Published in final edited form as:
Cogn Sci. 2008 June 1; 32(4): 643–684. doi:10.1080/03640210802066816.

Real-time investigation of referential domains in unscripted conversation: a targeted language game approach

Sarah Brown-Schmidt and **Michael K. Tanenhaus**
Department of Brain and Cognitive Sciences University of Rochester

Abstract

Two experiments examined the restriction of referential domains during unscripted conversation by analyzing the modification and on-line interpretation of referring expressions. Experiment 1 demonstrated that from the earliest moments of processing, addressees interpreted referring expressions with respect to referential domains constrained by the conversation. Analysis of eye movements during the conversation showed elimination of standard competition effects seen with scripted language. Results from Experiment 2 pinpointed two pragmatic factors responsible for restriction of the referential domains used by speakers to design referential expressions and demonstrated that the same factors predict whether addressees consider local competitors to be potential referents during on-line interpretation of the same expressions. These experiments demonstrate for the first time that on-line interpretation of referring expressions in conversation is facilitated by referential domains constrained by pragmatic factors which predict when addressees are likely to encounter temporary ambiguity in language processing.

Keywords

Eye-tracking; conversation; alignment; on-line; language processing; referential communication; cohort; point-of-disambiguation

1. Introduction

Most psycholinguistic research on spoken language comprehension can be divided into one of two traditions, each with its roots in seminal work from the 1960s (Clark, 1992; Trueswell & Tanenhaus, 2005), and each with its own characteristic theoretical concerns and dominant methodologies. The language-as-product tradition has its roots in George Miller's synthesis of the then emerging information processing paradigm and Chomsky's theory of transformational grammar (e.g. Miller, 1962; Miller & Chomsky, 1963). The product tradition emphasizes the individual cognitive processes by which listeners recover linguistic representations—the ‘products’ of language comprehension. Psycholinguistic research within the product tradition typically examines moment-by-moment processes in real-time language processing, using carefully controlled stimuli, scripted materials, and fine-grained on-line measures that are closely time-locked to the input.

Please direct correspondence to: Sarah Brown-Schmidt Beckman Institute 405 N Mathews Ave. University of Illinois at Urbana-Champaign Urbana, IL 61801 tel: (217) 244−4787 fax: (217) 333−2922 e-mail: brownsch@uiuc.edu.
Sarah Brown-Schmidt, Department of Brain and Cognitive Sciences, University of Rochester; Michael K. Tanenhaus, Department of Brain and Cognitive Sciences, University of Rochester.
Sarah Brown-Schmidt is now at the Department of Psychology, University of Illinois, Urbana-Champaign.

The motivation for on-line measures comes from two observations. The first is that speech unfolds over time as a series of rapidly changing acoustic events. The second is that listeners continuously integrate the input, making provisional commitments at multiple levels of representations (Marslen-Wilson, 1973; 1975; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). As a consequence, evaluating models of how linguistic representations are accessed, constructed and integrated requires data that can only be obtained from response measures that are closely time-locked to the input and sensitive to how the listener's representations change as the input unfolds in time.

One case-in-point is temporary ambiguity. One of the consequences of the combination of sequential input and continuous processing is that listeners are continuously faced with resolving temporary ambiguity at multiple levels of representation. For example, the initial sounds of *clown* are briefly consistent with both *cloud* and *clown.* Response measures that are closely time-locked to the input have revealed that as a listener hears *clown*, both *cloud* and *clown* are briefly activated, with activation to *cloud* decreasing as soon as coarticulatory information in the vowel becomes more consistent with *clown* (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Zwitserlood, 1989). Similarly, in a context that includes both a large silver fork and a large silver spoon, a listener hearing *the large silver spoon* will actively consider both as potential referents until the disambiguating sounds at the onset of *spoon* (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Hanna, Tanenhaus, & Trueswell, 2003). Response measures that are not closely time-locked to the input are of limited value for examining processes like these, which are central to language processing.

The language-as-action tradition has its roots in work by the Oxford philosophers of language use, e.g., Austin, (1962), Grice (1957) and Searle (1969), and work on conversational analysis, e.g., Schegloff and Sachs (1973). The action tradition focuses on how people use language to perform acts in conversation--the most basic form of language use. Psycholinguistic research within the action tradition typically examines unscripted interactive conversation involving two or more participants engaged in a cooperative task, typically with real-world referents and well-defined behavioral goals.

One reason is that many aspects of utterances in a conversation can only be understood with respect to the context of the language use, which includes the time, place and participant's conversational goals, as well as the collaborative processes that are intrinsic to conversation. Moreover, many characteristic features of conversation emerge only when interlocutors have joint goals and when they participate in a dialogue both as a speaker and an addressee.

Detailed analyses of participants' linguistic behavior and actions in cooperative tasks have provided important insights into how interlocutors track information to achieve successful communication. They demonstrate that many aspects of communication, establishing successful reference, for instance, are not simply individual cognitive processes; they are achieved as the result of coordinated actions among two or more individuals across multiple linguistic and non-linguistic exchanges (Bangerter, 2004; Clark & Wilkes-Gibbs, 1986; Clark & Krych, 2004; Gergle, Kraut, & Fussell, 2004; Schober & Brennan, 2003).

The coordination that emerges through participation in a conversation appears to change how language is understood in a conversation. One way to quantify the contribution of this coordination, while controlling for the contents of the language itself, is to compare the understanding of participants in a conversation, with the understanding of an 'overhearer'. For example, in Schober and Clark (1989), one participant, the 'director', told the other participant, the 'matcher', to arrange a set of twelve abstract tangram figures in a particular order. The task was repeated using the same twelve figures over a series of trials. The critical comparison was how well an overhearer would perform on the matching task; if overhearers performed as well

as matchers, this would indicate that the ability to interact and coordinate with the director was irrelevant to the process of understanding. In their first experiment, the overhearers listened to the conversations on audio-tape, and in the second experiment, the overhearers listened to the instructions while seated in the same room as the matcher and director. The results of both experiments were striking: Accuracy at placing the figures was significantly worse for overhearers, regardless of whether they listened to the conversation at their own pace on an audio-tape or to a live conversation in the same room. This suggests that the act of participating in the conversation affects the mental representations used to understand language, thus the language itself is not the only contribution to understanding.

A second example of the way in which conversation shapes language processes comes from analyses of the modification of referring expressions. Felicitous use of a definite referring expression requires that it uniquely identify its intended referent (Roberts, 2003). Referents are identified with respect to a specific domain, thus speakers must take into account both the properties of the referent and the relevant context when generating the referring expression (Olson, 1970). For example, consider a scenario in which Duane, who is dining with a friend at a restaurant, wants a glass of red wine, but the bottle is out of reach. With only a single bottle on the table, Duane could ask his friend *Please pass the wine*, using the definite referring expression, *the wine*. However, with two open bottles, one red and one white, he would need to use a more specific referring expression, such as *Please pass the red*. But he would not need to take into consideration other bottles of wine in the restaurant, including any bottles that might happen to be on other tables within his companion's reach. In this context, those bottles are not possible referents for *the wine* or *the red* because they are outside the set of potential referents, also called the **referential domain**.

Identifying the contents of referential domains, and understanding how speakers determine whether an entity is or is not in the referential domain has primarily been done through analyses of the speaker's modification patterns in combination with analyses of the discourse and broader context. For example, in work on machine-generation of referring expressions, Salmon-Alt and Romary (2000) use a subset of the global context in order to generate contextually appropriate expressions (also see Salmon-Alt, 2000; Landragin & Romary, 2003). The way the referential domain (e.g. the subset of the global context) is identified is based on human dialog, and uses factors such as the perceptual environment as well as gestures and the discourse history. In analyses of conversations during a construction task, Beun and Cremers (1998) found that factors including a spatial locus of attention (also see Grosz, 1977; Thórisson, 1994; Glenberg, Meyer, & Lindem, 1987; Rinck & Bower, 1995; Morrow, Bower and Greenspan, 1989), as well as information about the task predicted the speaker's pattern of modification, suggesting that these factors determined which entities were in the referential domain. Krahmer and Theune (2002) extended Beun and Cremer's (1998) results for use in a natural language generation system (also see Dale & Reiter, 1995; Kim, Hill, & Traum, 2005), which uses the proximity of potential referents to the last mentioned referent as a metric to predict likelihood of mention. These findings demonstrate that understanding how speakers construct their referring expressions in conversation will require analyzing both the language itself, as well as its context of use.

Recently, the language processing community has begun to show increased interest in bridging the product and action traditions (Pickering & Garrod, 2004; Trueswell & Tanenhaus, 2005). However, research that aims to bridge the two traditions has not traditionally combined on-line measures--the methodological cornerstone of the product tradition, with unscripted cooperative conversation--the central domain of inquiry in the action tradition (but cf. Brennan, 2005; Kraljic & Brennan, 2005). The research presented here was aimed at bridging these two traditions.

We designed two experiments in which on-line measures were combined with unscripted conversation to address the role of conversational processes in the on-line interpretation of referring expressions. Specifically, our experiments addressed the following two questions: First, how does the process of on-line ambiguity resolution for the addressee in a conversation compare to ambiguity resolution processes for language outside the context of conversation? While results from off-line experiments indicate that conversation serves to increase the efficiency of communication through increased shared knowledge (Schober & Clark, 1989), it is unknown if this information is available to on-line comprehension processes, and even if it is, if these benefits extend to early speech interpretation and ambiguity resolution processes, or only later post-lexical processes. One potential mechanism for increased efficiency of language processing in conversation is through the coordination of referential domains. Off-line analyses of conversations show that speakers modify referential expressions only with respect to those entities that are salient and task-relevant (Beun & Cremers, 1998). By limiting the number of potential discourse referents, constrained referential domains have the potential to eliminate multiple potential sources of temporary ambiguity for the addressee. Thus our second question is: Do referential domains constrain the on-line interpretation of referential expressions for addressees, and if so, do addressees use referential domains that are similar to the referential domains that the speaker used to construct these expressions? If the referential domains of the addressee and the speaker differ, we would expect to find that the addressee considers entities outside the speaker's referential domain to be potential referents during on-line interpretation. This result is also expected if the information that would be used to identify the referential domain is unavailable on-line or is only used at later stages of language processing (e.g. Keysar, et al. 1998).

In order to examine these questions, we monitored gaze and speech as pairs of naïve participants engaged in a referential communication task (Krauss & Weinheimer, 1966) to match the position of blocks on their respective game-boards. We adopted a “targeted language games” methodology in which the task was designed to generate sufficient trials in the conditions of interest to approximate a standard within-subjects factorial design, including control conditions, but without explicitly restricting what participants could say.

This class of dialogue has some clear benefits for examining real-time language processing in conversation. Most of the language is task-oriented. Moreover, the referential world, and the goals of the interlocutors are well defined. In addition, the referential communication task is closely related to task-oriented, or practical dialogues (Allen et al., 2001). Practical dialogue is one of the domains for which computational linguists are developing the most explicit models, instantiated as end-to-end dialogue systems in which human users interact with a system using unrestricted spoken language. This makes it a potential test bed for creating and evaluating explicit models of dialogue, potentially leading to a feedback loop between computational and experimental investigations (see Aist, Campana, et al., 2005).

Experiments 1 and 2 used a targeted language game approach to examine the referential domains used in the production and interpretation of temporarily ambiguous referring expressions. Experiment 1 examines these questions using ambiguities at the lexical level; Experiment 2 focuses on ambiguities at the phrasal level.

2. Experiment 1

Experiment 1 examined the time-course of interpreting expressions such as *the cloud* during a conversation compared to expressions produced outside the context of a conversation. Previous work by Allopenna, Magnuson, and Tanenhaus (1998) used eye movements to monitor the on-line interpretation of the same kind of expressions. They monitored the fixations that participants made while following pre-scripted instructions such as *Pick up the beaker*, in

contexts that included pictures of a beaker, a beetle, a speaker and a carriage. Previous work using eye movements and spoken instructions to manipulate objects in a co-present 'visual world' (e.g. Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberahrd & Sedivy, 1995) shows that soon after the onset of the object's name, listeners begin to fixate that object. Allopenna, et al. (1998) used this technique to test whether participants would briefly consider both the target and the cohort competitor to be potential referents as the interpreted the object name. They found that stimulus-driven fixations to the target object began as early as 200 ms after the onset of the noun. Crucially, eye movements launched at this point in the speech stream were equally likely to be directed to the eventual referent as other objects with names that were temporarily consistent with the speech signal, such as *beetle*. More recent work has demonstrated that looks to these cohort competitors are reduced or eliminated when the relevant referential context makes the cohort an implausible referent (Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Dahan & Tanenhaus, 2004).

Here, we examined whether looks to cohort competitors are reduced or eliminated when the relevant referential context established by unscripted task-based conversation makes the cohort competitor an implausible referent. Based on the previous findings by Schober and Clark (1989) that demonstrated improved understanding of language when participating in a conversation, we asked whether participating in the conversation could affect the on-line interpretation of a single word. In order to answer this question, we monitored the eye movements that participants made as they interpreted referring expressions like *the cloud* in contexts that included both a cloud and a clown. We compared the interpretation of expressions made by the conversational partner during the course of the conversation, with similar expressions made by the experimenter outside the context of the conversation.

2.1. Method

2.1.1. Participants—Twelve pairs of undergraduates from the Rochester, New York area participated in this experiment. Participants were paid $7.50 an hour. All twenty-four participants were native English speakers and had no known history of speech or hearing impairments. Additionally, each pair identified themselves as friends. The data from four participants were eliminated from analysis because of equipment failure, leaving the data from 20 participants for analysis.

2.1.2. Materials—The task we designed was a variant of a referential communication task (Krauss & Weinheimer, 1966). Pairs of participants worked together to arrange a set of Duplo™ blocks in a matching pattern. Partners were separated by a curtain and seated in front of a board with stickers and a resource area with blocks, as illustrated in Figures 1a-b. Unlike the standard referential communication task, each participant was both a director and a matcher.

