expected longer expressions in the 1K3N, 2K2N, 1K3N, and 1N compared to the 1K condition, but no significant difference between the first four conditions. In contrast, if speakers *Combine* the goals of efficiency and informativity (Hypothesis 2), then expression length should vary according to the mixture of common ground among the addressees. On this view, I would predict a general increase in expression length and disfluency rate as the combined knowledge state of the addressees decreases.

Results

The same measures as in Experiment 3a, expression length in entrainment and test trials, and rates of disfluency were analyzed (For full models, see Appendix A).

Entrainment Trials

Directors performed one round of entrainment with Matcher 1 in the 1K, 1K3N, and 1N condition, prior to test. Directors performed one round of entrainment with Matcher 1 and then one round with Matcher 2 in the 3K1N and 2K2N conditions. Directors completed three sequential rounds of entrainment (with Matchers 1, 2, and 3) in the 3K1N condition. Note that entrainment trials were always performed in a dialogue between the Director and one of Matchers. Thus, in the conditions where multiple Matchers had knowledge of the object labels, these entrainment trials were completed in sequence.

Consistent with previous findings, Directors shortened their referential expressions as they repeated their descriptions of the same images across trials (Figure 2.13). The pattern of referential shortening across trials was similar for Matchers 1, 2, and 3, despite the fact that the Directors' own egocentric knowledge accumulated across trials and rounds.

Our analysis of the entrainment trials evaluated whether the establishment of common ground with each particular Matcher across the four trials was influenced by Directors' egocentric knowledge. If Directors shorten utterances across trials based on the *addressee's* knowledge alone, we should see shorter expressions across trials, but no difference in expression length as a function of which Matcher the director is talking to. By contrast, if Directors are influenced by their egocentric familiarity with the object labels, then expressions should be significantly shorter with Matchers 2 and 3, compared to Matcher 1.

In a mixed-effect model, Matcher and trials (1-4) were included as fixed effects and the dependent measure was the number of words that Directors used to describe each image. The model showed a significant main effect of trial (z=-19.42, p<.05) and an interaction between Matcher (Matcher 1 vs. Matcher 2 and 3) and trial (z=4.23, p<.05), but not a main effect of Matcher (z=-1.18, p>.05). The significant main effect of trial was due to a reduction in description length across the four trials, consistent with gradual establishment of common ground. The interaction was mainly driven by a marginal difference between utterances designed for Matcher 1 vs. utterances designed for Matchers 2 or 3 on the fourth trial (z=1.82, p=0.07). On the fourth trial, Directors produced slightly shorter expressions for Matcher 1 than Matcher 2 and 3, but this difference was marginal. During trials 1-3, expression length did not differ as a function of the Matcher. These findings show that Directors design referential expressions based primarily on the current addressee's needs, rather being influenced by their own egocentric knowledge.

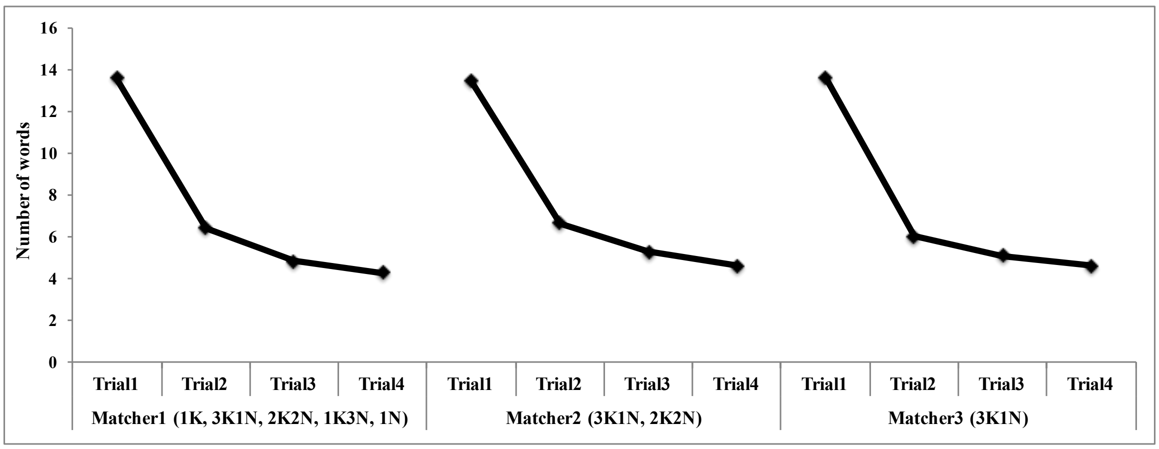


Figure 2.13. Experiment 3b: The average number of words used to describe each image across the four trials of entrainment. Note that the Director completed entrainment trials with Matcher 1 in all five conditions. Following entrainment with Matcher 1, in the 3K1N and 2K2N conditions, the Director completed 4 more entrainment trials with Matcher 2. In the 3K1N condition only, the Director then completed 4 more entrainment trials with Matcher 3.

Test Trials – Expression Length

During test trials, directors increased expression length as the combined naiveté of the Matchers increased (Figure 2.14). Directors produced the shortest expressions in the 1K condition and the longest expressions in the 1N condition. The expression lengths in the other conditions were in-between (3K1N, 2K2N, and 1K3N).

Expression length was analyzed in the mixed-effects model, including condition as a fixed effect. Expressions in the 1K conditions were shorter than the other four conditions (z=15.13, p<.05), and expressions in the 3K1N condition were shorter than expressions in the 2K2N, 1K3N, and N conditions (z=2.43, p<.05). There was no difference in expression length among the 2K2N, 1K3N, and N conditions (2K2N vs. 1K3N and N: z=0.83, p>.05; 1K3N vs. N: z=1.65, p>.05).

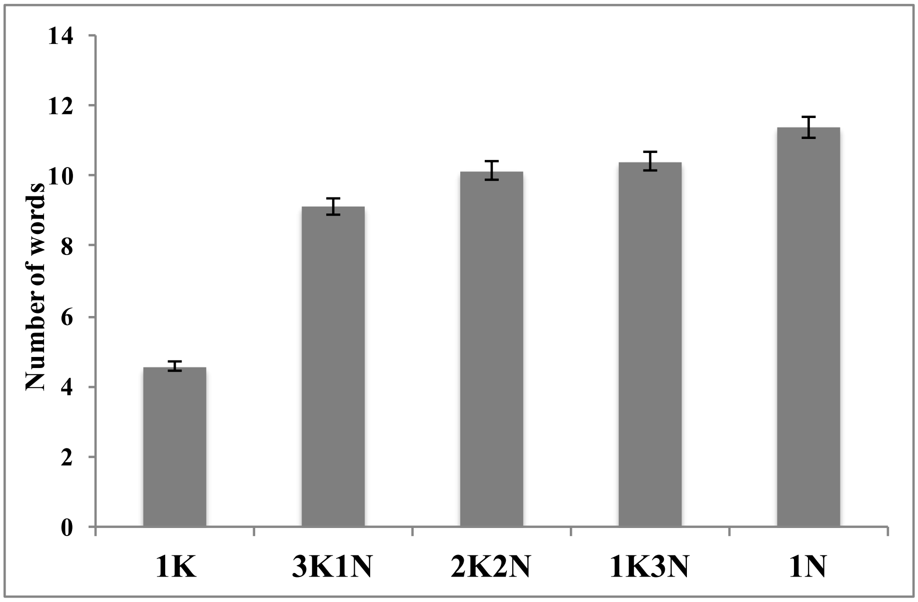


Figure 2.14. Experiment 3b: The average number of words used to describe each image on test trials. Error bars represent the standard error of the by-participant mean.

Test Trials – Disfluency

Disfluency was coded in the same way as in the previous experiments. The rate of disfluency increased as the combined knowledge states among addressees decreased (Table 2.6). Directors were least disfluent in the 1K condition and most disfluent in the 1N condition.

The disfluency rate was analyzed in a mixed-effect model that included condition as a fixed effect; a binary measure of disfluency was used as the dependent measure. Directors were significantly less disfluent in the 1K conditions than in the other four conditions (z=5.61, p<.05). The disfluency rate was lower in 3K1N condition than in the

2K2N, 1K3N, and 1N conditions (z=2.280, p<.05). The disfluency rate in the 2K2N condition did not differ from the rate in the 1K3N and 1N conditions (z=1.66, p>.05), while the disfluency rate between the 1K3N and 1N conditions did significantly differ (z=2.39, p<.05). These findings are generally consistent with the *Combining* hypothesis--that Directors balance the distinct needs of knowledgeable and naïve addressees when designing expressions.

Table 2.6. Experiment 3b: Percentage of disfluent expressions on test trials.

	1K	3K1N	2K2N	1K3N	1N
Disfluency	25.13%	46.33%	50.18%	51.25%	60.93%

Summary and Discussion of Experiments 3a-3b

During the entrainment phases of Experiments 3a-3b, Directors gradually established common ground, shortening expressions across repeated trials. Even when they had already completed four rounds of entrainment with a different Matcher, Directors designed referential expressions based on the current addressee's knowledge state. Directors produced long, descriptive utterances on the first round of entrainment, regardless of whether they were directing the 1st, 2nd, or 3rd Matcher. These utterances became shorter across successive trials of entrainment in a similar way for all three Matchers. Our findings are in contrast to results from earlier work by Gann and Barr (2012) that showed speakers designed shorter referential expressions for a naïve partner when the speaker had egocentric knowledge of the object labels compared to when they were naïve. In Gann and Barr (2012)'s study, they compared speakers' descriptions of old and new referents for a naïve partner on test trials, while in Experiments 3a-3b, speakers' descriptions on the first round of entrainment were compared with respect to the current partner (the 1st, 2nd, or 3rd Matcher). Across two studies of Experiments 3a and

3b, the primary determinant of utterance form was the *addressee's knowledge state,* inconsistent with the idea that utterance design is guided by egocentric knowledge.

During test trials, speakers considered the knowledge states of multiple addressees and flexibly designed expressions with respect to the mixtures of common ground among addressees. The results of Experiment 3a and b were most consistent with *Combine* hypothesis. As the combined knowledge states of the addressees decreased, expression length and the rate of disfluency numerically increased. In particular the transition from a single knowledgeable addressee (1K) to a group with one or more naïve addressees resulted in a large and significant increase in utterance length. We also observed in both studies a point at which the utterance length seemed to asymptote.

One explanation is that speakers base audience design on the dominant knowledge state among the set of addressees: When naïve partners are dominant (e.g., 1K2N and 1N in Experiment 3a; 1K3N and 1N in Experiment 3b), speakers may emphasize the needs of the naïve addressee(s) more than the needs of the knowledgeable addressee, and produce long, descriptive expressions. In contrast, when knowledgeable partners dominate the group, speakers consider both conversational efficiency and informativity when designing expressions.

Directors were particularly sensitive to the presence of at least one naïve participant in a group, demonstrating a dramatic increase in expression length and disfluency rate from the 1K condition to the 2K1N or 3K1N conditions. This implies that when combining knowledge states, speakers take informativity for naïve addressees into consideration more than efficiency for knowledgeable addressees. Of course, the task goals may be relevant. In the task, all of the addressees were required to identify the target for each trial; this task may have caused speakers to emphasize informativity more than efficiency. Thus, the way in which the speaker balances informativity and efficiency may reflect an optimization process in situations where addressees have different needs. The fact that speakers did design utterances differently depending on the composition of our participant groups shows that speakers do not always use an *Aim Low* strategy. Instead, they may flexibly modulate the approach to audience design depending on the mixtures of common ground among addressees.

Experiment 4

Speakers design referring expressions with respect to the local context and adjust their referring expressions when the addressee's immediate context differs from their own (Nadig & Sedivy, 2002). Less clear is how this contextual sensitivity scales up to multiparty conversation, where each interlocutor may have different background knowledge and a different immediate context. Experiment 3 showed that speakers are sensitive to the combined knowledge among addressees. However, less well understood is how speakers integrate different sources of information when designing utterances, such as immediate contextual information and the addressee's background knowledge. Here I tested the same two hypotheses in Experiment 3, *Aim low* and *Combine*, but this time I manipulated the immediate visual context, rather than the addressees' background knowledge. The overall knowledge state of the two addressees was the same as in the High+Low CG condition of Experiment 2 (one knowledgeable addressee and one naïve addressee). The question is the same as in Experiment 3 – whether audience design favors addressees who lack common ground, regardless of the complexity of the immediate context (*Aim low*) or whether speakers attempt to balance the different needs of addressees with respect to both immediate contextual information and background knowledge (*Combine*).

Method

Participants

Eighty-four undergraduates (twenty-eight groups) at the University of Illinois at Urbana-Champaign participated in the experiment in return for either partial course credit or cash payment ($12). Three participants took part in the study at the same time and were randomly assigned to the roles of Director, Matcher 1, and Matcher 2.

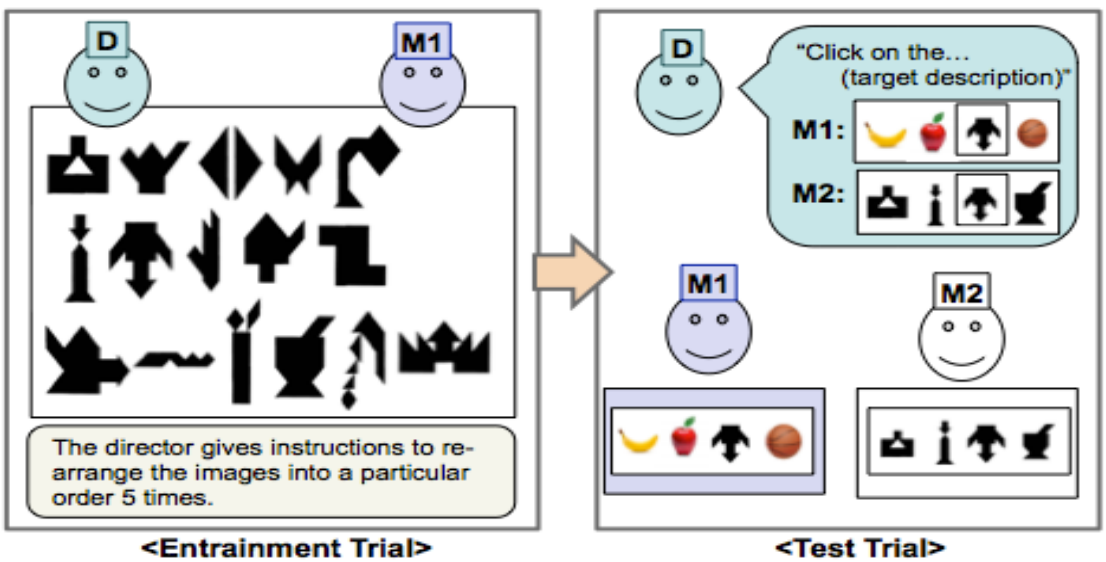


Figure 2.15. Experiment 4: Experimental procedure.

Materials and Procedure

The procedure of Experiment 4 was similar to the procedure of Experiment 2, which examined triadic conversation. Again, the task consisted of two phases: entrainment and test trials. The Director and Matcher 1 first played a tangram-sorting task as a dyad. The tangrams were black and white abstract figures from Slocum (2004). They repeated the same task five times, establishing labels for 16 abstract tangram images, while Matcher 2 was in another room. Thus, the Director and Matcher 1 shared common ground for the image labels, whereas Matcher 2 was naïve. Test trials as a triad followed the entrainment trials. At test, the Director simultaneously instructed Matcher 1 and Matcher 2 to click on one of the 4 images. Matcher 1 and Matcher 2 had separate computers that showed 4 images on each trial, and the Director had a screen that separately depicted the 4 images that each matcher had (8 total) (Figure 2.15). I manipulated the visual contexts that Matcher 1 and Matcher 2 encountered: Baseline: four old tangrams that the Director and Matcher 1 had previously labeled during entrainment. Simple: one old tangram and three new basic objects (e.g., ball). Complex: two old tangrams that the Director and Matcher 1 had previously labeled during entrainment, and two new tangrams that were similar to the old tangrams.

At test, each Matcher viewed a scene in one of these three conditions (trial order was pseudo-randomized within each of 4 blocks); see Table 2.7 (also Figure 2.16) for details of the experimental design.

In the Same-context condition (1), both Matcher 1 and 2 viewed same baseline context, including four old tangrams from entrainment. Critically, in this condition the Matchers view the same scene, but only Matcher 1 knows the names of the images (this is similar to the design of Experiment 2).

In the Simple/baseline condition (2), Matcher 1 had the simple context and Matcher 2 had the baseline context. This condition is similar to the same-context condition in that Matcher 2 is presented with an array of four novel tangrams that they do not know the names of.