Boards were divided into five distinct sub-areas, with 57 stickers representing the blocks. Stickers were divided between the boards; where one partner had a sticker, the other had an empty spot. Thirty-six blocks were assorted colored squares and rectangles. Twenty-one additional blocks had pictures of common objects, which included ten pairs of cohort competitors with names beginning with the same sounds (e.g., cloud/ clown)[1]. On the eye-tracked partner's board, the 21st block was a lobster, and was not a member of a cohort pair. This block differed from the other picture blocks in that the non eye-tracked partner did not have a lobster block, and was instead required to use a rat block as a placeholder. We used the lobster/ rat block in order to examine the naming conventions that partners would adopt, expecting that partners would agree on a name for that block, such as the "lobster" or "rat-lobster". However, many of the pairs joked about this block to the extent that the data, while

[1]Due to experimenter error, one pair had two additional blocks on their board, both of which had pictures of a comb.

amusing, were not informative. A complete list of blocks is included in Appendix A (all appendix materials are available for viewing at the following web address: http://www.cogsci.rpi.edu/CSJarchive/Supplemental/index.html).

Seven pairs of cohort blocks were placed in the same sub-area, the remaining three were placed in adjacent sub-areas. All cohort pairs were separated by approximately 8 cm. Each block subtended approximately 2−3° of visual angle when participants were seated at a comfortable distance from the workspace. Each picture was selected to be easily nameable. Twenty of the pictures were from a full-color version (Rossion & Pourtois, 2004) of a large corpus of pictures, normed for name agreement (Snodgrass & Vanderwart, 1980); the remaining picture was a similar clip-art image.

2.1.3. Procedure—The participants' task was to replace each sticker with a matching block, and instruct their partner to place a block in the same location to make their boards match. The positions of the stickers were determined by the experimenter, allowing for experimental control over the layout of the board. The shapes of the sub-areas and the initial placement of the stickers were designed to create conditions where the proximity of the blocks and the constraints of the task were likely to influence the strategies adopted by the participants. However, the initial placement of stickers and the non-standard shape of the sub-areas did not easily lend the game to simple spatial strategies such as working from left to right or top to bottom within a sub-area. When participants finished placing the blocks, we asked them to confirm the placement of each block with their partner. No other restrictions were placed on the interactions, which lasted approximately two hours.

In addition to monitoring the eye-tracked participant's interpretation of her partner's referring expressions, we also monitored the eye-tracked participant's interpretation of references made by the experimenter during periodic calibration checks. These calibration checks were added to create an opportunity for the experimenter to ask the eye-tracked participant to look at different objects on the board when the eye-track was already accurate (a fact we did not share with the participants).

Experimenter-generated trials: The experimenter performed two to three calibration checks per participant. Additional calibration checks were performed whenever the track needed to be corrected (usually one to two times per participant). For each calibration check, the experimenter (e.g. the first author) interrupted the conversation to announce that she needed to "check the track". She then asked the participant wearing the tracker to look at between five and ten of the picture blocks on the board, e.g., *Look at the clown, ok, now look at the penguin.* The experimenter needed to generate the instructions for each calibration check impromptu because there was no way to know in advance which blocks would be on the workspace and which would have cohort competitors. If, during a calibration check, the experimenter discovered the track needed to be corrected, the experimenter re-calibrated the eye-tracker and any affected trials were excluded from analysis.

Participant-generated trials: Our analysis of trials on which the non eye-tracked participant was speaking focused on the interpretation of definite references to objects paired with cohort competitors such as *the clown*, *that penguin*, *your snake*, and *snail*. Our analysis did not include references with indefinite articles such as *a clown* or pronouns such as *it* or *that*. We focused exclusively on definite referring expressions to increase the homogeneity of the dataset. Moreover, indefinites were typically used only for the first mention of a block, thus the referent was not on the board. We excluded trials where the speaker used an alternate name for a block. For example, we would exclude a trial in which the participant said *the writing instrument* instead of the more commonly used term, *the pencil*, because the alternative name no longer formed a cohort pair with the intended competitor, in this case a block with a picture of a

penguin. Like the analysis of experimenter-generated trials, trials with a poor eye-track (e.g. a majority of the trial was track-loss) were excluded from analysis. An example excerpt from one conversations is presented in (1) below:

(1) 1. okay... allright, so then I have ay..a blue block that's vertical...

2. uh, big one?

1. a big- blue.. block.. else it wouldn't matter um, (laughs) it goes over us here...so it's right above the snail?

2. uh-huh

1. but over one.. to the left

2. to the left

1. so it's kind of

2. I got it..

2.1.4. Equipment and Analysis—We recorded the speech of both partners and the eye movements of one partner. Halfway through the task, the eye-tracker was switched to the other partner in order to obtain eye-tracking data from both participants. Eye movements were monitored using an ISCAN visor-mounted system. The image of the eye-tracked partner's board with eye position superimposed, and the entire conversation (both participants' voices), was recorded using a frame-accurate video recorder. We coded eye movements made by addressees following definite references to objects paired with cohort competitors, and compared references made by the addressee's partner, with the experimenter's instructions to look at the same blocks.

2.2. Predictions

Given previous results by Schober and Clark (1989) and Beun and Cremers (1998), we hypothesized that the coordination gained by interacting in a conversation would facilitate interpretation of expressions like *the cloud* by constraining the referential domain to a small area of the board. If the speaker and the addressee's referential domains are constrained to a small area and coordinated, and addressees use referential domains to constrain on-line interpretation of referring expressions, then we would expect that addressees would be less likely to consider the cohort competitor to be a potential referent when interpreting expressions during the conversation compared to outside the conversation.

The predictions for the experimenter-generated referring expressions are straightforward. Experimenter-generated references should elicit the same pattern of looks as those typically observed in experiments using scripted instructions and displays limited to a small number of potential referents (Allopenna et al., 1998; Spivey-Knowlton, 1997). In a baseline region, before the referent is named, looks to the referent, the cohort competitor and blocks with unrelated names should be equivalent. During the portion of the referring expression where the name is phonetically consistent with both the referent and the competitor, looks to both should increase relative to unrelated distracters. After the referent is disambiguated by the phonetic input, looks to the referent should rise, and looks to the competitor should fall.

If our hypothesis that during conversation, referring expressions are interpreted with respect to conversationally constrained referential domains is correct, when addressees interpret referring expressions made by their partner, fixations to the referent, but not the competitor should increase during the ambiguous region of the referring expression. This pattern of results would indicate that (a) addressees can use referential domains to constrain on-line

interpretation, and (b) that the addressee's referential domain included the referent but not the competitor. In order for this result to obtain, the referential domains of the speaker and the addressee would have to be restricted to a small region of the workspace and closely coordinated. Alternatively, the pattern may be no different than that observed for the experimenter-generated utterances, with increased looks to both targets and competitors. This result is expected if, during the several hundred milliseconds when the expression is consistent with both the referent and the potential competitor, the addressee used a referential domain that included both the target and cohort. The fact that our task is complicated, provides a large amount of room for error, and that participants start out with different sets of stickers on their boards makes uncoordinated referential domains a real possibility. Competition effects would also be expected if participants do manage to coordinate referential domains, but these coordinated representations are not reliably used during on-line comprehension. Lastly, we might not see any systematic relationship between the fixations and the referring expression. This pattern would occur if data from tasks like these were too noisy to result in patterns of looks that are time-locked to words in spontaneously generated utterances.

2.3 Analysis and Results

Only trials with both the target block and the cohort competitor on the board were analyzed. “Other” blocks were carefully selected to serve as a baseline with which to compare fixations to cohort competitors. For each critical trial, the location of the cohort competitor relative to the target was identified. A second trial was then identified which had the same spatial relationship between the target block for that trial, and a picture block which was not the cohort competitor for that trial. This picture block was designated as the “other” block for the first trial. In the following analyses, the proportion of fixations to cohort competitor and other blocks are directly compared to establish whether cohort blocks were more likely to be fixated than would be expected based simply on the spatial relationship between target and cohort.

Eye movements associated with cohort references were analyzed in three 400 ms time regions relative to the onset of the target word. The first region captured fixations from 200 ms before until 200 ms after the onset of the head noun. Signal-driven fixations are not expected until 200 ms following word onset due to the time needed to program and launch an eye movement (Matin, Shao & Boff, 1993), thus this region is used as a baseline. The second window captured fixations made between 200 ms and 600 ms following the onset of the head noun; the third window captured the following 400 ms.

Figures 2 and 3 show the proportion of fixations to target, cohort competitor and other blocks for trials in which the non eye-tracked participant was speaking and the experimenter was speaking, respectively. Looks to cohort competitors increase after the onset of the target noun for trials in which the experimenter was speaking but not for trials in which the non eye-tracked participant was speaking. Looks to target blocks increase across the three windows for both types of trials, however this effect is more apparent for experimenter-speaking trials.

Separate, planned ANOVAs compared the proportion of fixations to targets, cohorts and other blocks for trials during which the experimenter and the non eye-tracked participant were speaking. Analyses by item were modeled on the experimental design used by Brennan (1995) where distinct video sequences were categorized as individual items. Here, we treat each individual block as an individual item. Because the pattern of stickers (and thus the blocks) was identical across participants, this approximates standard treatment of items. The Huynh-Feldt adjustment to degrees of freedom was applied when sphericity assumptions were not met; for clarity the unadjusted *F*-statistics and *df* are reported with the Huynh-Feldt epsilon (H-F ε) and adjusted *p*-value. We report generalized eta-squared (η^2_G; Bakeman, 2005, Olejnik and Algina, 2003) as a measure of effect size for analysis of variance (ANOVA) results. For *t*-tests, we report Cohen's *d* (Cohen, 1998).

2.3.1. Experimenter-generated trials—For experimenter trials, the ANOVA revealed a significant main effect of object type (target, competitor, other), H-F ε = .76, *F1*(2,38) = 57.60, *p*<.0001, η^2_G = .46; H-F ε = .70, *F2*(2,38) = 75.15, *p*<.0001, η^2_G = .59, due to significantly more looks to targets than either cohorts or other blocks (*p*'s<.0001), and an equal number of looks to cohort and other blocks (*p*'s>.1). There was also a main effect of time region, H-F ε = .76, *F1*(2,38) = 158.05, *p*<.0001, η^2_G =.55; *F2*(2,38) = 375.77, *p*<.0001, η^2_G =.66. Crucially, the interaction between object type and region was also significant, H-F ε = .43, *F1*(4,76) = 82.52, *p*<.0001, η^2_G =.61; H-F ε = .39, *F2*(4,76) = 115.01, *p*<.0001, η^2_G =.73. The interaction was explored with a planned ANOVA at each time region.

At the baseline region, we observed a significant effect of object type, *F1*(2,38) = 4.48, *p*<.05, η^2_G =.10; *F2*(2,38) = 4.96, *p*<.05, η^2_G =.14. Planned t-tests indicated that listeners fixated other blocks more than targets, *t1*(19) = 2.81, *p*<.05, *d*=.79; *t2*(19) = 3.13, *p*<.01, *d*=.79, and marginally more than cohorts, *t1*(19) = 1.73, *p*=.10, *d*=.36; *t2*(19) = 1.86, *p*=.08, *d*=.54, whereas looks to targets and cohorts were equivalent, *t1*(19) = 1.26, *p*=.22, *t2*(19) = 1.01, *p*=.32.

During the region beginning with the onset of the head noun, the effect of object type was also significant, *F1*(2,38) = 4.13, *p*<.05, η^2_G =.10; H-F ε = .80, *F2*(2,38) = 5.55, *p*<.05, η^2_G =.15. Here, looks to cohorts and targets were equivalent, *t1*(19) = .36, *p*=.72, *t2*(19) = .30, *p*=.77, but there were significantly more looks to cohorts and targets than other blocks; cohorts vs. other: *t1*(19) = 2.22, *p*<.05, *d*=.70; *t2*(19) = 3.50, *p*<.01, *d*=.77; targets vs. other: *t1*(19) = 2.54, *p*<.05, *d*=.81; *t2*(19) = 3.29, *p*<.01, *d*=1.08.

The object type effect was also significant at the final region, H-F ε = .64, *F1*(2,38) = 88.38, *p*<.0001, η^2_G =.76; H-F ε = .59, *F2*(2,38) = 120.13, *p*<.0001, η^2_G =.85, and was due to significantly more looks to targets than cohorts, *t1*(19) = 8.77, *p*<.0001, *d*=2.88; *t2*(19) = 9.47, *p*<.0001, *d*=3.85, and other blocks, *t1*(19) = 10.78, *p*<.0001, *d*=3.48; *t2*(19) = 14.91, *p*<.0001, *d*=5.07. Looks to cohorts were also significantly higher than other blocks at this region, *t1*(19) = 3.02, *p*<.01, *d*=.87; *t2*(19) = 2.49, *p*<.05, *d*=.66.

2.3.2. Participant generated trials—A different pattern of results obtained when eye-tracked participants interpreted references made by their non eye-tracked partners. An ANOVA for fixations to target, cohort and other blocks revealed a main effect of object type, H-F ε = .54, *F1*(2,38) = 66.18, *p*<.0001, η^2_G =.65; H-F ε = .67, *F2*(2,38) = 284.54, *p*<.0001, η^2_G =.85, and an effect of time region that was marginal in the items analysis, *F1*(2,38) = 4.75, *p*<.05, η^2_G =.01; *F2*(2,38) = 2.62, *p*=.09, η^2_G =.02. The main effect of object type was due to significantly more looks to targets than either cohorts or other blocks (*p's*<.0001), whereas looks to cohorts and other blocks were equivalent (*p*'s>.3). The main effects were qualified by a marginal interaction, H-F ε = .61, *F1*(4,76) = 2.08, *p*=.13, η^2_G =.01; H-F ε = .50, *F2*(4,76) = 2.82, *p*=.08, η^2_G =.03. Planned ANOVAs revealed a main effect of object type at each time region: baseline region, H-F ε = .51, *F1*(2,38) = 41.67, *p*<.0001, η^2_G =.59; H-F ε = .71, *F2*(2,38) = 197.00, *p*<.0001, η^2_G =.87; region beginning with the onset of the noun, H-F ε = .56, *F1*(2,38) = 54.92, *p*<.0001, η^2_G =.66; H-F ε = .67, *F2*(2,38) = 199.57, *p*<.0001, η^2_G =.86; final region, H-F ε = .59, *F1*(2,38) = 65.18, *p*<.0001, η^2_G =.70; H-F ε = .62, *F2*(2,38) = 159.90, *p*<.0001, η^2_G =.84. Each of these effects was due to significantly more looks to targets than either cohorts or other blocks (*p*'s <.0001). There were no reliable differences between cohorts and other blocks at either the baseline region or the noun region (*p*'s >.14), and at the late region there was a non-significant preference for cohorts over other blocks, *t1*(19) = 1.43, *p*=.17; *t2*(19) = 1.90, *p*=.07.