In the Baseline/simple condition (3), Matcher 2 had the simple context and Matcher 1 had the baseline context. In this condition, Matcher 2's task is significantly easier because there is only 1 tangram in their array.

Finally, in the Complex/simple condition (4), Matcher 1 had the complex context and Matcher 2 had the simple context.

There were 16 critical target trials and 8 filler trials in each of four experimental blocks. During filler trials, the two Matchers had the same simple context for half of trials and the same complex context for the other half of trials. The target on critical trials was always one of the old tangrams, while the target on filler trials was either one of the basic objects (e.g., the banana) or one of new tangrams that were similar to the old tangrams but that had not been seen during entrainment. The condition manipulation was within subjects, and each group of participants completed four different entrainment-testing blocks (a total of 64 target trials and 32 filler trials). Within each block the order of trials was pseudorandomized. The order of blocks was counterbalanced across groups of participants. Four lists were constructed to rotate the items through the four conditions (each participant only saw the items on a single list).

Table 2.7. Critical manipulations of visual context during test trials.

Condition (within-subjects)	M1's context	M2's context
(1) Same-context	Baseline	Baseline
(2) M1-simple / M2-baseline	Simple	Baseline
(3) M1-baseline / M2-simple	Baseline	Simple
(4) M1-complex / M2-simple	Complex	Simple

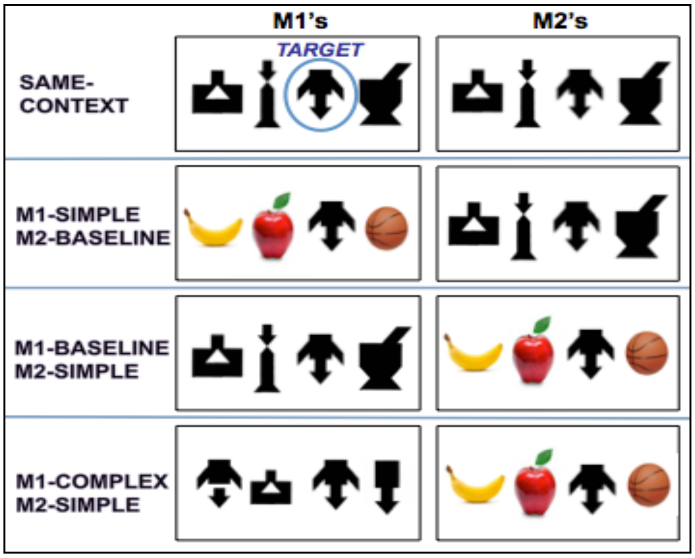


Figure 2. 16. Experiment 4: Critical manipulations of visual contexts that Matcher 1 and Matcher 2 encountered during test trials. The order of conditions was counterbalanced across groups.

Predictions

The results of Experiment 2 showed that speakers *Aimed Low* when they addressed one knowledgeable and one naïve addressee, both of whom had the same visual context. The primary manipulation of Experiment 4 was to vary the immediate contexts that two addressees encountered, while maintaining their overall knowledge state – one knowledgeable and one naïve addressee –across conditions. If speakers *Aim Low* as in Experiment 2, expression length should reflect Matcher 2's naiveté and context. On the *Aim Low* view, speakers should produce longer expressions in the (1) Same-context and (2) Simple/Baseline conditions than the (3) Baseline/Simple and (4) Complex/ Simple conditions, because it is in conditions (1) and (2) that the naïve Matcher 2 has the difficult task of distinguishing between four tangrams they have never seen before. By contrast, in conditions (3) and (4) the naïve addressee can easily identify the single tangram because it is the only one of its kind.

In contrast, if speakers *Combine*, speakers should show sensitivity to both Matcher 1 and Matcher 2's perspective. On a *Combine* view (like *Aim low* predicts), Directors should produce longer expressions in conditions (1) and (2) than condition (3), due to Matcher 2's lack of knowledge about old tangrams. In addition, the *Combine* view predicts that when Matcher 1 has a complex display (4), speakers should produce longer expressions to help Matcher 1's understanding, compared to the condition (3) Baseline/Simple.

Results

As in the previous experiments, expression length in the entrainment and test trials, and the rates of disfluency during the test trials were analyzed (see Appendix A for full models).

Entrainment Trials

During entrainment, the Directors produced shortened expressions across trials, consistent with previous findings that interlocutors gradually established common ground (z=-17.32, p<.05).

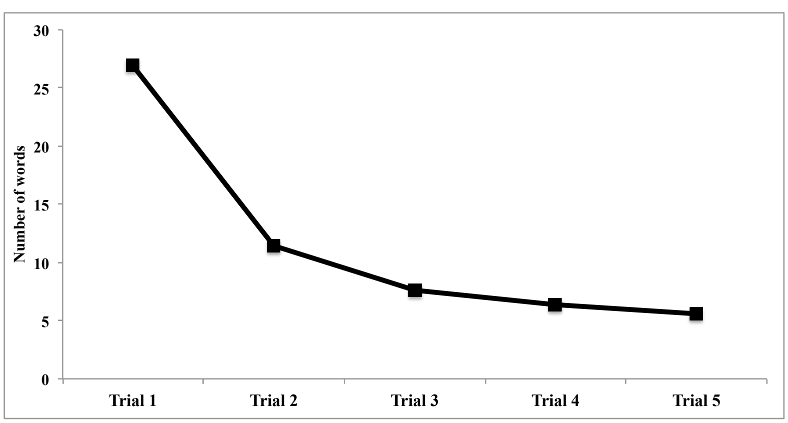


Figure 2.17. Experiment 4: The average number of words used to describe each image across the five trials of entrainment.

Test Trials – Expression Length

Directors produced the shortest expressions in the (3) Baseline/Simple condition and the longest expressions in the (4) Complex/Simple condition (Figure 2.18). Expression length in the other two conditions, (1) Same-context and (2) Simple/Baseline, were in-between.

Expression length at test was analyzed with maximal mixed-effects models. Directors produced similar-length expressions in the (1) Same-context and (2) Simple/Baseline conditions (z=1.33, p>.05); the lack of a significant difference between these conditions suggests that Directors prioritized Matcher 2's lack of common ground for the image labels. Expression length was significantly shorter in the (3) Baseline/Simple condition than in the (2) Simple/Baseline and (1) Same-context conditions (z=4.67, p<.05), indicating that Directors were sensitive to naïve Matcher 2's needs. Directors produced the longest expressions in the (4) Complex/Simple condition, compared to the other three conditions (z=-8.44, p<.05), suggesting that they considered Matcher 1's local context when designing expressions. Even though Matcher 1 was more knowledgeable than Matcher 2, Directors adjusted their referential expressions to support Matcher 1's needs. Thus, the findings suggest that Directors were sensitive to both Matcher 1 and Matcher 2's distinct perspectives and flexibly designed utterances accordingly.

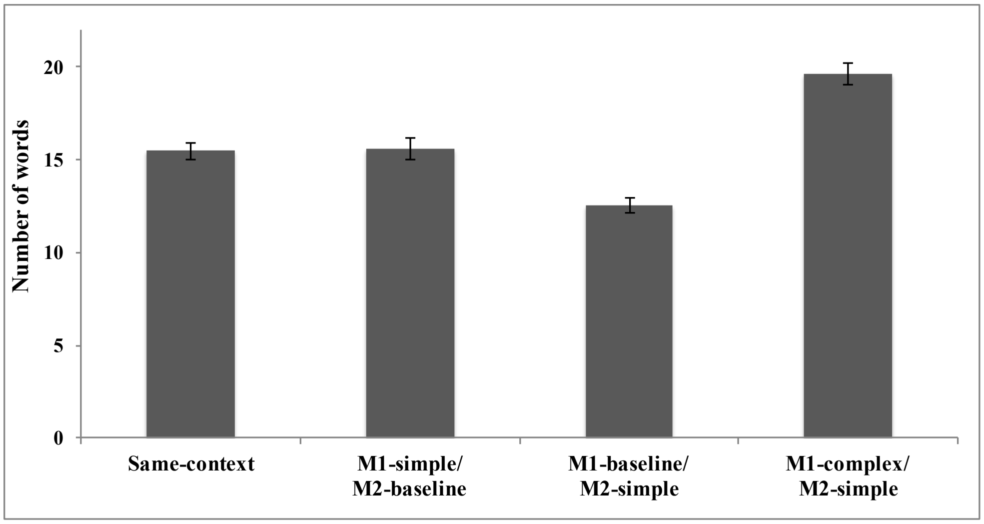


Figure 2.18. Experiment 4: The average number of words used to describe each image on test trials. Error bars represent the standard error of the by-participant mean.

Test Trials – Disfluency

Disfluency was coded in the same way as in the previous experiments. Overall, the rate of disfluency was high across the four conditions (Table 2.8).

I analyzed a binary measure of disfluency in a mixed-effect model, including condition as a fixed effect. Similar to the analysis of expression length, the disfluency rate was significantly higher in the (4) Complex/Simple condition compared to the other three conditions (z=-7.08, p<.05), and the disfluency rate was lower in the (3) Baseline/Simple than the (1) Same-context and (2) Simple/Baseline conditions (z=2.19, p<.05). The rate of disfluency in the (1) Same-context and (2) Simple/Baseline conditions did not significantly differ (z=0.31, p>.05).

Table 2.8. Experiment 4: Percentage of disfluent expressions on test trials.

	Same-context	M1-simple/ M2-baseline	M1-baseline/ M2-simple	M1-complex/ M2-simple
Disfluency	62.76%	62.62%	57.14%	83.80%

Summary and Discussion

In Experiment 4, Directors established common ground for a series of abstract image labels with one Matcher, and then subsequently engaged in a three-party conversation with this knowledgeable Matcher and a new, naïve matcher. The analysis of utterance form during this three-party conversation revealed that Directors were sensitive to both the knowledgeable and the naïve partner's needs based on their local contexts and each partner's background knowledge. When the task was difficult for the naïve Matcher, Directors produced long expressions. When the task was difficult for the knowledgeable Matcher, due to the introduction of novel distractors, Directors also produced long expressions. By contrast, when the task was easy for both Matchers, Directors produced short, efficient expressions.

Note that our manipulation of the local contexts was intermixed across test trials. Thus, to succeed, Directors had to promptly assess both partners' contexts and knowledge states, and design an appropriate referring expressions. The fact that our Directors were successful at assessing both the scene and the knowledge state of the addressees shows that these processes are quite rapid.

The finding that Directors took the perspective of both the naïve and knowledgeable partners into consideration is contrary to the prediction of *Aim Low* hypothesis. While my previous work suggested that in 3-party conversation that speakers may adopt an *Aim Low* approach (Yoon & Brown-Schmidt, 2014; Experiment 1-3), the results of this study, along with the results of Experiments 3a-3b show that speakers take the needs of both naive and knowledgeable addressees into consideration. When a knowledgeable partner encountered a complex context (4), the Complex/simple condition, Directors produced longer expressions despite the fact that the naïve partner had a very easy task (i.e., to identify a single tangram in a set of everyday objects). Taken together, the findings suggest that speakers represent the distinct perspectives of multiple addressees, and tailor expressions for all to understand with respect to each addressee's immediate context and their background knowledge in multiparty conversation.

General Discussion

Representation of Perspective

How do speakers represent common ground in multiparty conversation? If speakers cannot hold onto two distinct perspective representations at the same time, they should be unable to rapidly alternate between individual addressees in multiparty conversation. Similarly, if the representations of common ground associated with individual partners were to become blended and the distinct representations of the perspectives of these individuals lost (Figure 1.1), it would predict that speakers should have difficulty designing an utterance for one partner at a time, and that expressions designed for multiple addresses simultaneously should have similar properties to expressions designed for one of them. One reason to think speakers might fail to maintain distinct representations comes from findings that speakers experience difficulty distinguishing between different conversation partners when the partners share similar background knowledge (Horton & Gerrig, 2005). Further, egocentric biases in assessing others' knowledge (Fussell & Krauss, 1991, 1992), may make it difficult for speakers to distinguish their own knowledge from that of the addressees.

The results of Experiments 1-2 show that speakers can, in fact, actively maintain multiple representations of common ground held with different people. These representations are distinct from each other, and from one's own egocentric knowledge. Even in situations where the non-addressee was present at test, speakers designed longer expressions when addressing a low-knowledge partner, and shorter expressions when addressing a high-knowledge partner. Speakers did not discard the representations of the perspective of the individual addressees.

What form do these perspective representations take? On one view, audience design can be supported by a simple, one-bit model of the partner's knowledge, for example a representation that a label for an abstract image "is" or "is not" part of the discourse record (Galati & Brennan, 2010). This proposal is appealing because it provides a simple mechanism for encoding perspective that would presumably pose few resource demands. While a one-bit partner model may have explanatory power in other situations, it cannot account for the full pattern of findings here. Critically, in the non-recent conditions of Experiment 1, speakers designed expressions differently for addressees who had different degrees of experience with the image labels. This finding cannot be explained by a one-bit model because these image labels were part of the discourse record for both addressees—critically, the addressees varied in their *degree* of experience with those labels. Instead, our findings are consistent with the proposal that conversational partners represent gradient information about the degree to which a given piece of information is common ground with a specific discourse partner (Brown-Schmidt, 2012). The more firmly an object and its description were established as common ground, the shorter the expression.

An open question is how exactly these representations are encoded in memory, and if, in fact, they are encoded directly. One possibility is that the memorial record includes a separate representation of the common ground status of a given entity, e.g., a representation that encodes the fact that the term "*the 3D chair*" is common ground with a specific addressee, with these representations of common ground status per se guiding audience design. Alternatively, the common ground status of individual entities may not be stored per se, and instead, audience design might make use of memorial records of joint experience that are re-activated in the moment (see Horton & Gerrig, 2005). Regardless of whether they are represented per se or generated on-the-fly on the basis of memory access, our findings point to these representations as being partner-specific, gradient, and flexibly deployed in conversational settings. These representations (however they are encoded) likely draw on multiple memory systems including both declarative (e.g., episodic) and non-declarative (e.g., priming, procedural learning) memory mechanisms (Duff & Brown-Schmidt, 2012).

Together with other findings from the literature, the emerging picture, then, is that conversational partners maintain multiple representations of common ground held with different individuals, with these representations being partner-specific (Metzing & Brennan, 2003), gradient (Brown-Schmidt, 2012), and tailored on an item-by-item basis (Heller, et al., 2012; Gorman, et al., 2013).

Mechanisms of Audience Design

A separate set of questions concern how representations of common ground are used in conversation, and what cognitive processes support audience design.

When I compared the form of utterances designed for a single addressee as a function of whether the other partner was in the room, I found that the presence of the non-addressee did not significantly affect any aspect of referring I measured. This finding suggests that even if the common ground shared with that person was automatically activated (Horton, 2007), that utterance design was guided by those representations shared with the intended addressee. An open question is whether the presence of a non-addressee might influence other aspects of referring, and if so, in what circumstances this might be observed in (see Brown-Schmidt & Horton, 2014). One possibility is that representations associated with non-addressees may be more likely to guide language use in situations where the non-addressee is otherwise highly integral to the information at hand (see Eich, 1985). Another outstanding issue is the degree to which strategic choices are in play. To be sure, speakers can and do engage in strategic audience design, as in the case when actively deceiving someone. In the situations examined in the present experiments, however, audience design need not involve strategic choices. Instead, the attentional and memory mechanisms that support identification of the intended addressee, and retrieval of the mutual knowledge associated with them, could have produced the effects I saw here.

How do speakers approach the problem of audience design in multiparty conversation? Here I proposed, and evaluated three candidate hypotheses regarding the way in which speakers might approach the problem of audience design in 3-party conversation (Experiment 2): *Aim low* (design expressions for the most naïve addressee), *Aim high* (design expressions for the most knowledgeable addressee), and *Combining* (combine the knowledge states of all addressees). Previous findings by Yoon and Brown-Schmidt (2014) were equally consistent with the *Aim low* and the *Combine* accounts. By testing additional experimental conditions, Experiment 2 showed that in two-party conversation, speakers designed utterances for the naïve party. Utterances were equally long when addressing two addressees (one knowledgeable, one naïve) as when addressing one naïve addressee, consistent with the *Aim low* hypothesis. There was no evidence to support an *Aim high* approach, despite the fact that one of the addressees had the same relevant knowledge as the speaker.

Experiments 3-4 again examined whether speakers pursue an *Aim Low* approach in multiparty conversation, this time by manipulating addressees' knowledge states in a group or in their immediate conversational contexts. As combined knowledge states decreased in 4- and 5-party conversation, expression length and rate of disfluency gradually increased, supporting the *Combine* hypothesis. Further, speakers promptly updated representations for both partners' immediate contexts, and adjusted expressions depending on their immediate context and knowledge states. For instance, in Experiment 4, when the naïve addressee had an easy context within which to identify the target, speakers did not *Aim Low*, but produced fairly short expressions. In the same experiment, when the knowledgeable addressee had a complex context, speakers added extra details to help the knowledgeable partner identify the referent. These findings are consistent with a view in which audience design in multiparty conversation is based on a combination of the perspectives of the addressees.

How do speakers combine the addressees' perspectives? One possibility is that they calculate a mathematical average of knowledge states (while simultaneously maintaining the distinct representations of each addressee's perspective; see discussion above). In this view, speakers equally consider the distinct needs of knowledgeable and naïve addressees – efficiency and informativity. If we make the simplifying assumption that the naïve addressee has 0% of the relevant knowledge, and the knowledgeable addressee has 100% of the relevant knowledge, in the 3K1N, 2K2N, and 1K3N conditions, the overall mathematical average state is 75%, 50%, and 25% respectively. If audience design is based on the mathematical average knowledge state, in our task this would predict increased expression length as the average knowledge decreased.

Another possibility is that speakers maintain distinct representations of the knowledge of each addressee, and then identify the most dominant knowledge state. This view predicts that speakers will prioritize either informativity or efficiency, depending on whether the dominant knowledge state is naïve or knowledgeable, respectively. Evidence consistent with this idea was the finding that when more than half of addressees were naïve (e.g., 1K2N or 1K3N), audience design emphasized informativity more than efficiency. That is, in Experiments 3-4, expression length approached an asymptote as the combined knowledge state decreased; this finding could have resulted from a mechanism in which speakers design with respect to the dominant knowledge state.

It is not clear, however, how a dominance mechanism would account for the findings in balanced groups. When the group knowledge was perfectly balanced, such as the 1K1N (Experiment 2) and the 2K2N condition in (Experiment 3b), speakers designed equally long expressions as when they talked to a naïve addressee. Taken together, the results of Experiments 2-4 support a view of audience design in which speakers take into account the knowledge states of all of the addressees (at least up to 4 addressees), and combine them in some way, possibly by averaging, possibly by identifying the dominant knowledge state.

One thing to note in Experiment 3 is that the Director's experience of entrainment was not consistent across conditions. For some conditions, such as the 2K1N condition in

Experiment 3a and the 3K1N condition in Experiment 3b, Directors performed multiple rounds of entrainment, while they had only one round of entrainment in other conditions (e.g., 1K or 1N conditions). Thus in our design, when more addressees knew the object labels, the speaker also had more egocentric knowledge. We opted for this design choice (where the Director sequentially establishes common ground with each knowledgeable Matcher separately) in order to keep the process of entrainment for a given Matcher consistent across conditions (always involving two people entraining on names). Whether this confounding role of the Director's egocentric experience played a role in our findings during test trials remains to be explored. An important point to note however, is the fact that speakers produced equally long expressions for the first, second, and third Matchers during entrainment, regardless of their own egocentric knowledge. This finding suggests that the speaker's egocentric knowledge likely had little-to-no influence on how speakers designed expressions during test.

Another outstanding question is how group size affects the process of audience design. In Experiment 3, the number of partners in a group during test trials was not fully controlled: in the 1K and 1N condition, the Director addressed only one partner, whereas in the other conditions, the Director interacted with 3 or 4 partners. It remains an open question as to whether speakers design utterances differently when they talk to one or multiple partners with equivalent knowledge states (e.g., 1K vs. 3K (or 4K); 1N vs. 3N (or 4N)). A related question not addressed by the present research is what factors influence multiparty audience design. The answer to this question likely involves considerations of the relevant memory and attentional demands on the audience design process. For example, when addressing large groups, a summary extraction process may generate representations of clusters of individuals, analogous to high-level visual processes such as implicit detection of the average emotion in a set of faces (Haberman & Whitney, 2010). Communicative goals may focus attention on some addressees more than others, influencing the process of audience design (see Yoon, Koh, & Brown-Schmidt, 2012). When communication is *rapid*, faster activation of the mutual knowledge held with more knowledgeable addressees may bias the design process towards an *Aim High* approach. Fully understanding the malleability of the audience design process will likely involve comparisons of more complex communicative situations, with multiple potential addressees, and multiple types of communicative goals.

The last unanswered question concerns the time-course of these processes, and at what point sensitivity to the perspective of the addressee(s) comes into play. While we find clear evidence for sensitivity to the addressee(s) perspective in the speaker's initiating description of each object, an open question is when during the production process speakers access entrained terms, and plan additional descriptive material.

CHAPTER 3: MEMORY FOR THE DISCOURSE HISTORY

Lexical differentiation refers to a discourse phenomenon in which speakers take into account past reference when designing new referring expressions, differentiating two sequentially presented objects from the same category (Van der Wege, 2009). The goal of the present research is to examine the locus of the lexical differentiation effect and its relationship with memory for past discourse referents. In Experiments 5 and 6, I elicit a differentiation effect in language production, and examine the situations in which it does and does not occur in order to understand the influence of the historical discourse context on referring. In Experiment 7, we examine the same question in situations that include an unmentioned, but target-related context item. Measures of memory for the discourse history in Experiments 5-7 are used to test whether poor memory for past referents explains the low incidence of differentiation, and listeners' consequent lack of expectation for it (Yoon & Brown-Schmidt, 2013).

Experiment 5

According to Van Der Wege (2009), speakers lexically differentiate to prevent two different entities from having the same label, a process called "pre-emption by similar form" (also see Clark & Clark, 1979). For example, the speaker avoids giving the label, "*the shirt"*, to two different entities. By calling the second shirt "*the striped shirt"* or *"the blouse",* the speaker is able to differentiate the two labels. The idea behind pre-emption by similar form is that the previously used label "*the shirt*" pre-empts the subsequent use of the same label to refer to a different item, thus creating the need for lexical differentiation. If this view of the lexical differentiation effect is correct, differentiation should not be observed if the speaker had not labeled the first entity with the basic object label. For example, if the speaker were to refer to the first shirt with a locative expression, such as "*the top left one*", there would be no pre-emption of the label "*the shirt*", and thus no need to describe the second shirt with a modifier (instead "*the shirt"* would be an appropriate label for the second shirt). Alternatively, the locus of the differentiation effect might be an attempt to distinguish current from past *referents*, regardless of how they had been named. If so, any previous reference to a shirt—with a locative or a descriptive expression—should increase the likelihood that the speaker would differentiate the second shirt from the first.

In the current experiment, I manipulate the way in which the first referent was referenced—with a locative, "*the top left one",* or a descriptive phrase, *"the shirt"*—in order to test the locus of the lexical differentiation. In addition, I measure speakers' and listeners' memory for the discourse history.

Method

Participants

Seventy-two undergraduates (thirty-six pairs) at the University of Illinois at Urbana-Champaign participated in the experiment in return for partial course credit or cash payment ($8). Participants were native speakers of North American English with normal hearing and normal or corrected-to-normal vision.

Procedure

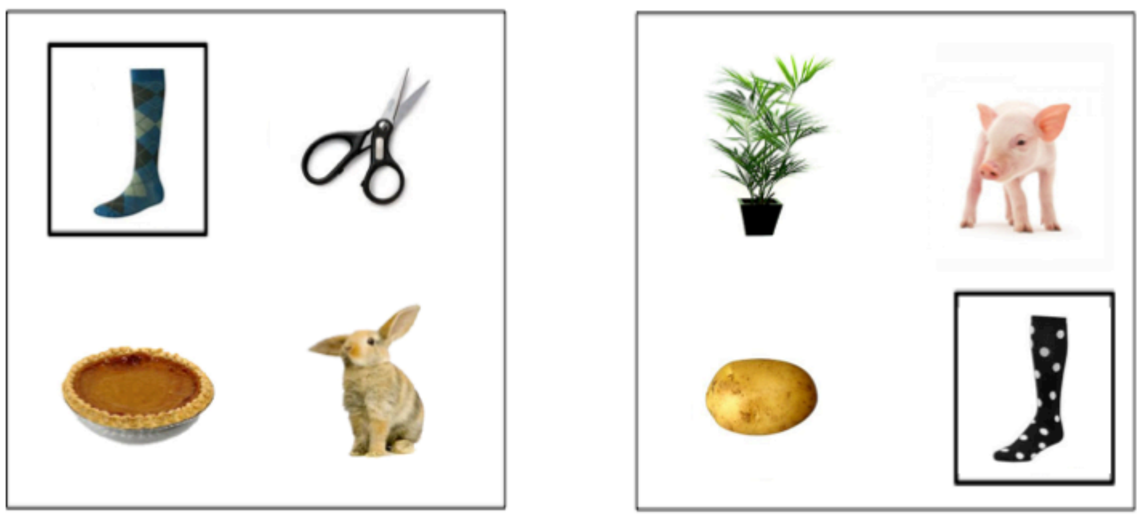
The experiment consisted of two phases: a referential communication task followed by a memory test. The entire experiment lasted approximately 50 minutes.

Referential Communication Task

During the referential communication task (Krauss & Weinheimer, 1966), two naïve participants were randomly assigned to the roles of speaking and listening and sat at separate computers in the same room. The computers were arranged so that participants could see each other's faces but could not see each other's computer screens. On each trial, the speaker and listener viewed a computer screen with four pictures. The pictures appeared in one of four different random positions on each screen. The target object was indicated to the speaker with a black box (see Figure 3.1). On the listener's screen, the target was not indicated. The speaker's task was to give the listener an instruction aloud to click on one of the four pictures (e.g., *Click on the sock*). The listener followed the speaker's instructions on her own screen. The speaker was allowed to describe the target object in any way they saw fit, except for locative phrases (e.g., *top right one*), which speakers were instructed not to use. The task was interactive and participants were allowed to ask questions for clarification as needed. After the listener clicked the target, the speaker clicked the mouse once to advance to the next trial. Participants maintained their assigned role (i.e., as speaker or listener) throughout the task. Recordings of the speaker's voice were saved directly to the computer.

Materials

During the referential communication task, each participant completed a total of 462 trials, including 198 <u>entrainment</u> trials, 33 <u>test</u> trials, and 231 <u>filler</u> trials. During the test trials (the focus of our analyses), the speaker described a "target" object from a particular category (e.g., a dotted sock). Each test trial was associated with 6 entrainment trials (see Figures 3.1a-1b). The type of entrainment trial served as the key experimental manipulation, and was manipulated within-subjects. These entrainment trials always preceded test trials, and allowed us to manipulate whether the basic object label for a given test object, e.g., "sock" had already been used to describe a different item during entrainment. During both entrainment and test trials, the target object was presented with three other unrelated objects such that given the *local* context, a bare noun was sufficient to identify the target. These unrelated objects were rotated pseudo-randomly such that the three unrelated items varied from trial to trial. The critical question, then, was whether speakers would design referring expressions at test based on the *historical* discourse context.



(a) (b)

Figures 3.1a-b. Experiments 5 and 6: Example stimuli from the referential communication task, including entrainment trials (1a) and test trials (1b). The target is indicated to the speaker by the black rectangle. The addressee's screen would show the same four pictures, but in different locations and without the black box. This example shows the "contrast" condition; in the "no-contrast" condition, the target during entrainment would be an unrelated item such as an apple.

In the Contrast-naming condition, the target on entrainment trials was an exemplar from the same category as the target on the test trial. For example, the speaker would describe an argyle sock 6 times during entrainment trials prior to describing the target dotted sock at test. During the entrainment trials I expected speakers to use the basic object label, e.g., "sock" to describe the target (argyle sock). If speakers are sensitive to the historical discourse context, then during the associated test trial, they should differentiate their referring expression and describe the target with a modifier (e.g., *the dotted sock*).

In the Non-contrast condition, participants described an unrelated object (e.g., apple) 6 times in entrainment, and then described the target object (e.g., dotted sock) at test. Note that for items in the Non-contrast condition, speakers never saw the contrast object.

Lastly, in the Contrast-location condition, speakers described the location of the contrast object (e.g., *top left one*) 6 times, and the target object at test, using a (modified) noun phrase. The speaker was cued to use a locative phrase through an on-screen text prompt (e.g., "LOCATION"). This condition was included in order to test whether it is naming *per se* that is critical to eliciting the differentiation effect.

Following entrainment, test displays showed four new (previously unseen) objects, including the target and three unrelated objects. While the target and the unrelated objects were all new tokens, the categories that they came from (e.g., socks, pigs, etc.) had all been previously experienced during the entrainment trials.

Filler trials contained two contrasting objects from the same category (e.g., two fish) and two unrelated objects. Half of the time the target was one of the two contrasting objects and on the other half of filler trials, the target was one of the two unrelated objects. Filler trials with contrasting objects were included so that speakers would sometimes need to produce modified noun phrases based on the immediate context.

Entrainment and test trials for the different item sets were interspersed with filler trials in a pseudo-random order such that from the participants' point of view, there was no distinction between the trial types, and so that all 7 trials associated with a single item set (6 entrainment trials followed by 1 test trial) occurred within a span of at most 120 trials (ranged from 73 to 116, M=100.37, SD=12.4). Note that for a given object set, the maximum time between the final entrainment trial and the test trial was at most 15 trials (range of 1 to 15, M=7.39, SD=4.46).

All visual stimuli were pictures of every-day objects. The stimuli included the 33 triplets of target (e.g., dotted sock), contrast (e.g., argyle sock), and unrelated items (e.g., apple) that were used as critical items across the three conditions. The remaining visual stimuli were used on filler trials and included 33 triplets of contrasting objects (e.g., opened box, wrapped box, stacked box) and 66 pairs of contrasting objects (e.g., sitting dog, jumping dog) from the same categories. All target items were counterbalanced across conditions across three lists. Each participant completed the items on one list.

Memory test

Following the communication task, participants performed a surprise recognition memory test on their own computer (Figure 3.2). Each trial presented two pictures from the same category: One was an "old" picture that was presented during the communication task, while the other was a "new" picture from the same basic object category. Participants were asked to click the old picture on each trial. The old and new pictures were rotated across two lists, so that the old picture on one list was the new picture on the other list. A total of 99 trials tested memory for the 33 contrast items (e.g., argyle sock or apple), 33 target items (e.g., dotted sock), and 33 unrelated filler items (e.g., drum). The correct answer for each trial was always the object that had been named or referenced with a locative phrase during the referential communication task. The speaker and listener were not allowed to talk to each other during this phase of the experiment.

Figure 3.2. Experiments 5 and 6: Example stimuli from the memory test.

Predictions

The communication task was designed to replicate the lexical differentiation effect in language production (Van der Wege, 2009; Yoon & Brown-Schmidt, 2013). If speakers lexically differentiate, the modification rate (e.g., production of a modified expression like *dotted sock* vs. an unmodified expression like *sock*) should be higher in the Contrast-naming condition than in the Non-contrast condition. The design of this experiment also allows us to address the nature of the lexical differentiation effect per se. According to the process "pre-emption by similar form" suggested by Van Der Wege (2009), speakers produce lexically differentiated labels, so that they do not assign the same label to two different objects. If this account of the lexical differentiation effect is correct, speakers should be more likely to modify their referring expressions in the Contrast-naming condition compared to the Contrast-location conditions because it is only in the Contrast-naming condition that the basic object label for the target had previously been used (e.g., “sock” had been used on entrainment trials for a different referent). On the other hand, if mere exposure to two items from the same basic object category, rather than previous *naming* per se, is sufficient to elicit lexical differentiation, the rate of modifiers on test trials should be higher in both the Contrast-naming and Contrast-location conditions compared to the Non-contrast condition.

The relationship between performance in the memory task and the referential communication task will allow us to address two questions about the nature of the memory representations that support language use in dialogue. The <u>first</u> question is why the differentiation rate in previous studies was so low. For example, speakers differentiated only 7.5% of the time in Yoon and Brown-Schmidt (2013) and 26% in Van der Wege (2009). One explanation is that successful differentiation requires memory for the contrast object, and that low rates of differentiation result from poor memory for the contrast. If so, I would predict that successful recognition of contrast objects will predict the higher differentiation rate. Such a finding would point to memory as the critical factor determining whether speakers will design expressions with respect to the historical discourse record.