The fact that targets were preferred at each of the time regions, with no reliable difference at any region between cohort and other blocks might suggest that fixations for partner speaking trials are not associated with processing of the noun phrase. In order to examine this possibility,

we compared the proportion of fixations to targets across the three time regions. If the addressee's gaze is not sensitive to the noun phrase itself, we should find that fixations to the target are equivalent at each region. In contrast, if certainty about the target increases as the name of the referent unfolds, then we should find an increase in fixations to the target across the three time regions. Crucially, the increase in target fixations from the baseline to the final region was significant, one-tailed $t1(19) = 2.51$, $p<.05$, $d=.38$; $t2(19) = 1.99$, $p<.05$, $d=.55$, demonstrating that addressees' gaze was sensitive to the linguistic input.

A separate analysis, which only included those trials on which the eye-tracked participant ultimately fixated the target, yielded an identical pattern of results. This analysis confirms that the observed differences in results for experimenter and participant-generated trials is not simply due to the fact that participants were asked to fixate the target during experimenter-generated trials.

Lastly, we examined the pattern of eye movements for trials on which the addressee was not fixating the target at the onset of the critical word to see if addressees would still converge on the target without considering the cohort competitor in cases where their attention had not already been attracted to the target before the referring expression. The results of this analysis yielded a similar pattern of results to our previous analyses, with the exception that the baseline preference to fixate the target in the participant-generated trials was reduced (see Appendix B).

2.4. Discussion

The aim of Experiment 1 was to test the on-line implications of previous work which indicates that the coordination which emerges during conversation improves language understanding. We compared the on-line interpretation of expressions like *the cloud* made during a conversation with expressions made outside the context of a conversation. While the linguistic signal (e.g. the noun phrase) was virtually identical in the two cases, the context of use changed how these words were understood. Outside the constrains of the conversation, addressees interpreted these expressions with respect to each of the entities in the global context (Salmon-Alt & Romary, 2000), showing the standard lexical competition effect. However, during the conversation, the same expressions were interpreted entirely differently--addressees did not fixate these same competitors. These results allow us to make two important conclusions. First, addressees use a subset of the global context to constrain on-line interpretation during conversation. This suggests that referential domains established by task-oriented conversation are relevant to on-line interpretation, and sets the stage for future investigations of how these referential domains interact with factors such as discourse structure and prosody (e.g. Dahan, et al. 2004; Arnold et al., 2004). Second, this work demonstrates that it is possible to examine on-line interpretation processes during unscripted conversation.

In Experiment 2, we examine whether the effects of constrained referential domains are limited to lexical competition or whether referential domains also constrain interpretation for longer-lasting phrasal ambiguities. More importantly, we focus on specific factors that influence circumscription of referential domains for speakers, and ask whether addressees are sensitive to the same factors as they interpret these expressions.

3. Experiment 2

The goal of this experiment was to use the combination of the speaker's utterances and the addressee's eye movements to determine how interlocutors circumscribe their referential domains. We examine both the speaker's and the addressee's referential domains by comparing the production and interpretation of definite references to color blocks, such as *the green horizontal block*. Unlike references to picture blocks with cohort competitors where the head

noun uniquely specified the intended referent (e.g. *penguin* uniquely referred to the single penguin block on the board), when referring to a color block, we expected that speakers would modify their expressions to distinguish the target from other color blocks that the speaker considered to be potential referents. Thus, the speaker's modification patterns will give us insight into which blocks he considers to be within the referential domain. A concurrent examination of the addressee's eye fixations as she interprets these referring expressions will reveal which entities she considers to be potential referents, allowing us to test whether addressees circumscribe their referential domains in a similar way. Specifically, this experiment was designed to test two hypotheses:

The first hypothesis was that speakers should modify their referential expressions with respect to referential domains constrained by three linguistic/ pragmatic factors: (1) proximity to the last mentioned block, with proximal blocks being more likely to be in the referential domain than less proximal ones. (2) relevance to the task; with blocks more relevant to the current task being more likely to be in the referential domain. (3) recency of mention in the discourse, with recently mentioned blocks more likely to be in the referential domain. Constraints similar to these have been identified in previous work using referential communication tasks (Beun & Cremers, 1998), as well as research on the mental models used in text processing (Morrow, Bower, & Greenspan, 1989). Because definite referring expressions must uniquely identify a referent with respect to a contextually defined referential domain (Olson, 1970), the speaker's modification pattern should indicate whether she considers a potential competitor block to be in the referential domain during utterance planning. For example, if the global context contained a green vertical rectangle, a green horizontal rectangle, and several blue blocks, and the speaker used the expression *the green horizontal rectangle*, we can infer that both rectangles were in the speaker's referential domain. In contrast, if the speaker used the expression *the green rectangle* to refer to the green horizontal rectangle, we could infer that the green vertical rectangle was not in the referential domain.

The second hypothesis motivating Experiment 2 was that during on-line interpretation of referring expressions, addressees use referential domains that are similar to the referential domains that speakers use to produce these expressions. If this hypothesis is correct, addressees should temporarily consider competitor blocks to be potential referents in the same situations speakers consider these competitors to be potential referents. Previous work using the visual world eye-tracking technique and sentences like *Point to the green horizontal rectangle* in contexts like the one above demonstrates that listeners initially fixate the potential referents that match the referring expression with equal likelihood until the point in the referring expression that uniquely identifies the referent (e.g., Eberhard, et al., 1995). In our example, we could expect listeners, upon hearing the onset of the word *green*, to fixate the two green rectangles with equal likelihood until the **point-of-disambiguation** at the word *horizontal*, at which point looks to the vertical rectangle would taper off and looks to the horizontal rectangle would continue to rise. For the purposes of the current experiment, if addressees tend to initially fixate both potential referents following the onset of the noun phrase, we can conclude that both entities were in the addressee's referential domain. In contrast, if at the onset of the noun phrase the addressee began to fixate the green horizontal rectangle but never fixated the green vertical rectangle, this would indicate that while the addressee was interpreting the words *the green*, she did not consider the green vertical rectangle to be in the referential domain.

Previous work using the visual world eye-tracking technique and scripted utterances has found evidence for the use of contextually constrained referential domains during interpretation of ambiguous referring expressions. For example, Chambers, Tanenhaus, Eberhard, Filip and Carlson (2002) gave participants instructions like *Put the cube inside the can*, in contexts that included a cube, two cans (one big, one small) and several unrelated objects. They manipulated whether the cube was small enough to fit in either can, or so large that it could only fit in one

of the cans. They found that interpretation of *the can* was constrained by whether the size of the block made one of the cans an implausible referent: In the large block condition, fixations to the target can were earlier and there were few looks to the competitor can. In contrast, when the cube would fit in either can, participants looked about equally to the two cans. This result suggests that listeners can use the lexical-semantic constraints of the instruction *put...inside* in combination with non-linguistic information about the properties of the entities in the global context to eliminate incompatible referents when interpreting the words *the can*.

While the results from Experiment 1 suggest that addressees will interpret referential expressions with respect to constrained domains, the possibility for misalignment of referential domains is greater in Experiment 2 because of the large number of competitor blocks in the global context in comparison to the single competitor we used in Experiment 1. If the addressee's referential domain contains any competitors that are not included in the speaker's referential domain, then, from the addressee's perspective, the referential expression will not uniquely specify a referent and should engender confusion.

3.1. Method

The targeted language game used in Experiment 2 was the same as that used in Experiment 1. Minor changes in the design are noted.

3.1.1. Participants—Twelve pairs of participants who were undergraduates in the greater Rochester, NY community participated in this experiment. None of the participants had participated in Experiment 1. All participants identified themselves as native speakers of North American English, and each pair identified themselves as friends.

3.1.2. Materials—The materials used in Experiment 2 were identical to those used in Experiment 1, with the following exceptions. Boards were divided into the same five distinct sub-areas, with 56 stickers representing the blocks. Thirty-seven blocks were assorted colored squares and rectangles. Nineteen additional blocks had pictures of common objects, including six cohort competitors with names beginning with the same sounds (e.g., cloud/ clown). Four pairs of cohort blocks were placed in the same sub-area, the remaining two were placed in adjacent sub-areas. A single block (lobster) was not shared by participants and was replaced by a place-holder block (rat) on the partner's board. Cohort pairs were separated by about 8 cm. All nineteen pictures were selected to be easily nameable. Seventeen of the pictures were from a full-color version (Rossion & Pourtois, 2004) of a large corpus of normed pictures (Snodgrass & Vanderwart, 1980), the remaining two were similar clip-art pictures.

3.1.3. Procedure—We recorded the eye movements of one partner and the speech of both. Eye movements were monitored using an ISCAN visor-mounted system. The image of the eye-tracked partner's board with eye position superimposed, and the entire conversation (both participants' voices), was recorded using a frame-accurate video recorder. Unlike Experiment 1, the eye-tracker was not switched to the second participant during the task. In addition, the experimenter only interrupted the task when re-calibration was necessary.

3.2 Results

The formal analysis focuses exclusively on the interpretation of references to the color blocks, however a preliminary analysis of references to cohort blocks by the non eye-tracked partner yielded a pattern of results identical to that seen in Experiment 1.

3.2.1. Specificity of referential expressions—Non eye-tracked partners generated 1467 definite references to color blocks. This figure does not include references to blocks that were not on the board, plural, indefinite or pronominal references, and references that occurred

within interrogative utterances. These references were excluded in order to increase the homogeneity of the dataset. Because speakers typically used indefinite expressions when mentioning a block for the first time, none of the expressions we used were the first mention to a block. Additionally, because repeated references to the same block were typically pronominalized, and we did not analyze pronominal references, the expressions we analyzed typically were not immediately repeated mentions.

Two coders independently coded each of the referential phrases for the point-of-disambiguation, defined as the beginning of the word that uniquely identified the referent, given the set of blocks on the addressee's board in the same sub-area as the target. For example, if the intended referent were a long green horizontal block in a sub-area with several long green vertical blocks, the point-of-disambiguation for *the long green horizontal block* would be the onset of *horizontal*. We defined the global context (Salmon-Alt & Romary, 2000) as the sub-area of the target because participants only worked on one sub-area at a time, so including all of the game-pieces from the entire board would likely overestimate the larger context. We hypothesized that the linguistic and pragmatic constraints would sometimes reduce the referential domain to a subset of the blocks in the sub-area.

Inter-coder agreement was high, and the few disagreements were resolved through discussion. Analysis of each definite referring expression revealed that 53% contained a linguistic point-of-disambiguation with respect to the entire sub-area; the remaining 47% did not uniquely specify the intended referent with respect to the sub-area (e.g., *the green piece* uttered in a sub-area containing multiple green blocks). For simplicity, we will refer to these expressions as ‘ambiguous’ because they are ambiguous with respect to the global context (clearly, however the use of a definite indicates that the speaker intended these expressions as fully specified with respect to his referential domain).

The intended referent of the speaker's referring expression was identified as the ‘target’ block. Competitor blocks were defined as blocks which were in the same sub-area as the target and were at least temporarily consistent with one or more content words in the referring expression (e.g. any horizontal block given the expression *the horizontal blue block*). Unrelated blocks were defined as color blocks that were not consistent or temporarily consistent with the target expression (e.g. a green block given the expression *the blue block*). Trials for which the eye-tracking data were accurate, and which had at least one competitor block and one other block on the board, were selected for further analysis. The presence of additional blocks was necessary in order to provide a trial-by-trial comparison of the probability of fixating a target compared to a competitor or other block in our eye-tracking analysis. These selection criteria excluded trials for which the track was lost or unreliable, trials for which there were no competitors (e.g. *the blue block*, uttered in the context of a single blue block, and multiple green blocks), and trials for which there were no unrelated blocks (e.g. *the red rectangle*, uttered in the context of a red square and a red rectangle, but no other blocks). After applying these criteria, 193 disambiguated and 558 ambiguous trials were available for further analysis.

We hypothesized that speakers use referential domains constrained by linguistic and pragmatic factors. If this hypothesis is correct, when an utterance is disambiguated with respect to the entire sub-area, the non-target blocks in the sub-area should be highly salient based on these constraints. In contrast, when an utterance is ambiguous with respect to the sub-area, the non-target blocks should have low salience. Additionally, we expected target blocks to be more salient the less specific the expression. In order to test these predictions, we coded the target, competitor, and unrelated color blocks along three dimensions: proximity, relevance to the task, and recency. For the purposes of this analysis, we operationalized the three constraints as follows:

Proximity: For a given reference to a block, the most recently mentioned block, prior to the target referring expression, was identified. This block was given a proximity score of zero. The other blocks in the current sub-area were then ranked in order of proximity to this most recently mentioned block, with the closet block receiving a score of 1. Blocks could tie in rank, and the mean rank was 3.02 (*SD* = .44). In example 2, *the black block* is the target reference.

(2) 1: up

2: ok

1: uh is a green...dark green...rectangle.

2: rectangle

1: and...it would be sitting on top of *the black block*

The most recently mentioned block, prior to the onset of the target referring expression is the rectangle, thus the rectangle would receive a proximity score of zero. The proximity scores for the other color blocks in this sub-area would be the ranked distance (e.g. 1, 2, 3) between each of these blocks and the rectangle.

Relevance to the task: Each block was coded as to whether the constraints of the task "did" or "did not" allow an upcoming reference to that block. This coding was based on a set of agreed upon heuristics, such as two blocks cannot be placed in the same location, blocks must be placed completely on the board, and the preferred place to put a new block is next to the last one that was placed. In Example 3, Speaker 2 uses a linguistically ambiguous expression, *the yellow*, which is nonetheless understood, in part because task constraints rule out reference to the competitor yellow block. Figure 4 shows the scene at the time of this exchange from the perspective of Speaker 2.