Our <u>second</u> question concerns potential asymmetries in the representation of the discourse record between speakers and listeners. Recall that Yoon and Brown-Schmidt (2013) observed that whereas speakers sometimes differentiated current from past referents, there was no evidence that listeners ever anticipated differentiation. Here I test if this asymmetry is due to speakers and listeners forming different memories of the discourse history. If speakers are more sensitive to the discourse history than listeners, this would explain the presence of a lexical differentiation effect for speakers but not listeners. Such a finding would be expected if the generation effect in memory research (Marsh, et al., 2001) extends to more natural conversational settings. By contrast, if conversational partners align at every level of representation (Pickering & Garrod, 2004), speakers and listeners should exhibit similar performance on the memory test.

Results

Referential communication task

Speakers' recorded object descriptions were transcribed and coded for whether their referring expressions were modified or not (Figure 3.3). I analyzed the data in a logistic mixed effects model with entrainment type (Non-contrast, Contrast-location, and Contrast-naming) as a fixed effect, and subjects and items as random intercepts. The dependent variable was whether the speaker used a modifier or not. Full model details are presented in Appendix B.

While bare noun phrases were common in all three conditions, speakers differentiated locally unique referents with modified noun phrases significantly more often when the contrast object had previously been discussed (21.9%, Contrast-naming vs. 20.8% and Contrast-location condition) than when it had not (14.1%, Non-contrast condition) (z=2.194, p<.01). The rate of modified noun phrases between the two contrast conditions was not significantly different (z=0.067, p>.05). The size of the differentiation effect was 7.8%[7], similar to the previous findings in Yoon & Brown-Schmidt (2013).

---

[7] One question about this finding is whether the lag between the last entrainment trial and the test trial modulates the differentiation effect (the lag varied between 1 and 15 trials). If speakers fail to differentiate after long lags, this would be consistent with the memory-based explanation for why speakers differentiate so infrequently. The results of this post-hoc analysis, however, showed that neither the main effect of lag (z=-0.826, p>.05), nor the interactions with condition (Non-contrast vs. Contrast-location & Contrast-naming: z=-0.301, p>.05; Contrast-location vs. Contrast-naming: z=0.994, p>.05) were significant.

In summary, these findings replicate the lexical differentiation effect (Van der Wege, 2009; Yoon & Brown-Schmidt, 2013), demonstrating that speakers design expressions with respect to the historical discourse record (Brennan & Clark, 1996). Critically, however, there was not a significant difference in modification rates between the Contrast-location and Contrast-naming conditions. This finding suggests that speakers differentiate not to avoid lexical conflict (cf., Van der Wege, 2009). Instead, speakers differentiate in order to distinguish current referents from past *referents*, regardless of how those referents had been labeled.

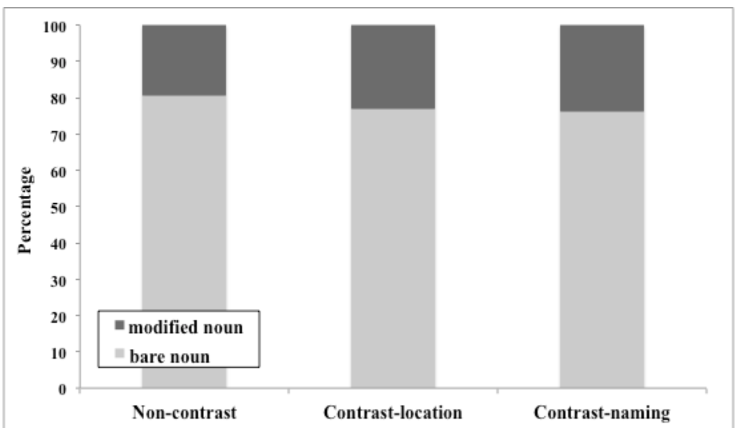


Figure 3.3. Percentage of each noun phrase type on target trials during communication task in Experiment 5.

Memory test

I test memory for the contrast to address the question whether the low rate of lexical differentiation is due to memory failures. I also compare memory performance by speakers and listeners to evaluate whether they developed distinct memory representations of the discourse. Overall, both speakers and listeners successfully recognized contrast and target items over 80% of the time (Figure 3.4), suggesting the low differentiation rates for target item in the communication task were not due to poor memory for past referents.

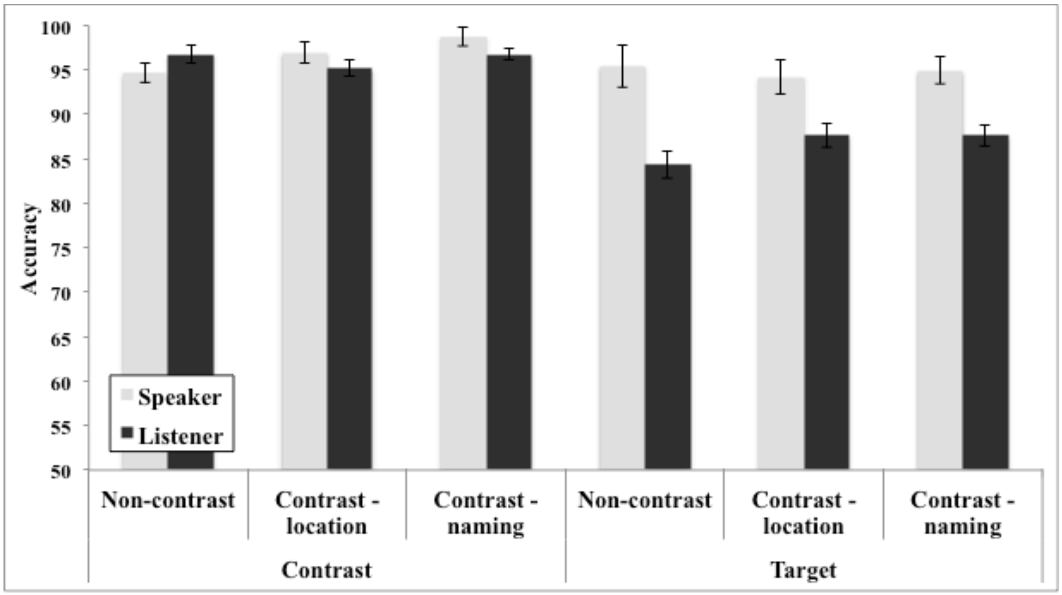


Figure 3.4. Experiment 5: Accuracy on two-alternative forced choice recognition memory task for contrast items (left bars) and target items (right bars) by speakers and listeners.

Memory accuracy data were analyzed in a logistic mixed effects model with role (listener vs. speaker), referent (contrast vs. target), and entrainment type (Non-contrast vs. Contrast-location vs. Contrast-naming) as fixed effects (see Appendix B for model details). Overall, speakers had better memory than listeners (z=2.781, p<.05). Memory was also better for contrast than target items, which is unsurprising, as contrast items were referenced 6 times during the communication task, whereas target items were referenced only once (z=-5.173, p<.05).

Planned comparisons explored a significant interaction between role and referent type (z=3.099, p<.05) by examining the role effect for contrast and target items separately. Memory for contrasts was high for both speakers and listeners (>95%) (z=-0.136, p>.05), likely reflecting a ceiling effect. By contrast, speakers remembered target items significantly better than listeners (94.9% vs. 86.5%; z=5.616, p<.05). These findings suggest that the gap between speakers' and listeners' memory for previous referents attenuates with repeated exposure.

Relationship between production and comprehension

The fact that recognition of contrast items was uniformly high is inconsistent with the idea that speakers fail to differentiate when memory traces for the contrast object were weak. I tested whether speakers' memory for the contrast predicted the differentiation rate and the results of that analysis revealed the differentiation rate was not significantly influenced by contrast memory ($z=-0.426$, $p>.05$).

While there was no evidence that *contrast memory* modulated use of a modifier on test trials, an open question is whether use a modifier on test trials improves memory for the target. One reason to think that modification might improve memory for the referent is that the process of selecting an appropriate modifier necessarily requires more elaborate encoding (Bradshaw & Anderson, 1982) of that object (e.g., saying *striped shirt* requires conceptualizing of the shirt as striped). However, the results of this post-hoc analysis showed that when speakers described target objects with modifiers, memory for targets did not improve for speakers or listeners ($z=0.027$, $p>.05$).

Summary and Discussion

In summary, the results of Experiment 5 showed that speakers were more likely to use modifiers when they had previously referenced an item from the same basic object category, replicating previous findings (Van der Wege, 2009; Yoon & Brown-Schmidt, 2013). However, naming was not necessary to elicit lexical differentiation. This finding is inconsistent with the argument that differentiation is motivated by pre-emption by a similar referential form (Van der Wege, 2009). Instead, speakers differentiate in order to distinguish the current referent from past referents—regardless of how those past referents had been labeled.

Recognition memory for past referents was high, consistent with findings in the memory literature that recognition memory for pictures tends to be quite good (Shepard, 1967). I also observed that speakers had better memory for target objects; whether memory for contrast objects follows a similar pattern was obfuscated by near-ceiling performance. The lack of a relationship between memory for the contrast item and the differentiation rate suggests that failures to link the current context with the past context are not due to failures to remember that past context.

Experiment 6

The aim of Experiment 6 was to replicate Experiment 5 with a lower number of entrainment trials in order to equate exposure to target and contrast items. In addition, participants in Experiment 6 alternated roles of speaking and listening to address the possibility that lower recognition rates on the part of listeners was due to poor task engagement. While the speaker-memory advantage in Experiment 5 is consistent with a generation effect, speakers also had a more active role in the conversation. The required commitment to both roles in a conversation in Experiment 6 should decrease the possibility of differences in engagement.

Method

Participants

Ninety-six undergraduates (forty-eight pairs) at the University of Illinois at Urbana-Champaign participated in return for partial course credit or cash ($8).

Participants were native speakers of North American English. None had participated in Experiment 5.

Materials and Procedure

The procedure of Experiment 6 was identical to Experiment 5 with the following exceptions: First, during the communication task, the two participants first completed 20 practice trials (2 blocks of 10 trials each), followed by two blocks of trials in the communication task. Participants alternated between the roles of speaking and listening in each block. During the practice trials, they were introduced to both roles of speaking and listening to get accustomed to the task. Second, I included a 20-minute break between the communication task and the memory test in an attempt to avoid a ceiling effect on the memory test.

The design of the communication task was also slightly different. Each participant completed a total of 200 trials, including 20 practice, 36 entrainment, 36 test, and 108 filler trials. Note that unlike Experiment 5, in Experiment 6 each target item was associated with a single entrainment trial in order to equate exposure to contrast and target objects prior to the memory test. The experimental conditions were the same as in Experiment 5 (Non-contrast, Contrast-location, Contrast-naming conditions), and were manipulated within-subjects.[8]

---

[8] I also manipulated the lag between the entrainment and the test trial (1 vs. 10 trials) to examine whether the differentiation effect might be more pronounced at shorter lags. In the 1-lag condition, the entrainment trial for a given item (e.g., the argyle sock) was immediately followed by the test trial for that item (e.g., the dotted sock). By contrast, in the 10-lag condition, the test trial occurred 10 trials after the entrainment trial. We expected that if lexical differentiation is sensitive to the recency of exposure to contrast item, speakers may be more likely to differentiate when the lag between the contrast and target is short. However, neither the main effect of lag (z=0.814, p>.05), nor the interactions between lag and entrainment type were significant (ps>.05), suggesting that the low rates of differentiation (~4%) were not simply due to long lags between

After the referential communication task, there was a 20-minute break during which participants completed an unrelated distractor test (quantitative math test). Following the break, participants completed the unexpected memory test. As in Experiment 5, each test trial on the memory test showed two pictures from the same category, one of which was "old" and one of which was "new". The task was to click the old picture. There were a total of 108 trials that were presented in a randomized order that tested memory for the 36 contrast items, 36 target items, and 36 of the filler items. The experiment lasted about 60 minutes in total.

Predictions

During the communication task, I predicted that speakers would produce modified noun phrases more often in Contrast-naming condition than the Non-contrast condition, replicating the lexical differentiation effect. If previous *reference* to the contrast item is sufficient to prompt lexical differentiation, modification rates should be equally high in the Contrast-naming and Contrast-location conditions, despite the fact that the contrast is not named in the Contrast-location condition.

In the present experiment, participants alternated roles between speaking and listening to better equate task engagement. If speakers continue to outperform listeners during the memory test, this would provide more convincing evidence that speakers and listeners formulate distinct memory representations for discourse referents during conversation. In contrast, if the distinct memory performance between speakers and listeners in Experiment 5 was due to differential engagement, I would expect speakers and listeners to perform equivalently on the memory test. Finally, by equating the number

---

trials. As a result we do not discuss the manipulation of lag further (and adding lag to the models did not improve model fit).

of exposures to contrast and target items, I anticipated that any effects of participant role (speaker vs. listener) should obtain for both targets and contrasts.

Results

Referential communication task

As in Experiment 5, I transcribed the speakers' productions on test trials and coded whether they produced a bare noun phrase or a modified noun phrase (Figure 3.5). In the Non-contrast condition, speakers modified their expressions 19.4% of the time, whereas modification rates were higher in both the Contrast-location (23.0%) and Contrast-naming conditions (23.7%).

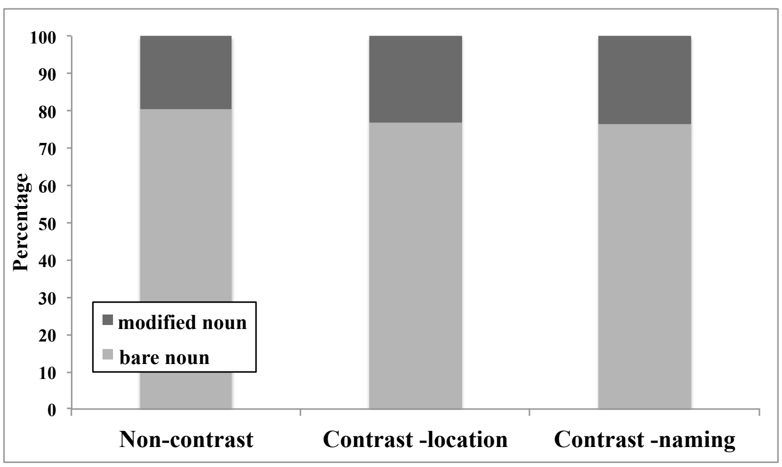


Figure 3.5. Percentage of each noun phrase type on test trials during entrainment in Experiment 6.

I analyzed the modification rates in a logistic mixed effects model with a binomial link function, using lag and entrainment type as fixed effects (see details in Appendix B). The analysis revealed a significant main effect of entrainment (Non-contrast vs. Contrast-location & Contrast-naming; z=2.270, p<.05), consistent with the results of Experiment 1. Speakers tended to differentiate the current referent from the past referent more when they had previously referenced the contrast item (regardless of whether it was named or spatially located), compared to when they had not been exposed to the contrast. The modification rates in the Contrast-location and Contrast-naming conditions did not differ significantly (z=0.203, p>.05).

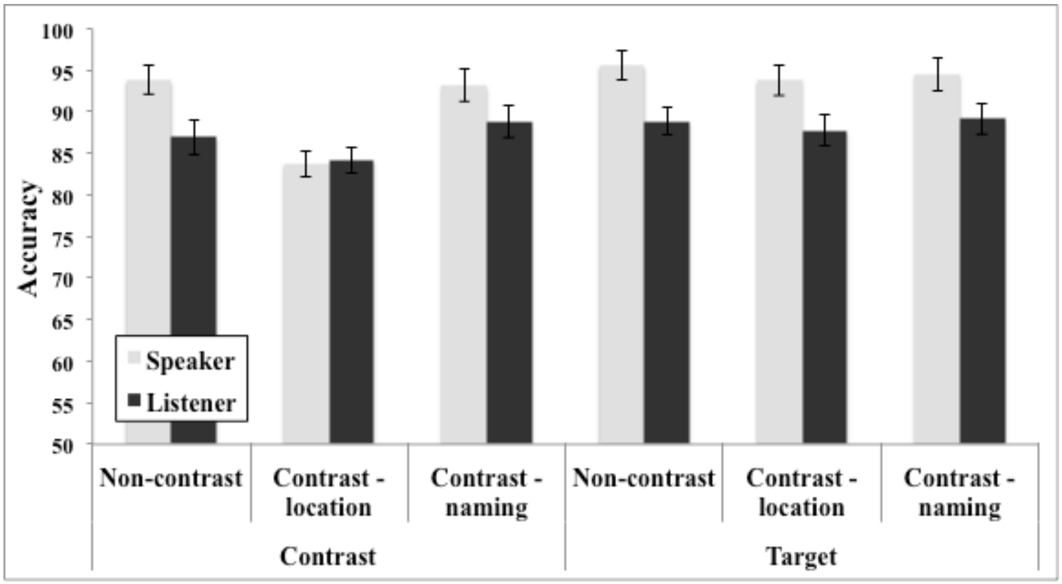


Figure 3.6. Accuracy on the memory test in Experiment 6.