(3) 1: it's right...kind of kitty corner...it's only touching one row on your right

2: ok

1: and it can't go anywhere else

2: so like if I were p- to put it to the right of *the yellow*? and then slide it up three?

1: uh.. yeah

In this exchange, partner 1 describes where to place a comb block. Despite the fact that there are two yellow blocks in the current sub-area, the noun phrase *the yellow* is only consistent with the horizontal yellow block because only the horizontal yellow block (the target) has space on the right hand side to place the comb. The competitor yellow block does not have room to the right to place a comb block. In this example, the target block was also mentioned more recently than the competitor. Using a coding scheme where 0 = predicted by the task; 1 = not predicted, the mean task rating was .73 (*SD* = .07).

Recency: Recency was defined as the number of conversational turns since the last reference to the block. A turn was defined as a word or sequence of words uttered by one partner which was not interrupted by the other partner. Each turn was on average seven words long. The mean recency score was 137.56 (*SD* = 67.48). We used this metric of recency rather than simply whether a block "was" or "was not" recently mentioned in order to capture the variability in discourse history for each block.

3.2.2. Constraint reliability—For each critical referring expression, the constraint scores for each block in the sub-area of the target block were coded. A second coder who was naïve to the experimental predictions independently coded the data from each of the 12 pairs of

participants. We then calculated the inter-coder reliability for the data from six of the 12 pairs. Inter-coder reliability was high; for each constraint, the data from the two independent coders was significantly more similar than would be expected due to chance.

Proximity: Despite the fact that proximity scores ranged from 0 to 10, the two coders assigned the same proximity rank to 53.15% of the 3157 blocks they coded. Two different metrics were used to quantify the degree of agreement while taking chance into account. The Concordance correlation coefficient (Lin, 1989; 2000) was .8418 (2-tailed *95% CI* lower bound = .8317). Krippendorff's alpha (Hayes, 2005) was .8415. A Concordance coefficient or Krippendorff's alpha higher than .7 is generally considered to be good agreement.

Relevance to the task: Agreement on task ratings was 83.94%; chance agreement was 50%. Taking chance into account, agreement was moderate; Krippendorff's alpha was .5598. A second measure, Cohen's Kappa, was .560. This measure was used instead of Lin's concordance coefficient because the task ratings were binary. While the agreement for task scores was lower than that observed for proximity, it was still above chance.

Recency: Recency scores by far had the largest range (0 to 2053 turns), but the coders still chose the same rating for 49% of blocks. The average deviation (including deviations of zero) was 25.79 turns. Lin's concordance coefficient was .9598 (2-tailed *95% CI* lower bound = . 9571). Calculating Krippendorff's alpha was infeasible due to the large size of the matrix required for that computation and limitations of the available software. The high degree of agreement was likely facilitated by the fact that each conversational turn was clearly numbered in our transcripts.

Disagreements in coding were resolved through discussion. The following analyses are based on the final, mutually accepted coding for each of the twelve pairs of participants.

3.2.3 Predicting reference specificity—Proximity, task, and recency scores for target, competitor and unrelated color blocks are show in Figures 56-7. Targets consistently showed an advantage for all three constraints[2], establishing their validity. However, the most consistent predictors of speaker specificity were the ratings of competitor blocks; speakers were more likely to disambiguate the target with respect to competitor blocks when the competitors were more proximal and fit the task constraints.

Three planned ANOVAs were used to analyze the constraint scores for target, competitor and unrelated blocks for each of the three constraints. Consistent with the structure of the items analysis of Experiment 1, each of the 37 color blocks was assigned a unique identification number, yielding 37 items. Each trial had one unique referent, thus the item for that trial was the referent number. Due to location on the board and color or shape of nearby blocks, some items were less likely to contribute usable trials. Items for which data were missing from one or more cells were excluded, leaving 23 items for this analysis. We were not missing data from any of the cells in the participants analysis.

The analysis of proximity scores revealed a significant main effect of object type (target, competitor, unrelated), $F1(2,22) = 196.61$, $p<.0001$, $\eta^2_G = .85$; H-F $\varepsilon = .73$, $F2(2,44) = 68.20$, $p<.0001$, $\eta^2_G = .62$, due to significantly lower (e.g. closer in proximity) scores for targets compared to either competitor or unrelated blocks (*p's*<.0001). Competitor blocks were numerically lower in proximity than unrelated blocks, but this effect was not significant in the participants analysis ($p=.37$) and was marginal in the items analysis ($p=.06$). An effect of

[2]The speaker's choice to use a definite noun phrase instead of a pronoun in cases where the target referent is already salient has been observed in previous research on unscripted conversation (Brennan, 1995), and may be related to the inherent ambiguity in the task.

specificity (ambiguous, disambiguated) was only significant in the participants analysis, *F1* (1,11) = 17.49, $p<.01$, $\eta^2_G = .16$; *F2*(1,22) = .76, $p=.39$, and was due to lower proximity scores for disambiguated utterances. These main effects were qualified by a significant interaction, *F1*(2,22) = 17.23, $p<.0001$, $\eta^2_G = .25$; H-F $\varepsilon = .74$, *F2*(2,44) = 6.90, $p<.01$, $\eta^2_G = .04$. A series of planned, two-tailed t-tests were used to directly compare the scores for ambiguous and disambiguated expressions. Proximity scores for target blocks were equivalent for ambiguous and disambiguated referring expressions, *t1*(11) = 1.48, $p=.17$; *t2*(22) = 1.64, $p=.12$. However, proximity scores for competitor blocks were significantly lower (e.g. closer in proximity) for disambiguated, compared to ambiguous expressions, *t1*(11) = 5.51, $p<.0001$, $d=2.25$; *t2*(22) = 3.73, $p<.01$, $d=.71$. The proximity scores for unrelated blocks were numerically lower when the referring expression was disambiguated, but this effect was only marginal in the participants analysis, *t1*(11) = 2.16, $p=.05$, $d=.51$; *t2*(22) = .18, $p=.86$.

Task relevance scores for competitor blocks also significantly predicted the form of the referring expression. The ANOVA for task scores revealed a main effect of object type, *F1* (2,22) = 600.75, $p<.0001$, $\eta^2_G = .95$; H-F $\varepsilon = .82$, *F2*(2,44) = 189.49, $p<.0001$, $\eta^2_G = .82$, due to significantly lower scores (e.g. more predicted) for targets than either competitor or unrelated blocks (*p's*<.0001). Competitor blocks had numerically lower scores than unrelated blocks, but this effect was marginal in the participants analysis ($p=.06$), and not significant in the items analysis ($p=.48$). The main effect of specificity was not significant, *p's*>.4, but specificity did significantly interact with object type, *F1*(2,22) = 17.53, $p<.0001$, $\eta^2_G = .22$; H-F $\varepsilon = .84$, *F2* (2,44) = 5.12, $p<.05$, $\eta^2_G = .07$. The targets of disambiguated expressions had marginally higher task scores (e.g. less relevant) compared to the targets of ambiguous expressions, *t1*(11) = 2.12, $p=.06$, $d=.67$; *t2*(22) = 1.74, $p=.10$, $d=.48$. We observed an opposite and stronger effect for competitor blocks, which were significantly more task relevant when the referring expression was disambiguated, *t1*(11) = 3.66, $p<.01$, $d=1.44$; *t2*(22) = 4.98, $p<.0001$, $d=1.17$. Unrelated blocks were less task-relevant when expressions were disambiguated, however this effect was only significant in the participants analysis, *t1*(11) = 2.71, $p<.05$, $d=.66$; *t2*(22) = .09, $p=.93$.

Unlike the proximity and task constraints, the recency of competitor blocks only marginally predicted the specificity of referring expressions. An ANOVA for recency scores revealed a main effect of object type, *F1*(2,22) = 24.72, $p<.0001$, $\eta^2_G = .37$; *F2*(2,44) = 33.02, $p<.0001$, $\eta^2_G = .32$, which was due to significantly lower recency scores for targets compared to competitors and unrelated blocks (*p's*<.0001), and equivalent recency scores for competitor and unrelated blocks (*p's*>.35). The specificity effect was only significant in the items analysis, *F1*(1,11) = 1.46, $p=.25$, *F2*(1,22) = 6.07, $p<.05$, $\eta^2_G = .04$, and was due to lower recency scores for disambiguated items. However, unlike the findings for proximity and task relevance, the effect of object type did not interact with ambiguity, *p's*>.13. Planned comparisons showed that recency scores for target and unrelated blocks were equivalent for ambiguous and disambiguated expressions (*p's*>.18). However, an effect of specificity was observed in the analysis of competitor blocks. Consistent with the findings from task and proximity ratings, when the noun phrase was disambiguated, the competitors were more recently mentioned, *t1* (11) = 2.14, $p=.06$, $d=.66$; *t2*(22) = 2.94, $p<.01$, $d=.74$.

The results of the constraints analysis were consistent with our first hypothesis that speakers modify their referential expressions with respect to pragmatically constrained referential domains. We found that the ratings of competitor blocks on two of the three constraints, proximity and task relevance, significantly predicted whether speakers would disambiguate their referring expressions with respect to these blocks. The marginal effect of specificity for the recency of competitor blocks suggests that recency of mention may also be an important factor in determining the contents of the referential domain. The fact that speakers were marginally more likely to use a more specific expression when targets were less relevant to the

task is also consistent with our hypothesis and suggests that speakers are sensitive to whether the intended referent is within the current domain.

We now turn to an analysis of the addressee's interpretation to examine the degree to which the addressee used the same referential domain as the speaker.

3.2.4. Reference interpretation—Our analysis of the addressee's interpretation of her partner's referring expressions was guided by our second hypothesis, that addressees would interpret referring expressions with respect to referential domains similar to those used by the speaker to construct the expression. We took three different approaches to examining the addressee's interpretation of her partner's utterances. First, we compared the overt (speech) response to expressions that were ambiguous or disambiguated with respect to the sub-area. In our second set of analyses, we examined the eye movements that addressees made following ambiguous and disambiguated expressions. In the third set of analyses, we directly combined the results of the constraints analysis with the eye-tracking data on a trial-by-trial basis to examine whether the salience of competitor blocks could be used to predict eye fixations directly.

In our first set of analyses, we examine the addressee's response to her partner's utterances, in order to ascertain whether in their final, off-line interpretation, addressees understood, or were generally confused by their partner's contributions. If the addressee's referential domain is significantly different than the speaker's, then it is likely that some expressions will be ambiguous with respect to the addressee's referential domain, because of the large number of potential competitors in the sub-area. Those expressions would confuse the addressee because they would not specify a unique referent from her perspective.

The two most common responses to a partner's utterance were backchannels such as *mm-hmm*, and *okay*. These forms were included in a category we call confirmations, in which the addressee positively replied to her partner's contribution. This category included what Bangerter and Clark (2003) term acknowledgment tokens (e.g. *yup, ok, mm-hmm*), agreement tokens (e.g. *right*), and consent tokens (e.g. *okay*), as well as acknowledgments in which the speaker repeats what was said (Traum, 1994; Zollo & Core, 1999), as in example (4) below. In examples (4) − (7), the target noun phrase is italicized.

(4) 1: ok ... RIGHT below the SNAKE is *the GREEN square*

2: the GREEN square .... mm hmm

Confirmations occurred in roughly equal proportions when the addressee was responding to an utterance that contained an ambiguous expression (71.1%) and a disambiguated expression (71.4%). Addressees continued the conversation without a confirmation 15% of the time for ambiguous expressions and 13% of the time for disambiguated expressions (example 5). This category included responses which started with terms like *Okay* and *yeah well* if they were included in the prosodic contour of the continuation; in these cases the terms appeared to serve a different purpose than the prosodically distinct confirmations. Addressees contradicted their partners 1.8% and 4.2% of the time for ambiguous and disambiguated expressions, respectively (example 6).

(5) 2: Alright I'll just move mine...alright...oh man, I just lost another block...ok. Alright, so I don't know where *the- the red the horizontal ... long one* goes.

1: Um, one space, between that red one you just put and the little red one.

(6) 1: So there's one row between ... *the ... light green rectangle* and the pencil.

2: That's not possible.

Addressees were slightly more likely to ask a general question about block placement like *Skip three?* or *To the right?* when responding to an utterance with a disambiguated expression (8.3%) than when responding to an ambiguous one (7%). On rare occasions, addressees acted confused or uttered only a disfluency in response, such as *What?* or *uhhh*; these responses occurred only 2.7% for ambiguous and 2.6% for disambiguated expressions. Finally, we examined the number of times that addresses asked a clarifying question specifically about the critical noun phrase (example 7).

(7) 1: directly...ABOVE *the red*, grab your lamp

2: Ok the red we just put in?

We observed slightly more confusions about the noun phrase following ambiguous expressions (2.0% /11 trials), compared to .5% for disambiguated ones (1 trial). However, the infrequent occurrence of noun phrase confusions suggests that addressees generally understood their partners.

This off-line analysis of the addressee's interpretation of her partner's utterances suggests that most of the time utterances containing disambiguated and ambiguous utterances were ultimately understood. There are some suggestions in the data that addressees might have been confused slightly more often for the ambiguous references. However, what is most striking is that addressees were generally not confused by referring expressions that should have been confusing if the addressee's referential domain were not similar to the speaker's. We now turn to an analysis of the addressee's eye movements to examine whether they experienced temporary confusion as they interpreted the expressions.

Our second set of analyses focused on the eye movements that addressees made as they interpreted the same set of referring expressions examined in the constraints analysis. Eye movements to the different blocks were grouped into three categories: Looks to (a) the **target** block--the intended referent of the noun phrase; (b) **competitor** blocks that (at least) temporarily matched the referring expression as the utterance unfolded (e.g., any long block in the same sub-area as the target would be a competitor for *the long green block*); and (c) any **other** blocks in the sub-area (including picture blocks and color blocks which never matched the target referring expression). Our analysis was guided by our second hypothesis, that addressees would interpret referring expressions with respect to referential domains similar to those used by the speaker to construct the expression. If addressees use the same referential domain as the speaker, addressees should identify the target referent more quickly when interpreting referring expressions which are produced with respect to pragmatically constrained referential domains. Faster interpretation for expressions which were constructed with respect to constrained referential domains--the ‘ambiguous’ expressions--would be indicated by an earlier preference to fixate the target during the time region immediately following the onset of the referring expression. In this same time region, we would also expect to find fewer fixations to competitor blocks compared to expressions which were disambiguated with respect to the entire sub-area.

Disambiguated referring expressions: In our on-line analysis, we first examine those referring expressions that were disambiguated with respect to the entire sub-area. If, like the speaker, the addressee's domain includes at least some of the competitors in the sub-area, then certainty about the referent should increase at the point-of-disambiguation, resulting in an increase in fixations to the referent.

Figure 8 shows the proportion of fixations to target, competitor and other blocks for the disambiguated utterances, aligned at the point-of-disambiguation. The proportion of fixations to the target rises shortly after the point-of-disambiguation. In addition, addressees had a

baseline preference to look at the target over the competitor and other blocks, even before the point-of-disambiguation, a point which we will return to later.

Eye-movement analyses were performed on the proportion of fixations to the blocks in the same sub-area as the target. For disambiguated utterances, a 495 ms baseline region plus three consecutive 800 ms time regions were analyzed; Figure 9 shows the proportion of fixations to target, competitor, and other blocks in the four regions.

A significant rise in fixations to the target and a drop in fixations to competitors (relative to other blocks) following the point-of-disambiguation would demonstrate that addressees used the disambiguating information to identify the intended referent. The baseline region encompassed the time between the average noun phrase onset and 600 ms before the point-of-disambiguation. The two central regions encompassed the 800 ms before and after the point-of-disambiguation, plus 200 ms. The fourth region ranged between 1000 and 1800 ms following the point-of-disambiguation. These regions were used to establish the pattern of baseline fixations, and how these fixations persisted over time. An ANOVA with region and object type as factors was used to analyze the fixations to the different objects over time. The items analysis was patterned after the constraints analysis and was restricted to the items for which we obtained observations from at least two participants in each cell (a reduced degrees of freedom in the items analyses reflects the loss of some items). For the participants analysis, we obtained observations from at least two items for each cell.

The ANOVA revealed a significant effect of object type, $F1(2,22) = 54.53$, $p<.0001$, $\eta^2_G = .68$; H-F $\varepsilon = .55$, $F2(2,36) = 41.18$, $p<.0001$, $\eta^2_G = .54$, which was due to significantly more fixations to targets than either competitors or other blocks (p's<.001), and an equivalent number of fixations to competitor and other blocks (p's>.7). A main effect of region was marginal, H-F $\varepsilon = .64$, $F1(3,33) = 3.01$, $p=.07$, $\eta^2_G = .03$; H-F $\varepsilon = .53$, $F2(3,54) = 3.56$, $p=.05$, $\eta^2_G = .02$. These main effects were qualified by an object type by region interaction that was reliable in the participants analysis, and marginal in the items analysis, H-F $\varepsilon = .55$, $F1(6,66) = 3.97$, $p<.05$, $\eta^2_G = .09$; H-F $\varepsilon = .32$, $F2(6,108) = 3.17$, $p=.06$, $\eta^2_G = .05$.

Planned ANOVAs explored this interaction by analyzing the region effect for each block type separately. The ANOVA for fixations to the target revealed a main effect of region that was marginal in the items analysis, $F1(3,33) = 3.73$, $p<.05$, $\eta^2_G = .13$; H-F $\varepsilon = .54$, $F2(3,54) = 3.29$, $p=.06$, $\eta^2_G = .06$. Planned one-tailed t-tests indicated that between the baseline region and the region immediately preceding the point-of-disambiguation, there was a non-reliable increase in target fixations, $t1(11) = 1.07$, $p=.15$, $t2(18) = 1.66$, $p=.06$. After the point-of-disambiguation, looks to the target clearly rise, $t1(12) = 2.57$, $p<.05$, $d=.74$; $t2(18) = 2.03$, $p<.05$, $d=.41$, replicating the pattern observed in experiments with scripted utterances and simple displays (e.g., Eberhard et al., 1995). Finally, between the third and the fourth regions, target fixations decreased, $t1(11) = 2.59$, $p<.05$, $d=.49$; $t2(18) = 3.78$, $p<.001$, $d=.77$.

The ANOVA for fixations to competitors across regions was also significant, $F1(3,33) = 3.74$, $p<.05$, $\eta^2_G = .11$; H-F $\varepsilon = .80$, $F2(3,54) = 3.62$, $p<.05$, $\eta^2_G = .06$. Planned one-tailed t-tests indicated that between the baseline region and the region immediately preceding the point-of-disambiguation, fixations to competitors rose significantly, $t1(11) = 2.76$, $p<.01$, $d=.66$; $t2(18) = 2.05$, $p<.05$, $d=.51$. After the point-of-disambiguation, looks to competitors fell, $t1(11) = 2.27$, $p<.05$, $d=.67$; $t2(19) = 3.69$, $p<.001$, $d=.53$. Finally, between the third and the fourth regions, competitor fixations remained stable, $t1(11) = .40$, $p=.35$, $t2(18) = .48$, $p=.32$. In contrast, looks to other blocks remained stable across these regions. The ANOVA for fixations to other blocks was not significant, $F1(3,33) = 1.57$, $p=.22$, H-F $\varepsilon = .75$, $F2(3,54) = 1.83$, $p=.17$.

If addresses considered competitor blocks to be potential referents before the point-of-disambiguation, than we would expect to find that addressees would be more likely to fixate

competitor blocks, compared to other blocks before the point-of-disambiguation, but not after. We tested this prediction by directly comparing the proportion of fixations to competitor and other blocks at each region in four planned, one-tailed t-tests. During the baseline region, the proportion of fixations to competitor and other blocks did not differ, $t1(11) = .16$, $p=.44$, $t2(18) = .48$, $p=.32$. However, in the region immediately preceding the point-of-disambiguation, there were more fixations to competitor blocks than to other blocks, $t1(11) = 1.86$, $p<.05$, $d=.86$; $t2(18) = 2.55$, $p<.05$, $d=.79$. In the two regions following the point-of-disambiguation, looks to competitors and other blocks did not differ, region 3: $t1(11) = .29$, $p=.39$, $t2(18) = .50$, $p=.31$; region 4: $t1(11) = .50$, $p=.31$, $t2(18) = .77$, $p=.23$.

In summary, for referring expressions that uniquely specified an intended referent with respect to each of the blocks in the sub-area, we replicated the general pattern of results previously observed in experiments with scripted utterances and simple displays (Eberhard et al., 1995). Following the point-of-disambiguation, fixations to the target increased. Before the point-of-disambiguation, looks to competitors were significantly higher than looks to other, non-target blocks, suggesting that the competitors were within the addressee's referential domain. In the time regions following the point-of-disambiguation, looks to competitors and looks to other blocks were equivalent, suggesting that the competitor blocks were no longer considered potential referents. The major difference between the current results and those with scripted utterances is that there is a baseline preference for the addressee to look at the intended referent early in the speaker's utterance. We return to the source for this preference, after presenting the results for ambiguous referring expressions.

Ambiguous referring expressions: For those referential expressions that were linguistically ambiguous with respect to the entire sub-area, we can assume that the speaker did not consider the competitors in that sub-area to be sufficiently proximal or task-relevant to influence the choice of referring expression. We have already seen that addressees are not overtly confused by these ambiguous utterances. We can now ask whether there was temporary confusion as the utterance unfolded in time. If the addressee used a referential domain that was very different than the speaker's, or could not use the pragmatic information that constrained the referential domain during on-line processing, we would expect to see competition between the target and competitor blocks that does not abate as quickly as it does following the point-of-disambiguation for the disambiguated utterances. In contrast, if the addressee's referential domain is similar to the speaker's, and therefore much smaller than the sub-area, addressees should make relatively few looks to competitor blocks, even though the referring expression is compatible linguistically with both the target and competitors.

Figure 10 shows the proportion of fixations to target, competitor and other blocks for the linguistically ambiguous referring expressions, aligned at reference onset. Pictured are the 558 trials for which the eye-track was accurate and there was at least one competitor and one other block on the board. Similar to the pattern of looks following disambiguated expressions, there was a baseline preference to fixate the target block. Because there was not a point at which the ambiguous expressions were disambiguated with respect to the blocks in the sub-area (point-of-disambiguation), we examined looks in two 800 ms regions: one ending at the mean point-of-disambiguation for the disambiguated utterances (1100 ms after the onset of the referential phrase), plus 200 ms, and one beginning at the point-of-disambiguation for the disambiguated utterances, plus 200 ms. We adopted this approach for two reasons; we wanted to have a principled method for defining regions and we wanted the analyses to be comparable to those for the disambiguated utterances. The same pattern of results obtained when we adopted a second analysis strategy: defining two 800 ms regions, one beginning at the onset of the ambiguous referring expressions, plus 200 ms and the second beginning at the offset of the first region (see Appendix C). As with the analyses of disambiguated referring expressions, items analyses were restricted to the items for which we obtained observations from at least

two participants in each cell. In the participants analysis, we obtained data from at least two items in each cell.

Figure 11 shows the proportion of fixations to target, competitor and other blocks in the two 800 ms regions centered on either side of the average point-of-disambiguation for disambiguated references. An ANOVA with region and object type as factors revealed a main effect of object type, H-F ε = .73, $F1(2,22) = 247.11$, $p<.0001$, $\eta^2_G = .94$; H-F ε = .60, $F2(2,58) = 125.05$, $p<.0001$, $\eta^2_G = .76$, due to significantly more looks to targets than either competitor or other blocks (p's<.0001), and equivalent looks to competitor and other blocks (p's>.4). The main effect of region was marginal in the participants analysis, $F1(1,11) = 3.35$, $p=.10$, $\eta^2_G = .01$; $F2(1,29) = 9.03$, $p<.01$, $\eta^2_G = .01$. The interaction was not reliable, H-F ε = .65, $F1(2,22) = 1.98$, $p=.18$, H-F ε = .73, $F2(2,58) = 2.96$, $p=.08$.

To mirror the analysis with disambiguated referring expressions, we performed planned 2-tailed tests on the proportion of fixations to targets over time, as well as the proportion of fixations to competitors compared to other blocks over time. Unlike the disambiguated utterances, the proportion of fixations to target blocks did not increase. Instead there was a decrease in the proportion of fixations to targets across the two regions that was significant only in the items analysis, $t1(11) = 1.59$, $p=.14$, $t2(29) = 2.13$, $p<.05$, $d=26$. Most remarkably, addressees were no more likely to look at competitor blocks than other blocks in either region, even though each referential expression was as consistent with competitor blocks as it was with the target block. In fact, we observed a non-significant preference to fixate other blocks more than competitor blocks in both regions, region A: $t1(11) = 1.63$, $p=.13$, $t2(29) = .14$, $p=.89$; region B: $t1(11) = 1.73$, $p=.11$, $t2(29) = 1.66$, $p=.11$.

Lastly, we examined the pattern of eye movements for trials on which the addressee was not fixating the target at the onset of the referring expression to see if addressees would still converge on the target without considering the competitors in cases where their attention had not already been attracted to the target before the referring expression. We performed this analysis for both disambiguated and ambiguous expressions. The pattern of results was similar to our previous findings (see Appendix D). For disambiguated expressions, we observed significant effects of object type, time region and a significant interaction. Target fixations rose significantly after the point-of-disambiguation, indicating that addressees were sensitive to this disambiguating information. Unlike the analysis on the full dataset, the increase in competitor fixations across the first two time regions was only marginal, likely due to the reduced amount of data. For ambiguous expressions, we analyzed the results in the same regions as our main analysis, before and after the average point-of-disambiguation for disambiguated expressions. We observed only a main effect of object, due to more target fixations than either competitor or other fixations, and equivalent fixations to competitor and other blocks. Again, target fixations remained stable across the two regions.

Our third set of analyses used the trial-by-trial analysis of constraints to test the hypothesis that the factors that predict whether the speaker considers competitor blocks to be in the referential domain also predict whether the addressee considers competitor blocks to be in the referential domain. While previous work by Beun and Cremers (1998), as well as our own constraints analysis indicate that the proximity and task-relevance of entities in the global context predict whether the speaker considers them to be in the referential domain, thus far we have not directly tested whether the salience of entities in the global context also predicts whether the addressee will consider these entities in the referential domain. The fact that addressees interpreted ambiguous and disambiguated expressions differently is consistent with this hypothesis, suggesting that when speakers consider competitors to be in the referential domain, addressees are likely to briefly consider these competitors while interpreting referring expressions.

However, we can go further than this by using the ratings of competitor blocks on the three constraints to predict competitor fixations directly.

This analysis was guided by two predictions. First, we predicted that when competitor blocks were more salient (as judged by the ratings from the constraints analysis), that addressees would be more likely to fixate competitors as they interpreted the referring expression. Second, we predicted the effect of task-relevance would be modulated by recency. Specifically, we expected that task-relevance would not be predictive of competitor fixations if the last mentioned block was a competitor, because a recently mentioned competitor is likely to still be highly salient in the addressee's model of the discourse, regardless of whether it remained task-relevant.

For each referring expression that was included in the eye-tracking analysis, we took the rating of the competitor with the smallest proximity rating (e.g. the closest competitor to the last mentioned entity), the rating of the competitor with the smallest task rating (e.g. most task-related), and the rating of the competitor with the smallest recency rating (e.g. the most recently mentioned), and used these scores to predict whether the addressee was likely to fixate a competitor block on that trial. We chose not to use the average ratings on the various constraints because this would underestimate the likelihood of fixating a salient competitor on a trial that had many competitors that were not relevant to the task, not mentioned recently and far from the last mentioned thing, but one very proximal, task-relevant and recently mentioned competitor. We analyzed eye movements during the critical regions in which competitor fixations were most likely to occur. These time-regions were the ‘competitor region’ for disambiguated expressions (e.g. the 800 ms immediately before the point-of-disambiguation), and Region A for ambiguous expressions (e.g. the 800 ms immediately before the average point-of-disambiguation for disambiguated expressions).