Memory test phrase

Accuracy on the memory test was overall high for both speakers and listeners, though not at ceiling, unlike Experiment 5 (see Figure 3.6). I analyzed accuracy in a maximal mixed effects model with participant role (listener vs. speaker), referent (contrast vs. target) and entrainment types (Non-contrast vs. Contrast-location vs. Contrast-naming) as fixed effects (see details in Appendix B). Overall, speakers had better memory than listeners (z=4.181, p<.05). Memory was notably lower in the Contrast-location condition than the other two conditions (Non-contrast vs. Contrast-location & Contrast-naming: z=-2.552, p<.05; Contrast-location vs. Contrast-naming: z=3.677, p<.05). These main effects were qualified by a significant interaction between referent type (contrast vs. target) and entrainment type (Contrast-location vs. Contrast-naming, z=-2.413, p<.05), and a significant three-way interaction between role (listener vs. speaker), referent type (contrast vs. target), and entrainment type (Contrast-location vs. Contrast-naming, z=-1.646, p<.05).

Separate planned analyses for contrast and target items were used to explore the interactions. For target items, there was only a main effect of role, due to better memory performance by speakers (z=3.842, p<.05). For contrast items, there was a significant interaction between role and entrainment type (z=3.292, p<.05): Whereas speakers had better memory in the Contrast-naming than the Contrast-location condition (z=2.363, p<.05), listeners' memory did not differ across these conditions (z=0.340, p>.05). These findings show that speakers' memory for the contrast item was boosted *more* by naming than listeners' memory for the contrast.

Relationship between production and comprehension

In Experiment 5, I did not observe a significant relationship between language use and memory for the discourse context, possibly due to ceiling memory performance for the contrast item. In Experiment 6, I again examined whether speakers' contrast memory predicts the rate of differentiation, and did not find any significant evidence for an influence of contrast memory on the differentiation rate (z=1.382, p>.05).

As in Experiment 5, I also examined whether the way in which a target item was described affected how it was remembered. A mixed-effects model included role, entrainment type, and whether the target item was modified as fixed effects; the dependent measure was whether the target object was correctly recognized during the memory test. The results of this analysis revealed that when speakers had described the target with a modifier, *both* speakers and listeners remembered the referent better, compared to situations when the referent was identified by a bare noun phrase (z=3.916,

$p<.05$). This result suggests that modification is helpful for future memory. There was no interaction with role or entrainment type.

Summary and Discussion

In summary, the results of Experiment 6 demonstrated a lexical differentiation effect that was not contingent on *naming* (cf. Van Der Wege, 2009). The small magnitude of the effect, a ~4% increase in modification when an object from the same basic object category had previously been referenced, may be due to the fact that the contrast item had only been seen once during entrainment (vs. 6 times in Experiment 5). Performance on the surprise memory test demonstrated that speakers had better memory for past referents than listeners. Naming (rather than locating) past referents boosted speakers' memory for contrast items. Lastly, memory was improved for items that were described with modified noun phrases, providing evidence for a link between referential language and memory for the discourse. One caveat to this last finding, however, is that speakers were in control of whether they modified or not, so uncontrolled item differences may be influencing this effect.

Experiment 7

Experiments 5 and 6 show that speakers and listeners form distinct memories of the discourse record. Speakers' memory for previous discourse referents was overall better, and naming more strongly influenced speaker memory. While naming per se was not directly related to lexical differentiation, it was helpful for future memory.

A question not addressed by Experiments 5 and 6 is whether speakers have superior memory for undiscussed aspects of the referential context as well. If successful reference involves consideration of the referential alternatives in the discourse context

(Olson, 1970; Pechmann, 1989), then speakers should have better memory for unmentioned items in the scene. However, some findings in the memory literature suggest that *listeners* should have better context memory than speakers (Gopie & MacLeod, 2009; Jurica & Shimamura, 1999). According to this view, the act of speaking puts the speaker's attentional focus on the referent, at the expense of attention to the context (Koriat, et al., 1991). By contrast, listeners distribute attention more broadly, supporting better context memory. However, the findings, which support this view, come from paradigms that are not representative of conversational settings (e.g. speakers read sentences to imagined addressees).

The aim of the present study was to examine whether the observed speaking benefit for memory for the discourse, extends to memory for unmentioned aspects of the discourse context as well.

Method

Participants

Ninety-six undergraduates (forty-eight pairs) at the University of Illinois at Urbana-Champaign participated in return for partial course credit or cash payment ($8). Participants were native speakers of North American English and had normal hearing and normal or corrected-to-normal vision. No participant took part in Experiments 5 or 6.

Materials and procedure

The general procedure of Experiment 7 was identical to Experiment 6. Pairs of participants completed a communication task during which they alternated between the roles of speaker and listener in each block, followed by a 20-minute break and then a surprise memory test.

(a) (b)

Figures 3.7a-b. Example stimuli from entrainment trials (7a) and test trials (7b) in Experiment 7. Note that the context item (e.g., striped sock) in entrainment trials was never mentioned.

As before, the three types of entrainment trials (Non-contrast, Contrast-location, and Contrast-naming) formed our conditions of interest. One small change to the format of the entrainment trials was made, in order to test speakers' and listeners' memory for previously *un*mentioned items from the discourse context. Unlike Experiments 5 and 6, I included a contrasting context item (e.g., striped sock) during entrainment trials (See Figure 3.7a). In the entrainment trials of the Contrast-location and Contrast-naming conditions, participants viewed two items from the same category (e.g., argyle sock and striped sock), as well as two unrelated items. Speakers referred to one of two items (e.g., argyle sock) during the entrainment trial; the other object (e.g., striped sock) was never mentioned. Because two items from the same basic object category were in the immediate context, I expected speakers to use modifiers during the entrainment phase in the Contrast-naming condition (e.g., "argyle sock", rather than "sock"). For consistency in terminology across the experiments, I will refer to the mentioned object as the "contrast item", and the unmentioned object from the same basic object category as the “context item”. As in Experiments 5-6, in the Non-contrast condition, neither the contrast nor the context object was shown to participants, and speakers named an unrelated object during entrainment (e.g., apple).

Following the communication task, participants performed the same unrelated task as in previous experiments for 20 minutes and then completed an unexpected memory test. Unlike Experiments 5 and 6, I used a yes/no recognition memory test. On each trial, there was a single picture on the screen and participants were instructed to press the “Y” key if the picture was an old one that they had seen during the communication task and to press the “N” key, if the picture was new. In the two-alternative forced choice task used in Experiments 5 and 6, each memory test trial contained an old object, thus participants were forced to choose the more familiar picture even in cases where they were highly uncertain. In contrast, the yes/no recognition task used in Experiment 7 allows us to continue testing recognition memory, but in a task that allows participants to reject pictures (and say “new”) in cases where they fail to recognize an item. Participants completed 264 randomly ordered recognition test trials. Half of the pictures were old items and the other half were new. The old items included 36 contrast (e.g., argyle sock or apple), 36 target (e.g., dotted sock), 24 context (e.g., striped sock), and 36 filler items. The 132 new objects were drawn from the same category as each old item (e.g., three different new socks). Thus, during the memory test, participants were exposed to 6 different items from the same category; three were old and the other three were new. The order of test trials was random.

Predictions

The entrainment trials in Experiment 7 contained two objects from the same category (the contrast and context objects), thus speakers needed to use a modifier in order to uniquely identify the contrast object in the Contrast-naming condition. On test trials, a modified noun phrase was not necessary given the local context, since all items were from different basic object category. If, as in Experiments 5-6, speakers lexically differentiate in order to distinguish the current referent from past referents, the modification rate should be higher in the two contrast conditions compared to the Non-contrast condition.

With respect to memory for the discourse, I expected that speakers would continue to have better performance than listeners for both target and contrast items. The main question addressed by this study is whether this speaker benefit for past referents extends to unmentioned items in the visual context. If the process of designing an appropriately informative referring expression supports encoding of both the referent and its context, speakers should remember context items better than listeners. In contrast, if the act of speaking focuses attention on the referent at the expense of items in the context, listeners should show better memory for context items than speakers (Gopie & MacLeod, 2009; Koriat, et al., 1991).

Results

Communication task

Speakers almost always used modifiers (99.5%) to identify the target on entrainment trials in the Contrast-naming condition. On test trials, in the Non-contrast condition, speakers modified their expressions 16.3% of the time, while modification rates were higher in both the Contrast-location (20.2%) and Contrast-naming conditions

(24.7%). The analysis of target descriptions on test trials revealed that speakers used modifiers to lexically differentiate past referents (Figure 3.8). Modification rates were analyzed in a mixed effects model with entrainment type as a fixed effect (see Appendix B). The results revealed a main effect of entrainment type (NC vs. CN & CL: z=4.116, p<.05), due to significantly higher rate of modified noun phrases in both contrast conditions compared to the Non-contrast condition. Consistent with the results of Experiments 5-6, the modification rate did not differ significantly between the two contrast conditions (z=1.798, p>.05).

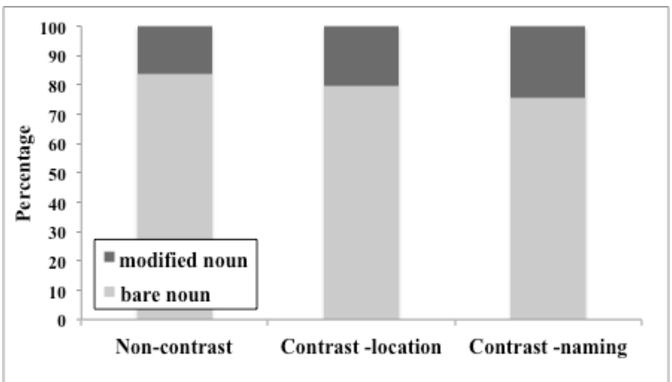


Figure 3.8. Percentage of each noun phrase type on test trials in Experiment 7.

Memory test phase

Performance on the memory task is plotted in Figure 3.9. in terms of the participants' ability to discriminate old from new items, or d'. Discriminability (d') was calculated by subtracting standardized false alarm rate from standardized hit rate. The use of the d' measure is preferred over other accuracy measures because it allows us to partial out the effects of response bias. The data were analyzed in three mixed effects models, which examined memory separately for target, contrast, and context items (note that an omnibus model with all three object types was not possible because the Non-contrast condition did not contain a context item). In all models, role (speaker vs. listener), entrainment (Non-contrast, Contrast-naming, Contrast-location) and item type (old vs. new) were included as fixed effects. Item type (old vs. new) was included as a fixed effect in order to separate correct acceptance of old items and correct rejections of new items from response bias (see Wright, Horry, & Skagerberg, 2008; Fraundorf, Watson, & Benjamin, 2010). Significant effects of item type show that participants were highly successful at discriminating old from new items; condition by item type interactions test whether condition influenced sensitivity to this distinction between old and new items. The dependent measure was binary; it coded whether the response on the memory test was 'yes' (old) or 'no' (new) (See Appendix for full model details).

For target items, speakers had better memory than listeners (target: z=-2.961, p<.01), consistent with Experiment 6. Target memory was also better in the Non-contrast condition than the other two conditions (z=7.113, p<.01), possibly due to confusion created by exposure to multiple items from the same category in the contrast conditions (i.e., Benjamin, 2001; Roediger & McDermott, 1995).

For contrast items, speakers had better memory than listeners (z=-3.503, p<.05). In addition, memory performance was substantially worse in the Contrast-location condition compared to the other two conditions (z=-9.618, p<.05). These effects were qualified by a three-way interaction between item type, role (speaker vs. listener), and entrainment type (z=2.046, p<.05): Speakers had better memory than listeners when the contrast had been labeled (Non-contrast: z=-3.500, p<.05; Contrast-naming: z=-3.773, p<.05), but not when it had been located (Contrast-location: z=-0.812, p>.05).

For context items, which were a newly included item category in Experiment 7, the analysis only included the two entrainment conditions for which a context item was present during entrainment (Contrast-naming and Contrast-location). Unlike target and contrast memory, speakers did not remember context items better (*z*=-0.173, *p*<.05). While context item memory was better in the Contrast-naming condition than in the Contrast-location condition (*z*=-9.162, *p*<.05), this effect did not differ as a function of participant role (*z* = -0.82, *p* = .42). Thus, naming the referent improved context memory for both speakers and listeners. These findings suggest that the speaker's superior memory for past referents does not extend to memory for unmentioned aspects of the discourse context.

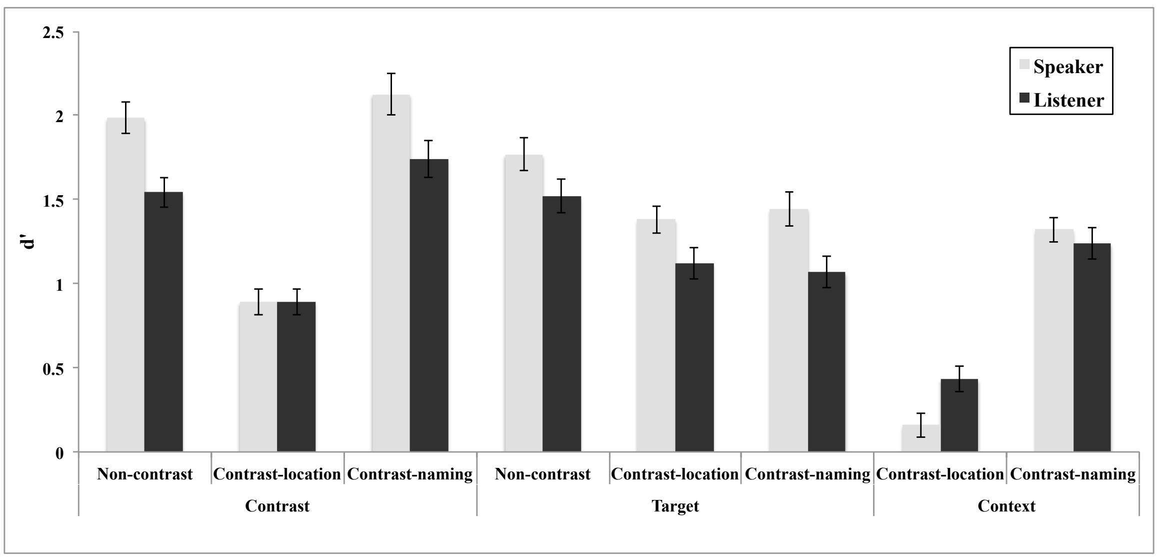


Figure 3.9. Discriminability (d') on the memory test in Experiment 3. Error bars indicate by-participant standard error of the mean.

Relationship between referential form and memory

As in Experiments 5 and 6, the relationship between referential form during the communication task and memory for past referents was examined. Consistent with the results of Experiments 5-6, the lexical differentiation rate was not related to contrast memory (*z*=-0.579, *p*>.05).

An analysis of the relationship between referential form on target trials and subsequent target memory revealed that interlocutors remembered target items better when they had been described during the test phase with modified noun phrases than when they had been described with bare noun phrases (z=-2.026, p<.05), consistent with the results of Experiment 6.

Summary and Discussion

The results of Experiment 7 again demonstrated a clear differentiation effect in the communication task that was not linked to naming. In addition, memory for previous discourse referents was better for speakers than for listeners. Though this benefit for past referents did not extend to unmentioned aspects of the discourse context (context items) selectively for speakers, *naming* (rather than locating) an intended referent did boost memory for unmentioned aspects of the discourse context for both speakers and listeners.

General discussion

Lexical differentiation and its source

The results of Experiments 5-7 replicate the lexical differentiation effect and reveal several new phenomena that hint at its origin. One explanation of the effect is that speakers differentiate in order to avoid using the same label (e.g., *sock*) to describe two different exemplars (Van der Wege, 2009). To test this hypothesis, the current studies included a condition in which the speaker used a locative phrase to identify the contrast object (e.g., *the top left one*), thus avoiding the potential for lexical conflict at test. Across the three experiments, however, the differentiation rate did not differ as a function of whether the contrast item had been located or labeled, suggesting that lexical conflict is not the source of differentiation (21.3% in the Contrast-location vs. 23.4% in the

Contrast-naming condition). Instead, it seems that speakers differentiate in order to more generally distinguish current from past referents.

The fact that speakers differentiated both in the locative and labeling conditions also clarifies an earlier puzzling finding. Yoon and Brown-Schmidt (2013) showed that listeners were equally likely to expect speakers to produce a modified noun phrase to describe a target referent when a contrasting referent had previously been located vs. labeled. Although they suggested that listeners did not expect speakers to differentiate, the present results offer a different interpretation of those findings. In that study, listeners may have been sensitive to the fact that differentiation can be prompted by previous reference to a similar object in the past, regardless of the form of that reference.