The data were analyzed with a hierarchical linear model (HLM) using restricted maximum likelihood estimation; the analysis was performed using the HLM 6 software by Scientific Software International (see Bryk & Raudenbush, 1992). Unlike ordinary least squares regression, HLM is appropriate for clustered datasets (the repeated-measured design used here resulted in a dataset clustered by participant). Task, proximity and recency scores (centered at grand means), as well as the specificity of the referring expression (ambiguous, disambiguated) were used to predict the proportion of fixations to competitor blocks in the critical region. The results of three different planned analyses are presented in Table 1. In the first model, each eye-tracking trial was included in the analysis. Proximity and ambiguity both significantly predicted competitor fixations. Consistent with the results from the analysis of speaker specificity, the closer the most proximal competitor was to the last mentioned entity, the more likely the addressee was to fixate competitors while interpreting that referring expression. This result suggests that proximity to the last mentioned entity was relevant to the speaker's assessment of whether an entity in the global context was a potential referent. Consistent with the results of our previous eye-tracking analyses, the significant effect of ambiguity was due to more competitor fixations when the expression was disambiguated.

The second and third models were used to test our second prediction, that task-relevance would have a stronger effect on competitor fixations when the last mentioned referent was not a competitor. The second model tested the effects of task and ambiguity for trials where the last mentioned block was a competitor. The third model tested the effects of task and ambiguity for trials where the last mentioned block was not a competitor (we did not include proximity in these models because it was calculated based on the most recently mentioned block).

In the second model, the only significant predictor of fixations was ambiguity, with more competitor fixations when the expression was disambiguated. In the third model, we observed

a very different pattern of results. Here, the only significant predictor of fixations was task-relevance, with more competitor fixations when there was at least one task-relevant competitor in the sub-area.

3.3. Discussion

The aim of Experiment 2 was to identify the factors speakers use to determine which entities are in the referential domain and test whether addressees are sensitive to the same factors when interpreting these expressions. Our analysis was guided by two hypotheses. First, we predicted that speakers would modify their referential expressions with respect to referential domains constrained by three linguistic/ pragmatic factors: proximity, relevance to the task and recency. The results of the constraints analysis were consistent with this prediction for two of the three factors. We saw that speakers disambiguated their expressions with respect to each of the blocks in the sub-area when competitors in that sub-area were proximal to the last mentioned block and when they were relevant to the current task. We also observed somewhat weaker effects of recency on referential specificity. These results are consistent with the well-establishing finding that the form of a speaker's referring expression is influenced by what is salient within the local domain (Beun & Cremers, 1998; Brown-Schmidt & Tanenhaus, 2006; Olson, 1970; Osgood, 1971; Pechmann, 1989; Salmon-Alt & Romary, 2000).

Our second hypothesis was that that during on-line interpretation of referring expressions, addressees use referential domains that are similar to the referential domains that speakers use to produce these expressions. The results were consistent with this hypothesis. Addressees had a strong preference to fixate the target and showed equivalent fixations to competitors and unrelated blocks when interpreting ambiguous expressions. The results for expressions which were disambiguated with respect to the sub-area showed competition effects and a rise in looks to the target which was time-locked to the point in the utterance which disambiguated the target from the competitors in the sub-area. The fact that we saw point-of-disambiguation effects for disambiguated utterances tells us that it should be possible to ask detailed questions about time course in unscripted conversation, on a par with those that have been examined with scripted utterances.

A notable difference between our results and the results of previous work is that addressees had a baseline preference to fixate the target regardless of whether the referential domain was constrained. The results of the constraints analysis for target blocks provides an explanation for this—regardless of whether or not the expression reflected a constrained referential domain, target blocks were always more salient that any of the other blocks in the sub-area. This may have given addressees a preference to look at the target block and possibly expect an upcoming reference to this block.

In our final set of analyses, we used the factors that predicted whether the speaker considered competitors to be in the referential domain to directly predict whether addressees considered these competitor blocks to be potential referents. We found clear effects for the proximity constraint: the proportion of fixations to competitors was higher the closer competitors were to the last mentioned entity. We also found that task-relevance did significantly predict competitor fixations, but only when the most recently mentioned block was not a competitor. In conjunction with the analysis of speaker specificity, these results demonstrate that speakers and listeners use the same factors to determine which entities in the global context are potential referents.

Finally, an important question that we could not address with our design is how the similarity of referential domains changes over time. For example, at the beginning of a conversation, one might expect larger and less constrained referential domains, which might include more competitors. Unfortunately, the current task did not lend itself to this type of analysis because

most of the target trials occurred well into the discourse. This was because we only used trials when the target referent of a definite noun phrase was on the board along with at least one competitor and unrelated block.

4. General Discussion

The present research adopted a targeted language games approach to examine two questions: (1) How does the process of on-line ambiguity resolution for the addressee in a conversation compare to ambiguity resolution processes outside the context of a conversation? And (2) what factors constrain referential domains for speakers, and are addressees sensitive to the same factors as they interpret their interlocutor's expressions?

Experiment 1 focused on the first question, examining interpretation of expressions like *the clown* in contexts which included both a clown and a cloud. Consistent with previous work using scripted utterances, we observed typical lexical competitor effects for expressions uttered by the experimenter outside the context of the conversation. We hypothesized that addressees would have decreased competition from lexical competitors when interpreting expressions within the conversation because of conversationally constrained referential domains. Consistent with this hypothesis, the fixation analyses revealed no competition from lexical competitors in these cases, and a baseline preference to fixate the target increased as they heard the referring expression.

Experiment 2 addressed the second question by focusing on factors that influence speaker and addressee referential domains for modified noun phrases like *the green horizontal block.* Our first hypothesis was that speakers should modify their referential expressions with respect to referential domains that were constrained by linguistic and pragmatic factors. Two of the factors we identified—proximity and relevance to the task—did significantly predict whether speakers would modify their expressions with respect to the entire sub-area, suggesting that these factors played a role in the speaker's decision as to what was in the referential domain. Moreover, we found clear evidence that the addressee interpreted expressions with respect to similarly constrained referential domains: the same factors that predicted whether the speaker disambiguated his expressions with respect to the competitor blocks predicted whether the addressee fixated these competitors as she interpreted the same expressions. When we examined referring expressions that were modified with respect to each of the blocks in the sub-area, addressees temporarily considered competitor blocks in the sub-area as they interpreted the expression. The increase in target looks and decrease in competitor looks following the point-of-disambiguation indicates that addressees were able to use the disambiguating information on-line during the conversation. This result in and of itself demonstrates that for certain referential contexts, the results observed in previous experiments do replicate in conversation. In contrast, when speakers used expressions which were ambiguous with respect to the blocks in the entire sub-area, addressees rarely fixated competitors. Together, the results of Experiments 1 and 2 tell us that speakers and addressees do use similar referential domains during conversation, that these representations facilitate on-line comprehension processes, and that two specific pragmatic factors contribute to the circumscription of the referential domains.

Most generally, our results demonstrate that it is possible to use eye movements to examine real-time processing in unscripted interactive conversation at a temporal grain comparable to that obtained in more traditional experiments with scripted utterances and simpler visual displays. This is important for two reasons. First, as we argued in the introduction, many questions about language processing can only be addressed by examining moment-by-moment processing using response measures that are closely time-locked to the utterance. Second, many of the phenomena that occur naturally in interactive conversation, including back-channel

responses, negotiation over referential expressions, perspective, and on-the-fly adjustment to feedback from an interlocutor, are difficult, if not impossible to create in scripted utterances and within traditional experimental paradigms. Thus, paradigms like the one used here can be used to examine a range of phenomena that emerge in natural conversation—phenomena that are central to conversation but problematic for standard experimental approaches. Moreover, the targeted language games methodology permits an examination of the form of utterances, as in more traditional dialogue analysis, in conjunction with real-time analyses, which can be used to evaluate specific hypotheses about real-time processing. To be sure, we view the investigation of language processing in unscripted conversation as a natural companion, not replacement to standard experimental paradigms. Here we have provided an existence proof that such a paradigm can be used to examine on-line processing in conversation, and that by doing so we not only demonstrate that findings from standard paradigms can extend to conversation, but are able to qualify when standard findings do not, as well as to provide novel observations and new insights into how and why processing of unscripted language is different than processing pre-scripted speech.

Finally, because tasks like the one we adopted are similar to tasks for which computational linguists are beginning to develop end-to-end dialogue systems, we believe that combining real-time measures, such as eye movements, with dialogue systems might create a potential test bed for creating and evaluating explicit models of dialogue, leading to a feedback loop between computational and experimental investigations (Campana, 2006).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by NIH grant HD 27206 to M. K. Tanenhaus. Thanks to Carol Faden, Courtney Pooler, Jesseca Aqui, Anne Tanenhaus, Theresa Pucci, and Sanjukta Sanyal for help transcribing and coding data. Special thanks to Ellen Campana, who collaborated on a preliminary version of Experiment 2 (Brown-Schmidt, Campana, & Tanenhaus, 2002; 2005).

References

Aist, G.; Campana, E.; Allen, JF.; Rotondo, M.; Swift, MD.; Tanenhaus, MK. Proceedings of the 27th Annual Meeting of the Cognitive Science Society, Stresa, Italy.. Lawrence Erlbaum Associates; Mahwah, NJ: 2005. Variations along the contextual continuum in task-oriented speech.; p. 79-84.

Allen JF, Byron DK, Dzikovska M, Ferguson G, Galescu L, Stent A. Towards conversational human-computer interaction. AI Magazine 2001;22:27–35.

Allopenna PD, Magnuson JS, Tanenhaus MK. Tracking the time course of spoken word recognition: Evidence for continuous mapping models. Journal of Memory and Language 1998;38:419–439.

Arnold JE, Tanenhaus MK, Altmann RJ, Fagnano M. The old and thee, uh, new. Psychological Science 2004;15:578–582. [PubMed: 15327627]

Austin, JL. How to do things with words. Sbisa, M.; Urmson, JO., editors. Harvard University Press; Cambridge, MA: 1962.

Bakeman R. Recommended effect size statistics for repeated measures designs. Behavior Research Methods, Instruments, and Computers 2005;37:379–384.

Bangerter A. Using pointing and describing to achieve joint focus of attention in dialogue. Psychological Science 2004;15:415–419. [PubMed: 15147496]

Bangerter A, Clark HH. Navigating joint projects with dialogue. Cognitive Science 2003;27:195–225.

Beun R-J, Cremers AHM. Object reference in a shared domain of conversation. Pragmatics & Cognition 1998;6:121–151.

Brennan, SE. How conversation is shaped by visual and spoken evidence.. In: Trueswell, JC.; Tanenhaus, MK., editors. Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions. The MIT press; Cambridge, MA: 2005. p. 95-129.

Brennan SE. Centering attention in discourse. Language and Cognitive Processes 1995;10:137–167.

Brown-Schmidt, S.; Campana, E.; Tanenhaus, MK. Real-time reference resolution by naïve participants during a task-based unscripted conversation.. In: Trueswell, JC.; Tanenhaus, MK., editors. Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions. MIT press; Cambridge, MA: 2005. p. 153-171.

Brown-Schmidt, S.; Campana, E.; Tanenhaus, MK. Reference resolution in the wild: How addressees circumscribe referential domains in a natural, interactive problem-solving task.. Proceedings of the 24th annual meeting of the cognitive science society.; Fairfax, VA.. 2002. p. 148-153.

Brown-Schmidt S, Tanenhaus MK. Mapping thoughts onto utterances: Eye movements predict the content and form of fluent and non-fluent referring expressions. Journal of Memory and Language 2006;54:592–609.

Bryk, AS.; Raudenbush, SW. Hierarchical linear models: Applications and data analysis methods. Sage Publications; Newbury Park, CA: 1992.

Campana, E. An Empirical Analysis of the Costs and Benefits of Naturalness in Spoken Dialog Systems. University of Rochester; Rochester, NY: 2006. Unpublished doctoral dissertation

Chambers CG, Tanenhaus MK, Eberhard KM, Filip H, Carlson GN. Circumscribing referential domains during real-time language comprehension. Journal of memory and language 2002;47:30–49.

Clark, HH. Arenas of language use. University of Chicago Press; Chicago: 1992.

Clark HH, Krych MA. Speaking while monitoring addressees for understanding. Journal of Memory and Language 2004;50:62–81.

Clark HH, Wilkes-Gibbs D. Referring as a collaborative process. Cognition 1986;22:1–39. [PubMed: 3709088]

Cohen, J. Statistical power analysis for the behavioral sciences. Vol. 2nd ed.. Erlbaum; Hillsdale, NJ: 1988.

Cooper RM. The control of eye fixation by the meaning of spoken language. Cognitive Psychology 1974;6:84–107.

Dahan D, Magnuson JS, Tanenhaus MK, Hogan EM. Tracking the time course of subcategorical mismatches: Evidence for lexical competition. Language and Cognitive Processes 2001;16:507–534.

Dahan D, Tanenhaus MK. Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. Journal of Experimental Psychology: Learning, Memory & Cognition 2004;30:498–513.

Dale R, Reiter E. Computational interpretations of the gricean maxims in the generation of referring expressions. Cognitive Science 1995;19:233–263.

Eberhard KM, Spivey-Knowlton MJ, Sedivy JC, Tanenhaus MK. Eye-movements as a window into spoken language comprehension in natural contexts. Journal of Psycholinguistic Research 1995;24:409–436. [PubMed: 8531168]

Gergle D, Kraut RE, Fussell SR. Language efficiency and visual technology minimizing collaborative effort with visual information. Journal of Language and Social Psychology 2004;23:491–517.

Glenberg AM, Meyer M, Lindem K. Mental models contribute to foregrounding during text comprehension. Journal of Memory and Language 1987;26:69–83.

Grice HP. Meaning. The Philosophical Review 1957;64:377–388.

Grosz, BJ. The representation and use of focus in a system for understanding dialogs.. Proceedings of the fifth international joint conference on artificial intelligence (IFCAI-77); Cambridge, MA. 1977. p. 67-76.