Lexical differentiation is also observed under conditions in which speakers had already used a modifier to describe the contrast object. In Experiment 7, describing the target item with a bare noun phrase would be sufficient to identify the target with respect to both the local and historical context, as contrast items in the Contrast-naming condition were described during entrainment with a modifier 99.5% of the time (e.g., *argyle sock*). As a result, an unmodified description of the target (e.g., *sock*, rather than *dotted sock*) would suffice to distinguish the target from the contrast. The fact that speakers produced modifiers on test trials at similar rates in Experiments 6 and 7 (23.7% and 24.7%, respectively) adds to the evidence that it is the distinction between *referents* and not *referential form* that gives rise to the lexical differentiation effect.

Lastly, recognition memory for the contrast item was high for speakers and listeners alike. These findings suggest that the low differentiation rate (about 5% in these experiments) is not caused by the speaker's failure to remember the contrast item. An alternative explanation of the low differentiation rate may be that speakers viewed the previous discourse as irrelevant to the current discourse context—that is, that the memories were intact but not accessed. In a conversation in which the past was relevant to the topic at hand, differentiation might be more pronounced. A final consideration is that memory was tested using a recognition paradigm. An open question is whether the relationship between memory and lexical differentiation would be more pronounced if memory was measured in a free-recall task. While the recall rate for conversational memory is generally low (Stafford & Daly, 1984), the act of recalling information from memory may be more similar to what speakers do when they bring to mind past discourse contexts and relate them to the current topic of conversation.

Distinct memory representations for speakers and listeners

Interlocutors are thought to maintain representations of the discourse context in the form of rich representations of joint experiences (Clark & Marshall, 1978; Brown-Schmidt, 2012), and through the automatic association of partners and referents (Horton & Gerrig, 2005a). Here we examined whether speakers and listeners develop distinct memory representations while communicating.

Memory performance was quite good for both speakers and listeners, even in Experiments 6 and 7 where a 20-minute break was introduced between the communication task and the unexpected memory test. This finding is generally consistent with the idea that discourse representations are maintained over time, even in a task-based dialogue where there is little pressure to remember previous topics. However, consistent differences in memory performance between speakers and listeners imply that the representation of the discourse record varies as a function of one's role in a discourse.

The fact that the speaker advantage obtained even in Experiments 6 and 7, where participants alternated roles of speaking and listening, suggests that this speaker-benefit operates at the level of individual utterances within the discourse (see also McKinley et al., 2015). Notably, the speaker benefit did not extend to unmentioned aspects of the discourse context (Experiment 7, context memory).

Why does speaking improve referent memory? One explanation is that speakers invest more effort into the planning of utterances compared to the amount of effort needed to interpret the same utterance. This asymmetric effort explanation is consistent with the idea that the act of generating material increases the mental effort or depth of processing during encoding (Graf, 1980; McFarland et al., 1980, Slamecka & Graf, 1978). This view predicts that speakers who invest more effort during encoding should perform better on subsequent memory tests. While we have no direct evidence that would speak to this hypothesis, the fact that the use of a modified noun phrase on target trials (e.g., *spotted sock* vs. *sock*) improved subsequent target memory is generally consistent with this idea.

An alternative possibility is that interlocutors develop different strategies to encode information depending on their role as speaker or listener. The act of speaking requires utterance planning and may place more focus on the referent at the expense of attention to the immediate context (see Gopie & MacLeod, 2009; Jurica & Shimamura, 1999; Koriat et al., 1991; cf. Brown-Schmidt & Tanenhaus, 2006; McKinley et al., 2015). By contrast, listening requires evaluating the unfolding speech signal with respect to the candidate referents in the discourse context (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995), and as a result may improve encoding of both the intended referent as well as the alternatives in the local context. While our findings do show the predicted speaker benefit on this account, there was no evidence that this was at the expense of memory for the context. Similar to recent findings by McKinley et al. (under review), speakers and listeners showed equivalent context memory that was improved—regardless of role—when the speaker named the object rather than located it.

Context memory

It is known that the historical record of a discourse (i.e., memory for the discourse history) includes representations of events and their participants, as well as how past referents have been described (Brennan & Clark, 1996; Greene, Gerrig, McKoon, & Ratcliff, 1994; Nieuwland, Otten, & Van Berkum, 2007). Less clear is whether unmentioned objects in the discourse context are encoded in memory as well, and if so, how they are bound or linked to referenced items. In Experiment 7, when speakers had to describe the target using a modified noun phrase to distinguish it from the context item (Contrast-naming condition), memory for the context items was comparable to that for memory for targets, and similar for speakers (target d'=1.45; context d'=1.32) and listeners (target d'= 1.12; context d'=1.23). By contrast, memory for context items decreased dramatically when the speaker produced a locative (Context-location condition; speaker: context d'=0.16, listener: d'=0.43). These findings show that there is flexibility in how a discourse context is (or is not) encoded. Speakers and listeners do not automatically encode everything in the context, but instead selectively encode information that is conversationally-relevant.

Context in the present studies was defined as the *referential context.* Designing an appropriate referring expression in Experiment 7 required taking into consideration the properties of both the target and the context item in order to select words that would uniquely identify the target. For example, given the scene at entrainment in Figure 3.7a, modified expressions such as *the tall sock* or *the multicolored sock* would not suffice to identify the target; instead the speaker would have to take both the target and the context item into consideration to select an expression that uniquely picks out the target such as *the argyle sock*. This joint consideration of both an intended referent and its local context may support context memory in natural conversation. Whether other types of contexts (such as memory for one's discourse partner) may be similarly boosted in natural conversation remains an open question. The present research also does not address how unmentioned and irrelevant aspects of the discourse context are encoded in memory. For example, in Figure 3.7a, the bunny and the pie were never mentioned and were largely irrelevant to describing the target (sock); whether naming (rather than locating) the target would boost memory for these irrelevant context objects is unknown. If the memory boost for the context object in the naming condition was due to the relevance of the context object to the target, this naming benefit would likely not extend to irrelevant aspects of the context.

CHAPTER 4: CONCLUSION

During conversation, interlocutors establish representations of what each other do and don't know; these representations of mutual knowledge guide language use (Brennan & Clark, 1996; Metzing & Brennan, 2003; Brown-Schmidt, 2009). Speakers design their utterances based on the knowledge and perspective of the addressee, and memory representations of this mutual knowledge, or common ground, are necessary for successful audience design. Previous studies of two party conversation showed that speakers are sensitive to the common ground shared with their partners (Schober & Clark, 1989; Isaacs & Clark, 1987; Krauss & Weinheimer, 1964, 1966; Wilkes-Gibbs & Clark, 1992; Van der Wege, 2009; Yoon & Brown-Schmidt, 2014). How these processes scale up to more complex settings, including multiparty conversation is unknown. Further, little is known about how mutual knowledge is stored in memory during natural conversation. One goal of the present research was to use multiparty conversation as a test case in order to provide insight into how perspective is represented, and how these representations are brought to bear on the process of audience design. Another goal was to use the phenomenon of lexical differentiation in order to explore links between language use and memory for the discourse context.

In Chapter 2, I show that speakers develop and maintain distinct and gradient representations of the common ground held with different individuals, and use these representations to guide audience design. In multiparty conversation, speakers emphasized both the knowledgeable and the naïve addressees' distinct perspectives and designed their referential expressions accordingly. Taken together, the findings point to an audience design process that considers all of the addressees' distinct representations of common ground, where speakers flexibly take into consideration all of the addressees' knowledge states to maximize the understanding of all addressees in multiparty conversation.

In multiparty conversation, identifying the circumstances in which speakers depart from a *Combine* approach to audience design remains a goal for future work; such inquiries will likely benefit from considering the relevant behavioral goals that are motivating the conversation in the first place. Speakers are under pressure to maintain both conversational efficiency and informativity. Depending on conversational goals, they may approach audience design in a way that emphasizes one of these goals more than the other. Likewise, there are likely to be limits on the number of individuals for whom speakers can maintain distinct representations of common ground; identifying these limits is a goal for future work.

In Chapter 3, the present research replicates and extends previous findings, demonstrating that speakers differentiate their referring expressions to distinguish current from past referents. By investigating the relationship between language use in dialogue and memory for the discourse, I show that low rates of differentiation are not due to failures to remember the past context; instead recognition memory for past discourse referents is quite good. Further, the findings demonstrate a generation effect for item memory in a natural conversational setting, pointing to an asymmetry in memory between speakers and listeners. This is inconsistent with Pickering and Garrod's (2004; 2013) proposal that speakers and listeners align at every level of linguistic representation in conversation. While language production is clearly influenced by previous language comprehension in dialogue (Haywood, Pickering, & Branigan, 2005; Heller & Chambers, 2013), my findings show that the memorial representations created during the processes of language comprehension and language production are distinct. The fact that memory for past referents varied on an item-by-item basis depending on a person's role at the time in the conversation, and how that item was described, points to a high degree of flexibility (and variability) in how discourse referents will be encoded across the course of a conversation. This finding implies that successfully designing utterances based on the knowledge state of one's addressee likely requires conversational partners to appreciate the fact that listener memory may be fleeting. In some cases then, a failure to assume common ground for previously discussed information may, in fact, reflect successful modeling of the listener's mental state.

In conclusion, the present research offers new insights into the mechanisms of conversation including how conversational partners use language in multiparty settings, and how memory and language interact to support conversational processes. This work also exemplifies how understanding the relevant memory processes involved in conversation is crucial for fleshing out a theory of how discourse history guides language use. This integrative approach to the study of language use in dialogue with measures of memory for the discourse history represents a key step forward in developing a unified theoretical framework of the cognitive processes underlying language use in conversation.

REFERENCES

Arnold, J. E., Hudson Kam, C., & Tanenhaus, M. K. (2007). If you say thee uh- you're describing something hard: the on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 914–930.

Award, M., Warren, J. E., Scott, S. K., Turkheimer, F. E., & Wise, R. J. S. (2007). A common system for the comprehension and production of narrative speech. *The Journal of Neuroscience, 27,* 11455-11464.

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,* 255–278.

Ben Shalom, D. & Poeppel, D. (2008) Functional anatomic models of language: assembling the pieces. *The Neuroscientist, 14,* 119–27.

Bersch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: a meta-analytic review. *Memory & Cognition, 35,* 201-210.

Bradshaw, G. L., & Anderson, J. R. (1982). Elaborative encoding as an explanation of levels of processing. *Journal of Verbal learning and Verbal Behavior, 21*, 165-174.

Brennan, S. E., & Clark, H. H. (1996). Conceptual Pacts and Lexical Choice in Conversation. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *22*, 482-493.

Bromme, R., Jucks, R., & Wagner, T. (2005). How to refer to 'Diabetes'? Language in online health advice. *Applied Cognitive Psychology, 19*, 569-586.

Brown-Schmidt, S. (2009a). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language, 61*, 171-190.

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes, 27,* 62-89.

Brown-Schmidt, S. & Horton, W. S. (2014). The Influence of Partner-Specific Memory Associations on Picture Naming: A Failure to Replicate Horton (2007). PLOS ONE 9(10): e109035. doi:10.1371/journal.pone.0109035.

Brown-Schmidt, S. & Tanenhaus, M.K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language, 54,* 592-609.

Brown-Schmidt, S., Yoon, S., & Ryskin, R. A. (2015). People as Contexts in Conversation. *Psychology of Learning and Motivation, 62,* 59-99.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clark, E. V., & Clark, H. H. (1979). When nouns surface as verbs. *Language*, 55, 767-811. *Reprinted* in: F. Katamba (Ed.), *Critical concepts in linguistics*, vol. 5*: Morphology: its relation to semantics and the lexicon*. London: Routledge, 2003. Pp. 128-183.

Clark, H. H., & Marshall, C. R. (1978). Reference diaries (pp. 57-63). In D. L. Waltz (Ed.), *Theoretical issues in natural language processing* (Vol. 2). New York: Association for Computing Machinery.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A.

K. Joshi, B. Webber, & I. Sag (Eds.), *Elements of discourse understanding* (pp. 10-63). Cambridge: Cambridge University Press.

Clark. H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In J.-F. L. Ny & W. Kintsch (Ed.), *Language and comprehension* . New York: North Holland.

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology, 37,* 201–242.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-39.

Dell, G. S., & Brown, P. M. (1991). Mechanisms for listener-adaptation in language production: limiting the role of the "model of the listener". In D. Napoli & J. Kegl (Eds.), *Bridges between psychology and linguistics.* New York: Academic Press.

Donaldson, W., & Bass, M. (1980). Relational information and memory for problem solutions. *Journal of Verbal Learning and Verbal Behavior, 19,* 26-35.

Duff, M. C. & Brown-Schmidt, S. (2012). The hippocampus and the flexible use and processing of language. *Frontiers in Human Neuroscience, 6,* 1-9.

Duff, M.C., Hengst, J., Tranel, D., & Cohen, N.J. (2006). Development of shared information in communication despite hippocampal amnesia. *Nature Neuroscience*, *9,* 140-146.

Duff, M. C., Hengst, J., Tranel, D., and Cohen, N. J. (2008). Collaborative discourse facilitates efficient communication and new learning in amnesia. *Brain Lang.* 106, 41–54.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistics Research, 24*, 409-436.

Eich, E. (1985). Context, memory, and integrated item/context imagery. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 764–770.

Engelhardt, P. E., Bailey, K. G. D., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language, 54,* 554–573.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language, 30,* 210–233.

Fraundorf, S. H., Watson, D. G., & Benjamin, A. S. (2010). Recognition memory reveals just how CONTRASTIVE contrastive accenting really is. *Journal of Memory and Language, 63,* 367-386.

Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fussell, S.R. & Krauss, R.M. (1991). Accuracy and bias in estimates of others' knowledge. *European Journal of Social Psychology, 21,* 445-454.

Fussell, S.R. & Krauss, R.M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology, 62*, 378-391.

Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, 62, 35-51.

Gann, T. M., & Barr, D. J. (2012). Speaking from experience: Audience design as expert performance. *Language and Cognitive processes, 2012*, DOI:10.108 0/01690965 .2011.6413 88.

Gopie, N., & MacLeod, C. M. (2009). Destination memory: Stop me if I've told you this before. *Psychological Science, 20,* 1492-1499.

Gorman, K. S., Gegg-Harrison, W., Marsh, C. R., & Tanenhaus, M. K. (2013). What's learned together stays together: Speakers' choice of referring expression reflects shared experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 843-853.

Graf, P. (1980). Two consequences of generating: Increased inter and intra word organization of sentences. *Journal of Verbal Learning and Verbal Behavior, 19,* 316-327.

Graf, P. (1982). The memorial consequences of generation and transformation. *Journal of Verbal Learning and Verbal Behavior, 21,* 539-548.

Greene, S. B., Gerrig, R. J., McKoon, G., & Ratcliff, R. (1994). Unheralded pronouns and management by common ground. *Journal of Memory and Language, 33*, 511-526.

Haberman, J. & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics, 72,* 1825-1838.

Heller, D., Gorman, K., & Tanenhaus, M. K. (2012). To name or to describe: Shared knowledge affects referential form. *Topics in Cognitive Science, 4,* 290-305.

Herrmann, D. J., & Pearle, P. M. (1981). The proper role of clusters in mathematical models of continuous recall. *Journal of Mathematical Psychology, 24*, 139-162.

Holler, J., & Wilkin, K. (2009). Communicating common ground: how mutually shared knowledge influences the representation of semantic information in speech and gesture in a narrative task. *Language and Cognitive Processes, 24*, 267-289.

Horton, W. S. (2007). Metaphor and readers' attributions of intimacy. *Memory & Cognition*, *35*, 87-94.

Horton, W. S., & Gerrig, R. J. (2002). Speakers' experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language, 47*, 589-606.

Horton, W. S., & Gerrig, R. J. (2005a). Conversational common ground and memory processes in language production. *Discourse Processes, 40,* 1–35.

Horton, W. S., & Gerrig, R. J. (2005b). The impact of memory demands on audience design during language production. *Cognition, 96,* 127–142.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*, 91-117.

Horton, W. S., & Spieler, D. H. (2007). Age-related effects in communication and audience design. *Psychology and Aging*, *22*, 281-290.

Isaacs, E., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General, 116*, 26-37.

Johnson, M., Raye, C , Foley, H., & Foley, M. (1981). Cognitive operations and decision bias in reality monitoring. *American Journal ofPsychology, 94,* 37-64.

Jurica, P. J., & Shimamura, A. P. (1999). Monitoring item and source information: Evidence for a negative generation effect in source memory. *Memory & Cognition, 27*, 648-656.

Knutsen, D., & Le Bigot, L. (2012). Managing dialogue: How information availability affects collaborative reference production. *Journal of Memory and Language, 67,* 326-341.

Knutsen, D., & Le Bigot, L. (2014). Capturing egocentric biases in reference reuse during collaborative dialogue. *Psychonomic Bulletin and Review, 21,* 1590-1599.

Koriat. A., Ben-Zur, H., & Druch, A. (1991). The contextualization of input and output events in memory. *Psychological Research, 53,* 260-270.

Kraljic, T. & Samuel, A.G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*, 141-178.

Krauss, R, M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction. A preliminary study. *Psychonomic Science , 1,* 113-114.

Krauss, R, M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, *4*, 343–346.

Levelt, W. J. M. (1993). Lexical selection, or how to bridge the major rift in language processing. In: Theorie und Praxis des Lexikons, ed. F. Beckmann & G. Heyer. De Gruyter.

Lichtheim, L. (1885). On aphasia. *Brain, 7,* 433-484.

MacLeod. C. Gopie, N. Hourihan, K. L, Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. Journal of Experimental Psychology: *Learning, Memory, and Cognition, 36,* 671-685.

Marsh, E. J., Edelman, G., & Bower, G. H. (2001). Demonstrations of a generation effect in context memory. *Memory & Cognition, 29,* 798 – 805.

Marsh, R. L., & Hicks, J. L. (2002). Comparisons of target output monitoring to source input monitoring. *Applied Cognitive Psychology, 16,* 845-862.

Mayr, U. (2006). What matters in the cued task-switching paradigm: Tasks or cues? *Psychonomic Bulletin & Review, 13*, 794-799.

McDaniel, M. A., Waddill, P. J. & Einstein, G. O. (1988). A contextual account of the generation effect: A three-factor theory. *Journal of Memory and Langauge, 27*, 521-536.

McFarland, C, Frey, T., & Rhodes, D. (1980). Retrieval of internally versus externally generated words in episodic memory. *Journal of Verbal Learning and Verbal Behavior, 19,* 210-225.

McKinley, G. L., Brown-Schmidt, S., & Benjamin, A. (2015). Common ground in conversation: Evidence from item and context memory. Poster presented at the annual meeting of the Psychonomic Society, Chicago, IL.

McNamara, D. S., & Healy, A. F. (2000). A procedural explanation of the generation effect for simple and difficult multiplication problems and answers. *Journal of Memory & Language, 43,* 652-679.

Metzing, C. & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects in the comprehension of referring expressions. *Journal of Memory* *and Language, 49*, 201-213.

Nadig, A. & Sedivy, J. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, *13,* 329-336.

Nieuwland, M. S., Otten, M., & Van Berkum, J. J. A. (2007). Who are you talking about? Tracking discourse-level referential processing with event-related brain potentials. *Journal of Cognitive Neuroscience, 19*, 228-236.

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review, 77,* 257–273.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics, 27,* 89–110.

Pickering, M. J., & Garrod, S. (2004). The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences, 27,* 212-225.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36,* 329-347.

Richardson, D.C & Dale, R. (2005). Looking To Understand: The Coupling Between Speakers' and Listeners' Eye Movements and its Relationship to Discourse Comprehension. *Cognitive Science, 29,* 1045–1060.

Richardson, D.C., Dale, R. & Kirkham, N.Z. (2007) The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science, 18,* 407-413.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*, 207-231.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of*

Personality and Social Psychology, 37, 322-336.

Ryskin, R.A., Brown-Schmidt, S., Canseco-Gonzalez, E., Yiu, E.K., & Nguyen, E.T. (2014). Visuospatial Perspective-taking in Conversation and the Role of Bilingual Experience. *Journal of Memory and Language, 74,* 46-76.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50,* 696-735.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*, 211-232.

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences, 111,* 4687-4696.

Shepard, R. N. (1967). Recognition memory for words, sentences and pictures. *Journal of Verbal Learning and Verbal Behavior, 6,* 156-163.

Slamecka, N. J. & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning & Memory, 4*, 592-604.

Slocum, J. (2004). *The tangram book*. New York, NY: Sterling Publishing.

Stafford, L. Burggraf, C. S., Sharkey, W. F. (1987). Conversational memory: the effects of time, recall mode, and memory expectancies on remembrances of natural conversations. *Human Communication Research, 14,* 203-229.

Stafford, L., & Daly, J. A. (1984). Conversational memory: The effects of recall mode and memory expectancies on remembrances of natural conversations. *Human Communication Research, 10,* 379-402.

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension*. Cambridge, MA: MIT Press.

Trueswell, J. C., & Tanenhaus, M. K. (2005). *Approaches to studying world-situated language use: Birding the language-as-product and language-as-action traditions*. Cambridge, MA: MIT Press.

Van der Wege, M. M. (2009). Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, *60*, 448-463.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*, 183-194.

deWinstanley, P.A., Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition, 32,* 944-955.

Wright, D. B., Horry, R., & Skagerberg, E. M. (2008). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods, 41,* 257–267.

Yoon, S., & Brown-Schmidt, S. (2013). Lexical differentiation effects in language production and comprehension. *Journal of Memory and Language, 69,* 397-416.

Yoon, S., & Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 40*, 919-937.

Yoon, S., & Brown-Schmidt, S. (submitted). Aim Low: Mechanisms of audience design in multiparty conversation.

Yoon, S., Koh, S., & Brown-Schmidt, S. (2012). Influence of perspective and goals on reference production in conversation. *Psychonomic Bulletin and Review, 19,* 699-707.

Young, C. J. (2004). Contributions of metaknowledge to retrieval of natural categories in semantic memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30*, 909-916.

APPENDIX A: Model results for Chapter 2

Statistical Analysis

In Experiments 1-4, the statistical analyses focus on the length of the Director's referring expressions (the number of words), as well as the fluency of the referring expression at test. In Experiment 2, I additionally examined whether the Director reconceptualized their expressions at test.

Analyses of expression length used Poisson-link mixed effects models. Analyses of disfluency and reconceptualization used logit-link mixed effects models, which model the log odds of disfluency or reconceptualization, respectively. In all analyses, models were fit using the lmer package in R, with the maximal random effects structure for subjects and items. In cases where the maximal model did not converge, a backwards-fitting procedure was used to identify the model with the largest random effects structure that would converge (see Barr, Levy, Scheepers, & Tily, 2013). Fixed effects were coded with mean-centered Helmert contrasts.

Table A.1. *Experiment 1 test trials.* Mixed effect model with degree of common ground and recency as fixed effects. The dependent measure is the number of words in the referring expression at test.

	Estimate	SE	*z*-value	*p*-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	1.721	0.092	18.674	<.0001	*Subject*	(Intercept)	0.138	0.372
CG	-0.033	0.038	-0.854	0.393		CG	0.007	0.085
Recency	-0.005	0.326	-0.152	0.879		Recency	0.007	0.084
CG*Recency	0.166	0.053	**3.132**	**0.002**	*Trial*	(Intercept)	0.091	0.302
						CG	0.011	0.103
						Recency	<.001	0.011
						CG*Recency	0.038	0.196

Table A.2. *Experiment 1 Disfluency.* Mixed effect model with degree of common ground and recency as fixed effects. The dependent measure (binary) is whether the expression was fluent or not.

	Estimate	SE	*z*-value	*p*-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	-0.931	0.305	-3.049	0.002	*Subject*	(Intercept)	1.679	1.296
CG	-0.246	0.154	-1.595	0.111		CG	0.016	0.127
Recency	-0.230	0.149	-1.545	0.122		Recency	0.047	0.217
CG*Recency	0.106	0.403	0.263	0.792		CG*Recency	1.604	1.266
					Trial	(Intercept)	0.223	0.472
						CG	0.041	0.202
						Recency	0.039	0.198
						CG*Recency	0.000	0.008

Table A.3. *Experiment 2 test trials:Reference Length.* Mixed effect model with repetition and partner (High CG, Low CG, High+Low CG) as fixed effects. The dependent measure is the number of words in each referential expression at test.

	Estimate	SE	*z*-value	*p*-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	1.773	0.073	24.397	<.0001	*Subject*	(Intercept)	0.092	0.304
Repetition	0.325	0.071	**4.586**	**<.0001**		Repetition	0.050	0.223
Partner1	0.557	0.083	**6.751**	**<.0001**		Partner1	0.117	0.342
(High vs. Low & High+Low)								
Partner2 (Low vs. High+Low)	0.072	0.062	1.158	0.247		Partner2	0.061	0.247
Repetition*Partner1	0.329	0.054	**6.072**	**<0.001**	*Trial*	(Intercept)	0.047	0.216
Repetition*Partner2	0.087	0.051	1.696	0.090		Repetition	0.085	0.291
						Partner1	0.011	0.107
						Partner2	0.021	0.146
						Repetition*Partner1	0.085	0.292
						Repetition*Partner2	0.019	0.137

Table A.4. *Experiment 2 test trials: Reference Length.* Mixed effect model with repetition and partner (High/Low CG switching) as fixed effects. The dependent measure is the number of words in each referential expression at test.

	Estimate	SE	*z*-value	*p*-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	1.835	0.081	22.614	<.0001	*Subject*	(Intercept)	0.114	0.338
Repetition	0.102	0.060	1.717	0.086		Repetition	0.004	0.064
Partner (High-Switching vs. Low-Switching)	0.615	0.088	**7.019**	**<.0001**		Partner	0.161	0.341
Repetition*Partner	0.104	0.092	1.125	0.261		Repetition*Partner	0.032	0.178
					Trial	(Intercept)	<.0001	<.0001
						Repetition	0.289	0.538
						Partner	<.001	0.003
						Repetition*Partner	0.351	0.593

Table A.5. *Experiment 2 test trials: Disfluency.* Mixed effect model with repetition and condition (High CG, Low CG, High+Low CG) as fixed effects. The dependent measure (binary) is whether the expression was fluent or not.

	Estimate	SE	*z*-value	*p*-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	-0.877	0.255	-3.443	0.001	*Subject*	(Intercept)	0.326	0.571
Repetition	1.401	0.279	**5.020**	**<.0001**		Repetition	0.096	0.309
Partner1 (High vs. Low & High+Low)	**3.102**	**0.456**	**6.810**	**<.0001**		Partner1	0.736	0.858
Partner 2 (Low vs. High+Low)	0.171	0.266	0.644	0.520		Partner2	0.142	0.377
Repetition*Partner1	2.623	0.577	**4.547**	**<.0001**	*Trial*	(Intercept)	0.301	0.549
Repetition*Partner2	-0.122	0.450	-0.270	0.787		Repetition	0.583	0.764
						Partner1	0.853	0.924
						Partner2	0.051	0.226

Table A.6. *Experiment 2 test trials: Disfluency.* Mixed effect model with repetition and condition (High/Low CG switching condition) as fixed effects. The dependent measure (binary) is whether the expression was fluent or not.

	Estimate	SE	*z*-value	*p*-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	-0.325	0.230	-1.414	0.157	*Subject*	(Intercept)	0.494	0.703
Repetition	0.365	0.285	1.283	0.200		Repetition	0.013	0.113
Partner (high-switching vs. low-switching)	**2.969**	**0.346**	**8.585**	**<.0001**		Partner	0.000	0.019
Repetition*Partner	0.689	0.624	1.104	0.270	*Trial*	(Intercept)	0.123	0.351
						Repetition	0.513	0.716

Table A.7. *Experiment 2 test trials: Reference reconceptualization.* Mixed effect model with repetition and condition (High CG, Low CG, High+Low CG) as fixed effects. The dependent measure (binary) is whether the expression was reconceptualized or not.

	Estimate	SE	*z*-value	*p*-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	-0.264	0.243	-1.084	0.278	*Subject*	(Intercept)	0.919	0.959
Repetition	0.560	0.356	1.574	0.116		Repetition	1.469	1.212
						Partner1	1.154	1.074
Partner1 (High vs. Low & High+Low)	**2.245**	**0.279**	**8.039**	**<.001**				
Partner 2 (Low vs. High+Low)	0.516	0.218	**2.361**	**0.018**		Partner2	0.520	0.721
Repetition*Partner1	0.195	0.272	0.716	0.474	*Trial*	(Intercept)	0.528	0.727
Repetition*Partner2	0.442	0.283	1.565	0.118		Repetition	2.662	1.631
						Partner1	0.050	0.228
						Partner2	0.049	0.221

Table A.8. *Experiment 2 test trials: Reconceptualization.* Mixed effect model with repetition and condition (High/Low CG switching condition) as fixed effects. The dependent measure (binary) is whether the expression was reconceptualized or not.

	Estimate	SE	*z*-value	*p*-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	-0.270	0.248	-1.089	0.276	*Subject*	(Intercept)	0.954	0.977
Repetition	0.603	0.246	1.259	0.158		Repetition	0.135	0.367
Partner (high-switching vs. low-switching)	2.335	0.266	**8.793**	**<.0001**		Partner	0.474	0.688
Repetition*Partner	0.727	0.426	0.704	0.388		Repetition*Partner	0.092	0.304
					Trial	(Intercept)	0.433	0.658
						Repetition	0.013	0.113
						Partner	0.382	0.618
						Repetition*Partner	0.175	0.418

Table A.9. *Experiment 3a test trials: Reference Length.* Mixed effect model with degree of common ground as a fixed effect. The dependent measure is the number of words in the referring expression at test.

	Estimate	SE	z-value	p-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	2.112	0.073	29.646	<.0001	*Subject*	(Intercept)	0.104	0.322
CG1 (K vs. 2K1N, 1K2N, N)	0.069	0.067	**10.261**	**<.0001**		CG1	0.095	0.307
CG2 (2K1N vs. 1K2N, N)	0.179	0.051	**3.541**	**<.001**		CG2	0.051	0.226
CG3 (1K2N vs. N)	0.063	0.064	0.984	0.325		CG3	0.088	0.297
					Trial	(Intercept)	0.081	0.285
						CG1	0.095	0.307
						CG2	0.051	0.226
						CG3	0.088	0.297

Table A.10. *Experiment 3a test trials: Disfluency.* Mixed effect model with degree of common ground as a fixed effect. The dependent measure (binary) is whether the expression was fluent or not.

	Estimate	SE	*z*-value	*p*-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	0.339	0.251	1.350	0.177	*Subject*	(Intercept)	1.462	1.209
CG1 (K vs. 2K1N, 1K2N, N)	1.882	0.205	**9.169**	**<.0001**		CG1	0.330	0.575
CG2 (2K1N vs. 1K2N, N)	0.597	0.266	**2.247**	**0.025**		CG2	1.253	1.119
CG3 (1K2N vs. N)	0.321	0.216	1.483	0.138		CG3	0.249	0.499
					Trial	(Intercept)	0.390	0.625
						CG1	0.258	0.508
						CG2	0.007	0.086
						CG3	0.054	0.232

Table A.11. *Experiment 3b test trials: Reference Length.* Mixed effect model with degree of common ground as a fixed effect. The dependent measure is the number of words in the referring expression at test.

	Estimate	SE	*z*-value	*p*-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	2.068	0.058	35.48	<.0001	*Subject*	(Intercept)	0.077	0.278
CG1	0.719	0.048	**15.13**	**<.0001**		CG1	0.044	0.211
(K vs. 3K1N, 2K2N, 1K3N, N)								
CG2	0.100	0.041	**2.43**	**0.015**		CG2	0.045	0.212
(3K1N vs. 2K2N, 1K3N, N)								
CG3 (2K2N vs. 1K3N, N)	0.042	0.051	0.83	0.406		CG3	0.072	0.267
CG4 (1K3N vs. N)	0.068	0.041	1.65	0.098		CG4	0.037	0.193
					Trial	(Intercept)	0.091	0.301
						CG1	0.034	0.183
						CG2	0.008	0.087
						CG3	0.012	0.108
						CG4	0.014	0.117

Table A.12. *Experiment 3b test trials: Disfluency.* Mixed effect model with degree of common ground as a fixed effect. The dependent measure (binary) is whether the expression was fluent or not.