Hanna JE, Tanenhaus MK, Trueswell JC. The effects of common ground and perspective on domains of referential interpretation. Journal of Memory and Language 2003;49:43–61.

Hayes, AF. An SPSS Procedure for Computing Krippendorff's Alpha [Computer Software]. 2005. Available from http://www.Comm.Ohio-State.edu/ahayes/macros.htm

Kim, Y.; Hill, RW.; Traum, DR. Controlling the focus of perceptual attention in embodied conversational agents.. Proceedings of the 4th international joint conference on autonomous agents and multiagent systems (AAMAS-05); Utrecht, Netherlands. 2005. p. 1997-1098.

Krahmer, E.; Theune, M. Efficient context-sensitive generation of referring expressions.. In: vanDeemter, K.; Kibble, R., editors. Information sharing: Reference and presupposition in language generation and interpretation. CSLI Publications; Stanford, CA: 2002. p. 223-264.

Kraljic T, Brennan SE. Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? Cognitive Psychology 2005;50:194–231. [PubMed: 15680144]

Krauss RM, Weinheimer S. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. Journal of Personality and Social Psychology 1966;4:343–346. [PubMed: 5969163]

Landragin, F.; Romary, L. Referring to objects through sub-contexts in multimodal human-computer interaction.. Proceedings of the seventh workshop on the semantics and pragmatics of dialogue (DiaBruk'03); Saarbrucken, Germany. 2003. p. 67-74.

Lin LI. A note on the concordance correlation coefficient. Biometrics 2000;56:324–325.

Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255–268. [PubMed: 2720055]

Marslen-Wilson W. Sentence perception as an interactive parallel process. Science 1975;189:226–228. [PubMed: 17733889]

Marslen-Wilson W. Linguistic structure and speech shadowing at very short latencies. Nature 1973;244:522–523. [PubMed: 4621131]

Matin E, Shao KC, Boff KR. Saccadic overhead – information-processing time with and without saccades. Perception & Psychophysics 1993;53:372–380. [PubMed: 8483701]

Metzing C, Brennan SE. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. Journal of Memory and Language 2003;49:201–213.

Miller GA. Some psychological studies of grammar. American Psychologist 1962;17:748–762.

Miller, GA.; Chomsky, N. Finitary models of language users.. In: Luce, RD.; Bush, RR.; Galanter, E., editors. Handbook of mathematical psychology v2. Wiley & Sons, Inc.; New York, NY: 1963. p. 419-491.

Morrow DG, Bower GH, Greenspan SL. Updating situation models during narrative comprehension. Journal of Memory & Language 1989;28:292–312.

Olejnik S, Algina J. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. Psychological Methods 2003;8:434–447. [PubMed: 14664681]

Olson DR. Language and thought: Aspects of a cognitive theory of semantics. Psychological Review 1970;77:257–273. [PubMed: 5448408]

Osgood, CE. Where do sentences come from?. In: Steinberg, DD.; Jakobovits, LA., editors. Semantics: An interdisciplinary reader in philosophy, linguistics and psychology. Cambridge University Press; Cambridge, MA: 1971.

Pechmann T. Incremental speech production and referential overspecification. Linguistics 1989;27:89–110.

Pickering MJ, Garrod SC. Towards a mechanistic theory of dialog. Behavioral and Brain Sciences 2004;7:169–190. [PubMed: 15595235]

Rinck M, Bower GH. Anaphora resolution and the focus of attention in situation models. Journal of Memory and Language 1995;34:110–131.

Roberts C. Uniqueness in definite noun phrases. Linguistics and Philosophy 2003;26:287–350.

Rossion B, Pourtois G. Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. Perception 2004;33:217–236. [PubMed: 15109163]

Salmon-Alt, S. Interpreting referring expressions by restructuring context.. Presented at the 12th European summer school in logic, language and information (ESSLLI-00).; Birmingham, UK. Aug. 2000

Salmon-Alt, S.; Romary, L. Generating referring expressions in multimodal contexts.. Paper presented at the Workshop on Coherence in Generated Multimedia (INLG-00); Mitzpe Ramon, Israel. 2000.

Schegloff E, Sachs H. Opening up closings. Semiotica 1973;7:289–327.

Schober, MF.; Brennan, SE. Processes of interactive spoken discourse: The role of the partner.. In: Graesser, AC.; Gernsbacher, MA., editors. Handbook of discourse processes. Erlbaum; Mahwah, NJ: 2003.

Schober MF, Clark HH. Understanding by addressees and overhearers. Cognitive Psychology 1989;21:211–232.

Searle, J. Speech acts: An essay in the philosophy of language. Cambridge University Press; Cambridge: 1969.

Snodgrass JG, Vanderwart M. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. Journal of Experimental Psychology: Human Learning and Memory 1980;6:174–215. [PubMed: 7373248]

Spivey-Knowlton, MJ. Integration of Visual and Linguistic Information: Human Data and Model Simulations. University of Rochester; Rochester, NY: 1997. Unpublished doctoral dissertation

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information in spoken language comprehension. Science 1995;268:1632–1634. [PubMed: 7777863]

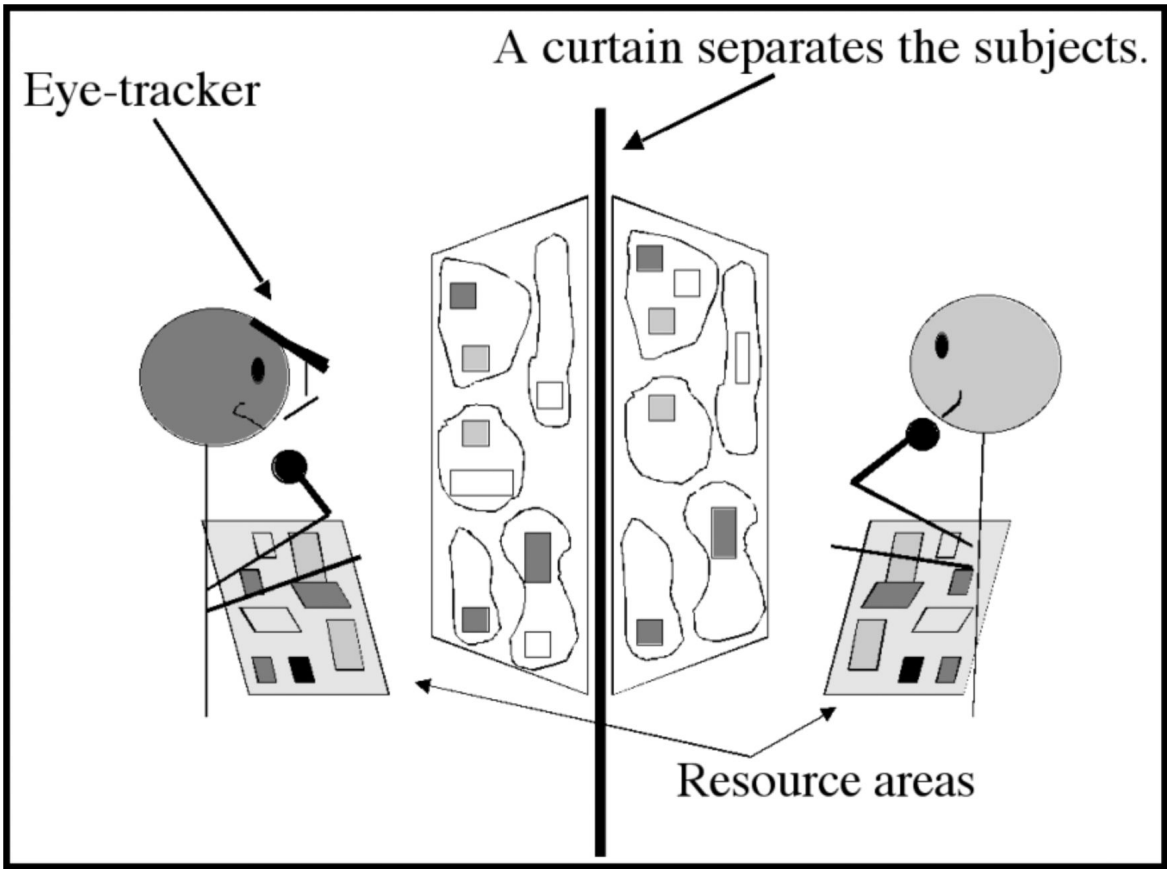
Thórisson, KR. Simulated perceptual grouping: An application to human-computer interaction.. Proceedings of the 16th annual conference of the cognitive science society; Atlanta, GA. 1994. p. 876-881.

Traum, DR. A Computational Theory of Grounding in Natural Language Conversation. University of Rochester; Rochester, NY: 1994. Unpublished doctoral dissertation

Trueswell, JC.; Tanenhaus, MK. Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions. MIT Press; Cambridge, MA: 2005.

Zollo, T.; Core, M. Automatically extracting grounding tags from BF tags.. Proceedings of the 37th meeting of the association for computational linguistics (ACL) workshop on standards and tools for discourse tagging; College Park, Maryland. 1999. p. 109-114.

Zwitserlood P. The locus of the effects of sentential-semantic context in spoken-word processing. Cognition 1989;32:25–64. [PubMed: 2752705]



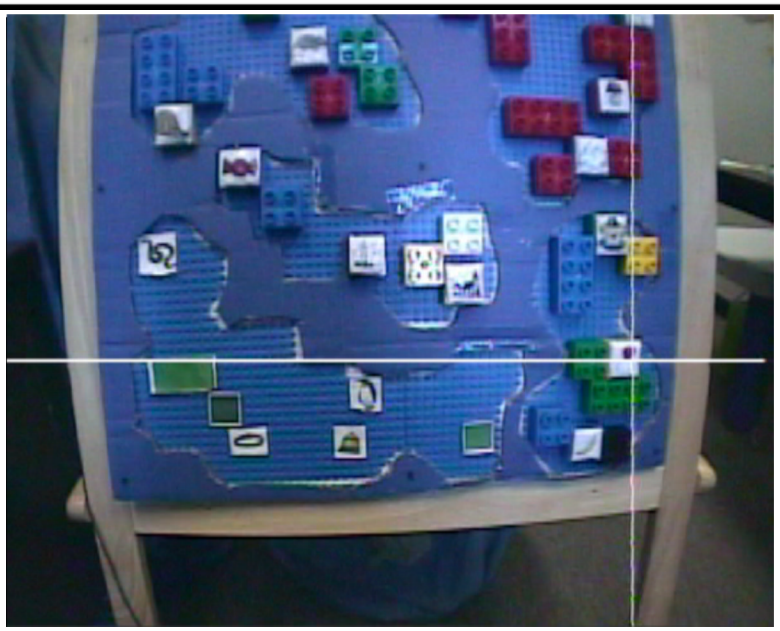


Fig. 1a.
Schematic illustration of the task used in Experiments 1 and 2.
Fig. 1b: Image of the game-board from the eye-tracked participant's viewpoint, mid-way through the task in Experiment 1. Eye position is superimposed, indicated by the white crosshair.

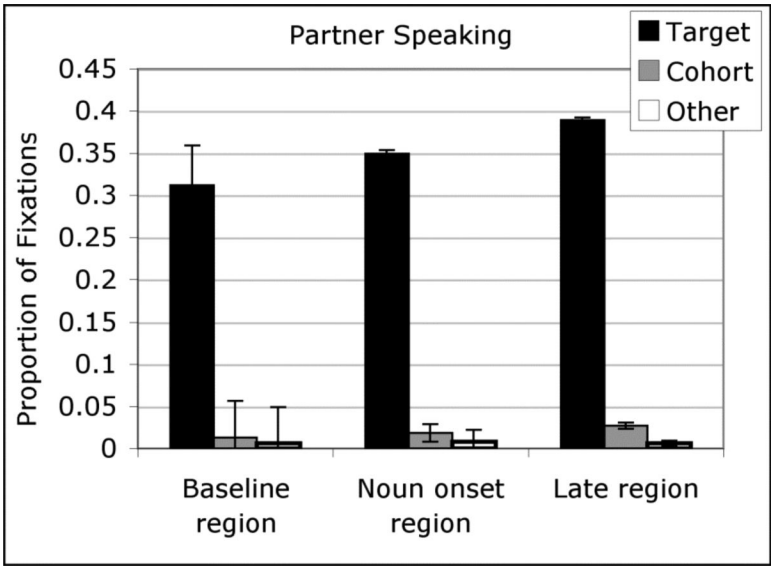


Fig. 2.

Partner-generated expressions. Mean proportion of fixations to target, cohort and other blocks in the baseline region, the region beginning with the onset of the head noun, and the final region. Error bars indicate standard error of the mean (SEM).

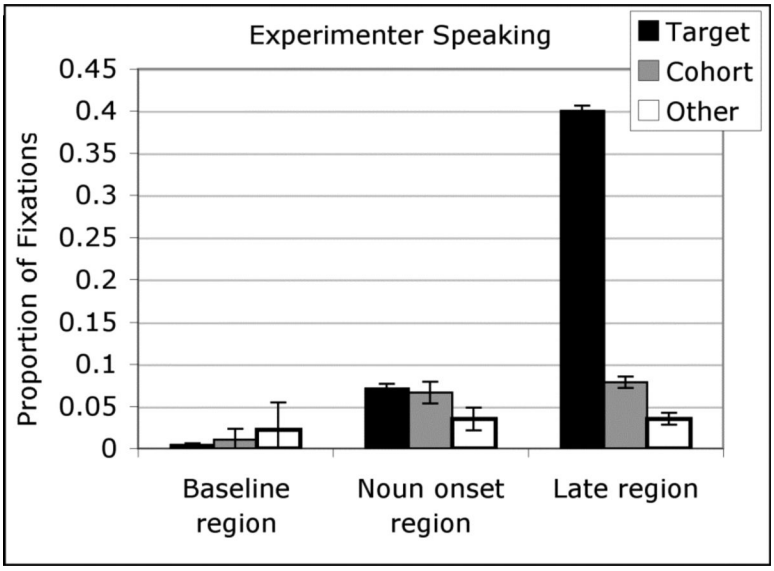


Fig. 3.

Experimenter-generated expressions. Mean proportion of fixations to target, cohort and other blocks in the baseline region, the region beginning with the onset of the head noun, and the final region. Error bars indicate standard error of the mean (SEM).