	Estimate	SE	*z*-value	*p*-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	-0.129	0.243	-0.532	0.595	*Subject*	(Intercept)	1.845	1.358
CG1 (K vs. 3K1N, 2K2N, 1K3N, N)	1.413	0.252	**5.607**	**<.0001**		CG1	1.518	1.232
CG2 (3K1N vs. 2K2N, 1K3N, N)	0.366	0.161	**2.280**	**0.022**		CG2	0.336	0.580
CG3 (2K2N vs. 1K3N, N)	0.328	0.198	1.657	0.098		CG3	0.738	0.859
CG4 (1K3N vs. N)	0.497	0.208	**2.389**	**0.017**		CG4	0.602	0.776
					Trial	(Intercept)	0.274	0.523

Table A.13. *Experiment 4 test trials: Reference Length.* Mixed effect model with visual context as a fixed effect (Context1: (4) M1-complex/M2-simple vs. (1) Same-context, (2) M1-simple/M2-baseline, & (3) M1-baseline/M2-simple; Context2: (3) M1-baseline/M2-simple vs. (1) Same-context, & (2) M1-simple/M2-baseline; Context3: (1) Same-context vs. (2) M1-simple/M2-baseline). The dependent measure is the number of words in the referring expression at test.

	Estimate	SE	*z*-value	*p*-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	2.621	0.047	55.56	<.0001	*Subject*	(Intercept)	0.048	0.219
Context1 (4) vs. (1),(2),(3)	-0.380	0.045	**-8.44**	**<.0001**		Context1	0.025	0.220
Context2 (3) vs. (1),(2)	0.192	0.041	**4.67**	**<.0001**		Context2	0.023	0.150
Context3 (1) vs. (2)	0.054	0.041	1.33	0.185		Context3	0.012	0.112
					Trial	(Intercept)	0.030	0.172
						Context1	0.060	0.244
						Context2	0.037	0.193
						Context3	0.057	0.239

Table A.14. *Experiment 4 test trials: Disfluency.* Mixed effect model with visual context as a fixed effect. (Context1: (4) M1-complex/M2-simple vs. (1) Same-context, (2) M1-simple/M2-baseline, & (3) M1-baseline/M2-simple; Context2: (3) M1-baseline/M2-simple vs. (1) Same-context, & (2) M1-simple/M2-baseline; Context3: (1) Same-context vs. (2) M1-simple/M2-baseline). The dependent measure (binary) is whether the expression was fluent or not.

	Estimate	SE	*z*-value	*p*-value			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	1.040	0.263	3.952	<.0001	*Subject*	(Intercept)	1.770	1.330
Context1 (4) vs. (1),(2),(3)	-1.701	0.240	**-7.076**	**<.0001**		Context1	0.546	0.739
Context2 (3) vs. (1),(2)	0.330	0.150	**2.187**	**0.028**		Context2	0.015	0.122
Context3 (1) vs. (2)	0.053	0.171	0.311	0.756		Context3	0.038	0.194
					Trial	(Intercept)	0.072	0.269
						Context1	0.273	0.523
						Context2	0.156	0.395
						Context3	0.033	0.180

APPENDIX B: Model results for Chapter 3

Table B.1. Experiment 5: Mixed effect model with entrainment type during the communication task as a fixed effect. The dependent measure is binary whether the expression on test trial was modified or not. (NC: Non-contrast, CN: Contrast-naming, CL: Contrast-location)

	Estimate	SE	*z*-value	Pr(>|z|)			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	-2.047	0.278	-7.362	<0.001	*Subject*	(Intercept)	1.371	1.171
Entrainment 1 (NC vs. CN & CL)	0.439	0.120	**2.194**	**0.028**		Entrainment 1	0.103	0.320
Entrainment 2 (CN vs. CL)	0.014	0.211	0.067	0.946		Entrainment 2	0.011	0.107
					Trial	(Intercept)	0.959	0.979
						Entrainment 1	0.009	0.094
						Entrainment 2	0.059	0.243

Table B.2. Experiment 5: Mixed effect model with role, referent, and entrainment type in memory test as fixed effects. The dependent measure is binary whether the response on memory test was correct or not.

	Estimate	SE	*z*-value	Pr(>|z|)			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	3.508	0.138	25.354	<.0001	*Subject*	(Intercept)	0.138	0.372
Role (Speaker vs. Listener)	0.749	0.269	**2.781**	**0.005**		Role	1.584	1.259
Referent (Target vs. Contrast)	-1.490	0.288	**-5.173**	**<.0001**		Referent	0.932	0.965
Entrainment 1 (NC vs. CN & CL)	0.227	0.189	1.199	0.231		Entrainment 1	0.191	0.437
Entrainment 2 (CN vs. CL)	0.237	0.214	1.106	0.269		Entrainment 2	0.147	0.383
Role*Referent	1.088	0.351	**3.099**	**0.002**	*Trial*	(Intercept)	0.360	0.560
Role*Entrainment 1	0.008	0.334	0.023	0.982		Role	0.017	0.130
Role*Entrainment 2	0.392	0.407	0.962	0.336		Referent	0.264	0.514
Referent*Entrainment 1	-0.471	0.351	-1.340	0.180		Entrainment 1	0.151	0.388
Referent*Entrainment 2	-0.479	0.409	-1.172	0.241		Entrainment 2	0.017	0.129
Role*Referent*Entrainment 1	-1.277	0.661	-1.930	0.054				
Role*Referent*Entrainment 2	-0.318	0810	-0.392	0.695				

Table B.3. Experiment 5: Mixed effect model with role and entrainment type for contrast item in memory test as fixed effects. The dependent measure is binary whether the response on memory test was correct or not.

	Estimate	SE	*z*-value	Pr(>|z|)			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	4.855	0.265	18.328	<.0001	*Subject*	(Intercept)	0.392	0.626
Role (Speaker vs. Listener)	-0.071	0.523	-0.136	0.892		Role	5.652	2.377
Entrainment 1 (NC vs. CN & CL)	0.092	0.454	0.202	0.840		Entrainment 1	0.761	0.872
Entrainment 2 (CN vs. CL)	0.358	0.422	0.849	0.396		Entrainment 2	1.064	1.032
Role*Entrainment 1	1.114	0.646	1.723	0.085	*Trial*	(Intercept)	0.626	0.791
Role*Entrainment 2	0.860	0.787	1.093	0.274		Role	0.028	0.167
						Entrainment 1	2.338	1.529
						Entrainment 2	0.097	0.312

Table B.4. Experiment 5: Mixed effect model with role and entrainment type for target item in memory test as fixed effects. The dependent measure is binary whether the response on memory test was correct or not.

	Estimate	SE	*z*-value	Pr(>|z|)			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	2.775	0.155	17.951	<.0001	*Subject*	(Intercept)	0.255	0.505
Role (Speaker vs. Listener)	1.450	0.258	**5.616**	**<.0001**		Role	1.043	1.021
Entrainment 1 (NC vs. CN & CL)	-0.126	0.207	-0.609	0.542		Entrainment 1	0.297	0.545
Entrainment 2 (CN vs. CL)	-0.018	0.213	-0.083	0.933		Entrainment 2	0.189	0.435
Role*Entrainment 1	-0.708	0.366	-1.934	0.053	*Trial*	(Intercept)	0.281	0.530
Role*Entrainment 2	0.155	0.400	0.388	0.698		Role	0.189	0.435
						Entrainment 1	0.023	0.151
						Entrainment 2	0.001	0.025

Table B.5. Experiment 6: Mixed effect model with entrainment type and lag during the communication task as fixed effects. The dependent measure is binary whether the expression on test trial was modified or not.

	Estimate	SE	*z*-value	Pr(>|z|)				Variance	Std.Dev.
Fixed					*Random*				
(intercept)	-1.869	0.264	-7.072	<.0001		*Subject*	(Intercept)	1.443	1.201
Lag (1 vs. 10 trial)	0.139	0.171	0.814	0.416			Lag	0.409	0.640
Entrainment 1 (NC vs. CNM & C)	0.388	0.170	**2.270**	**0.023**			Entrainment 1	0.185	0.430
Entrainment 2 (CNM vs. C)	0.038	0.187	0.203	0.839			Entrainment 2	0.261	0.511
Lag*Entrainment 1	-0.357	0.304	-1.175	0.240			Lag*Entrainment 1	0.041	0.203
Lag*Entrainment 2	-0.383	0.336	-1.143	0.251			Lag*Entrainment 2	0.167	1.091
						Trial	(Intercept)	1.190	1.091
							Lag	0.022	0.147
							Entrainment 1	0.060	0.246
							Entrainment 2	0.037	0.191
							Lag*Entrainment 1	0.024	0.154
							Lag*Entrainment 2	0.049	0.220

Table B.6. Experiment 6: Mixed effect model with role, referent, and entrainment type in memory test as fixed effects. The dependent measure is binary whether the response on memory test was correct or not.

	Estimate	SE	*z*-value	Pr(>|z|)			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	2.746	0.101	27.169	<.0001	*Subject*	(Intercept)	0.376	0.613
Role (Speaker vs. Listener)	0.577	0.138	**4.181**	**<.0001**		Role	0.730	0.854
Referent (Target vs. Contrast)	0.248	0.162	1.527	0.127		Referent	0.196	0.443
Entrainment 1 (NC vs. CN & CL)	-0.403	0.158	**-2.552**	**0.011**		Entrainment 1	0.147	0.384
Entrainment 2 (CN vs. CL)	0.473	0.129	**3.677**	**0.000**		Entrainment 2	0.314	0.561
Role*Referent	0.341	0.203	1.677	0.093	*Trial*	(Intercept)	0.186	0.431
Role*Entrainment 1	-0.366	0.202	-1.815	0.070		Role	0.139	0.373
Role*Entrainment 2	0.399	0.213	1.873	0.061		Referent	0.304	0.551
Referent*Entrainment 1	-0.103	0.304	-0.339	0.735		Entrainment 1	0.862	0.928
Referent*Entrainment 2	-0.543	0.225	**-2.413**	**0.016**		Entrainment 2	0.085	0.291
Role*Referent*Entrainment 1	0.439	0.399	1.100	0.271				
Role*Referent*Entrainment 2	-0.693	0.421	**-1.646**	**0.010**				

Table B.7. Experiment 6: Mixed effect model with role and entrainment type for contrast item in memory test as fixed effects. The dependent measure is binary whether the response on memory test was correct or not.

	Estimate	SE	*z*-value	Pr(>|z|)			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	2.756	0.141	19.587	<.0001	*Subject*	(Intercept)	0.551	0.742
Role	0.444	0.181	**2.456**	**0.014**		Role	1.084	1.041
Entrainment 1 (NC vs. CN & CL)	-0.222	0.261	-0.850	0.395		Entrainment 1	0.008	0.090
Entrainment 2 (CN vs. CL)	0.803	0.223	**3.605**	**0.000**		Entrainment 2	1.586	1.260
Role*Entrainment 1	-0.470	0.276	-1.707	0.088	*Trial*	(Intercept)	0.294	0.542
Role*Entrainment 2	1.017	0.309	**3.292**	**0.001**		Role	0.122	0.349
						Entrainment 1	1.619	1.272
						Entrainment 2	0.183	0.428

Table B.8. Experiment 6: Mixed effect model with role and entrainment type for target item in memory test as fixed effects. The dependent measure is binary whether the response on memory test was correct or not.

	Estimate	SE	*z*-value	Pr(>|z|)			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	2.955	0.121	24.338	<.0001	*Subject*	(Intercept)	0.342	0.584
Role	0.789	0.205	**3.842**	**0.000**		Role	1.328	1.152
Entrainment 1 (NC vs. CN & CL)	-0.328	0.185	-1.780	0.075		Entrainment 1	0.250	0.450
Entrainment 2 (CN vs. CL)	0.139	0.207	0.671	0.502		Entrainment 2	0.888	0.942
Role*Entrainment 1	-0.453	0.303	-1.495	0.135	*Trial*	(Intercept)	0.201	0.448
Role*Entrainment 2	0.053	0.321	0.167	0.867		Referent	0.244	0.493
						Entrainment 1	0.283	0.532
						Entrainment 2	0.249	0.499

Table B.9. Experiment 7: Mixed effect model with entrainment type during the communication task as a fixed effect. The dependent measure is binary whether the expression on test trial was modified or not.

	Estimate	SE	*z*-value	Pr(>|z|)			Variance	Std.Dev.
Fixed					*Random*			
(intercept)	-1.670	0.203	-8.356	<.0001	*Subject*	(Intercept)	0.526	0.725
Entrainment 1 (NC vs. CN & CL)	0.637	0.155	**4.116**	**<.0001**		Entrainment 1	0.059	0.243
Entrainment 2 (CN vs. CL)	0.277	0.154	1.798	0.072		Entrainment 2	0.003	0.054
					Trial	(Intercept)	0.885	0.940
						Entrainment 1	0.023	0.152
						Entrainment 2	0.001	0.023

Table B.10. Experiment 7 target memory: Mixed effect model with role, entrainment, and item type as fixed effects. The dependent measure is binary -- whether the response on the memory test was 'yes' (old) or 'no' (new).

	Estimate	SE	*z*-value	Pr(>|z|)			Variance	Std.Dev.
Fixed					*Random*			
(intercept) (response bias)	0.647	0.088	7.334	<.0001	*Subject*	(Intercept)	0.385	0.621
(effect on response bias)						Role	0.001	0.033
Role	-0.185	0.064	-2.888	0.004		Item	1.077	1.038
Entrainment 1 (NC vs. CN & CL)	0.116	0.086	1.346	0.178	*Trial*	(Intercept)	0.113	0.337
Entrainment 2 (CN vs. CL)	-0.042	0.085	-0.490	0.624		Role	0.003	0.052
Role*Entrainment 1	-0.048	0.142	-0.337	0.736		Item	0.726	0.852
Role*Entrainment 2	0.030	0.150	0.197	0.843				
Item (old vs. new) (sensitivity)	-2.802	0.162	**-17.267**	**<.0001**				
(effect on sensitivity)								
Item*Role	-0.381	0.129	**-2.961**	**0.003**				
Item*Entrainment 1	1.222	0.172	**7.113**	**<.0001**				
Item*Entrainment 2	-0.004	0.170	-0.021	0.983				
Item*Role*Entrainment 1	-0.167	0.284	-0.587	0.557				
Item*Role*Entrainment 2	-0.389	0.300	-1.297	0.195				

Table B.11. Experiment 7 contrast memory: Mixed effect model with role, entrainment, and item type as fixed effects. The dependent measure is binary -- whether the response on the memory test was ‘yes’ (old) or ‘no’ (new).

	Estimate	SE	*z*-value	Pr(>|z|)			Variance	Std.Dev.
Fixed					*Random*			
(intercept) (response bias)	0.491	0.088	5.564	<.0001	*Subject*	(Intercept)	0.349	0.590
(effect on response bias)						Role	0.011	0.103
Role	-0.099	0.069	-1.441	0.149		Item	1.056	1.028
Entrainment 1 (NC vs. CN & CL)	0.256	0.087	2.932	0.003	*Trial*	(Intercept)	0.194	0.440
Entrainment 2 (CN vs. CL)	-0.908	0.091	-9.948	<.0001		Role	0.017	0.130
Role*Entrainment 1	0.158	0.142	1.112	0.266		Item	0.528	0.726
Role*Entrainment 2	-0.446	0.161	-2.771	0.006				
Item (old vs. new) (sensitivity)	-2.994	0.166	**-18.018**	**<.0001**				
(effect on sensitivity)								
Item*Role	-0.477	0.136	**-3.503**	**<.001**				
Item*Entrainment 1	0.690	0.173	**3.964**	**<.0001**				
Item*Entrainment 2	-1.754	0.182	**-9.618**	**<.0001**				
Item*Role*Entrainment 1	0.580	0.283	**2.046**	**0.041**				
Item*Role*Entrainment 2	-0.550	0.322	-1.706	0.088				

Table B.12. Experiment 7 context memory: Mixed effect model with role, entrainment, and item type as fixed effects. The dependent measure is binary -- whether the response on the memory test was 'yes' (old) or 'no' (new).

	Estimate	SE	*z*-value	Pr(>|z|)			Variance	Std.Dev.
Fixed					*Random*			
(intercept) (response bias)	1.032	0.088	11.789	<.0001	*Subject*	(Intercept)	0.360	0.599
(effect on response bias)						Role	0.020	0.143
Role	0.287	0.082	3.518	<.001		Item	0.147	0.383
Entrainment (CN vs. CL)	-0.726	0.085	-8.508	<.0001	*Trial*	(Intercept)	0.098	0.313
Role*Entrainment	-0.380	0.153	-2.486	0.013		Role	0.062	0.249
						Item	0.488	0.699
Item (old vs. new) (sensitivity)	-1.726	0.131	**-13.189**	**<.0001**				
(effect on sensitivity)								
Item*Role	-0.028	0.161	-0.173	0.862				
Item*Entrainment	-1.540	0.168	**-9.162**	**<.0001**				
Item*Role*Entrainment	-0.249	0.306	-0.815	0.415				