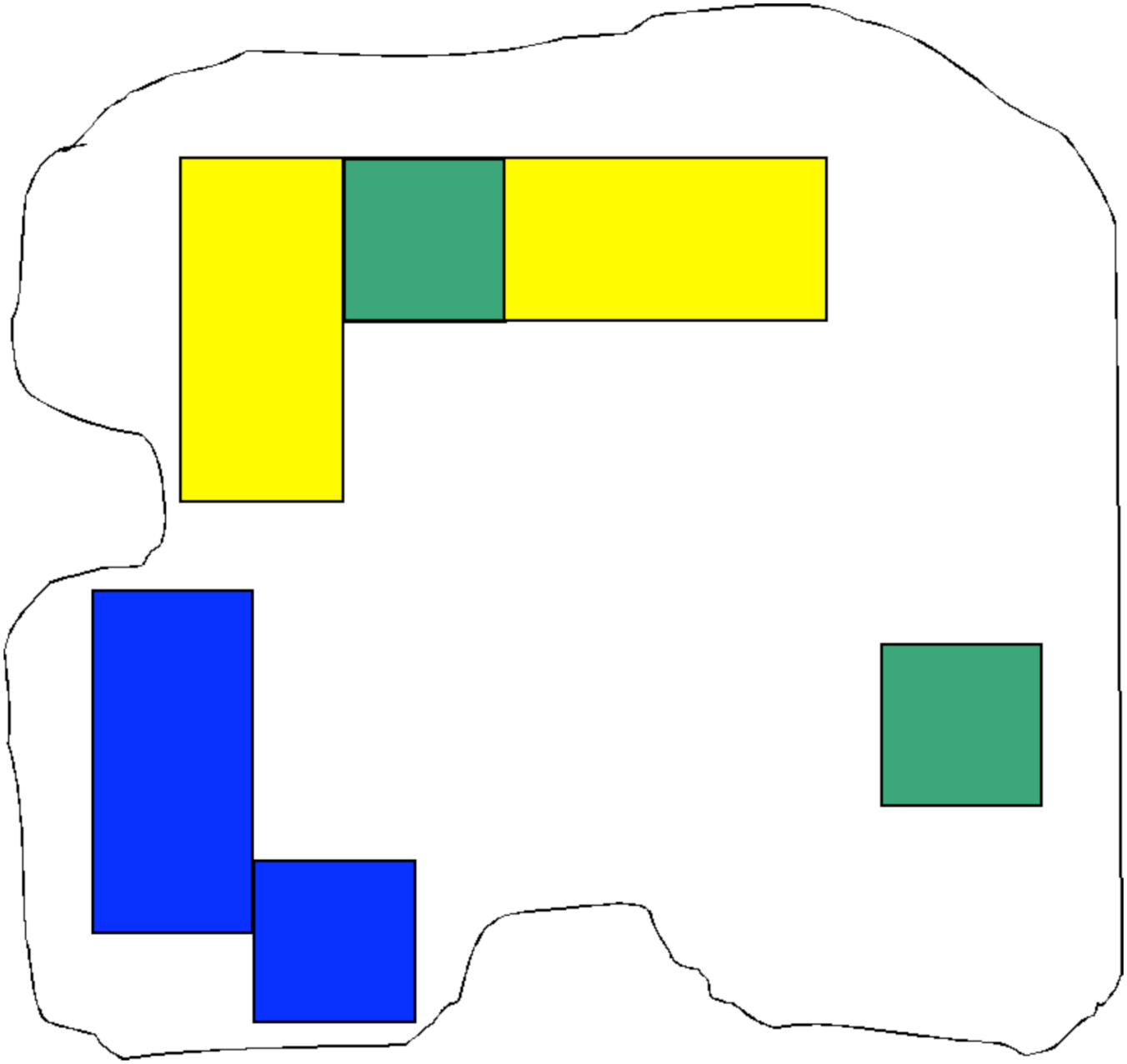


Fig. 4.
Schematic of a portion of the workspace during the exchange in (7) from Speaker #2's perspective. The yellow blocks are the two lightly-shaded rectangles at the top.

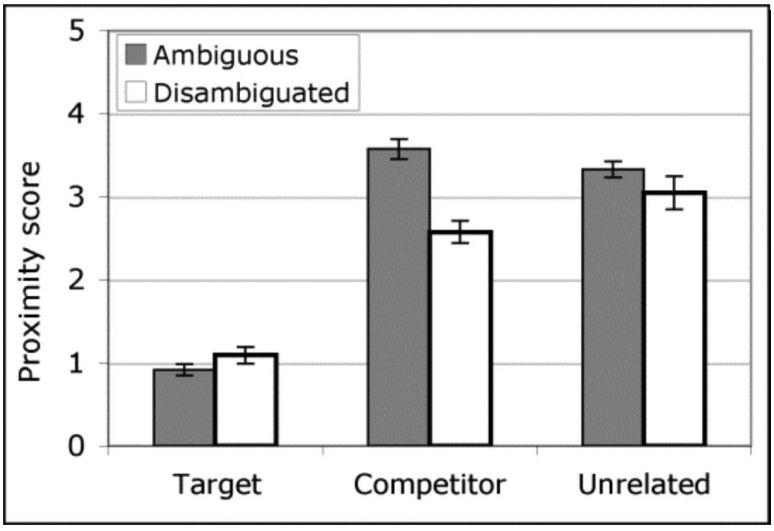


Fig. 5.
Proximity scores for target, competitor and unrelated blocks in the sub-area of the target during ambiguous and disambiguated expressions. Lower numbers indicate the block was more proximal to the last mentioned block. Error bars indicate standard error of the mean (SEM).

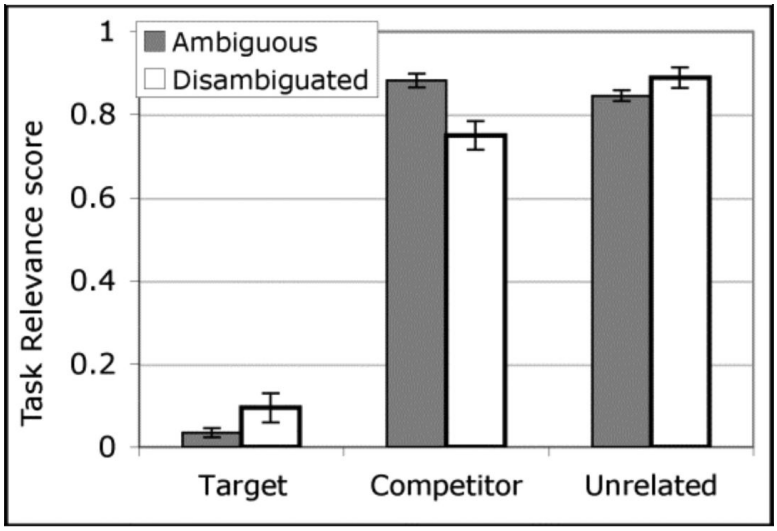


Fig. 6.
Task scores for target, competitor and unrelated blocks in the sub-area of the target during ambiguous and disambiguated expressions. Lower numbers indicate the block was more task-relevant. Error bars indicate standard error of the mean (SEM).

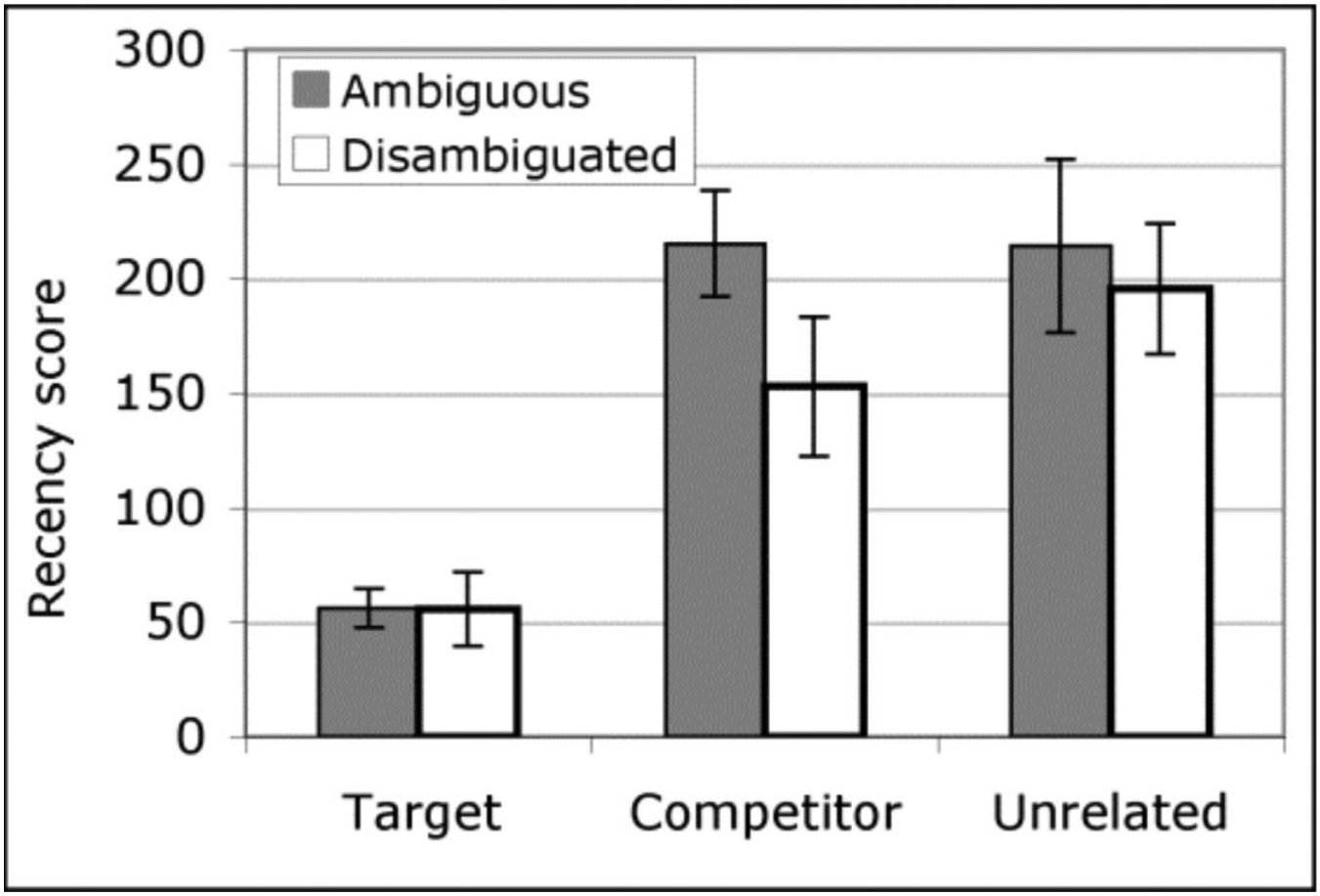


Fig. 7.

Recency scores for target, competitor and unrelated blocks in the sub-area of the target during ambiguous and disambiguated expressions. Lower numbers indicate the block was mentioned more recently. Error bars indicate standard error of the mean (SEM).

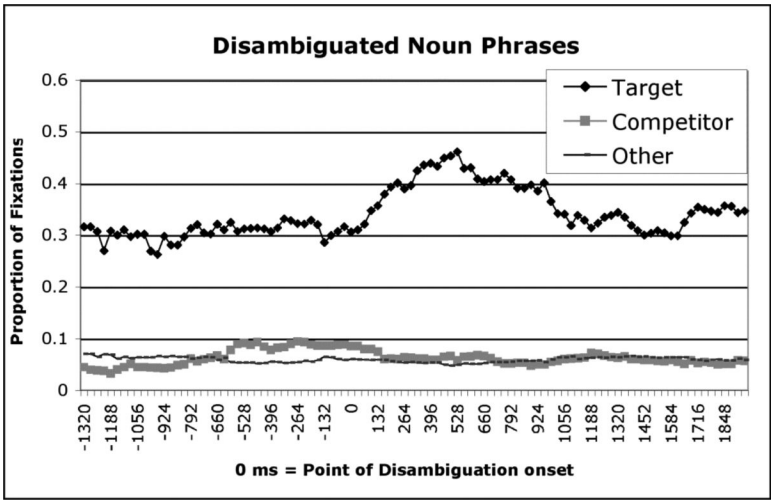


Fig. 8.
Disambiguated noun phrases. Proportion of fixations to target, competitor and other blocks. 0 ms = the point-of-disambiguation.

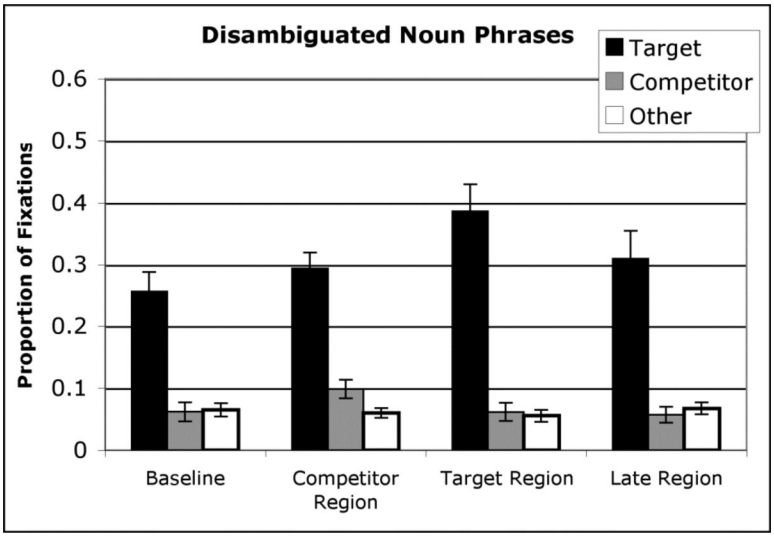


Fig. 9.

Disambiguated noun phrases. Mean proportion of fixations to target, competitor and other blocks in the four critical time regions, aligned at the point-of-disambiguation. Error bars indicate standard error of the mean (SEM).

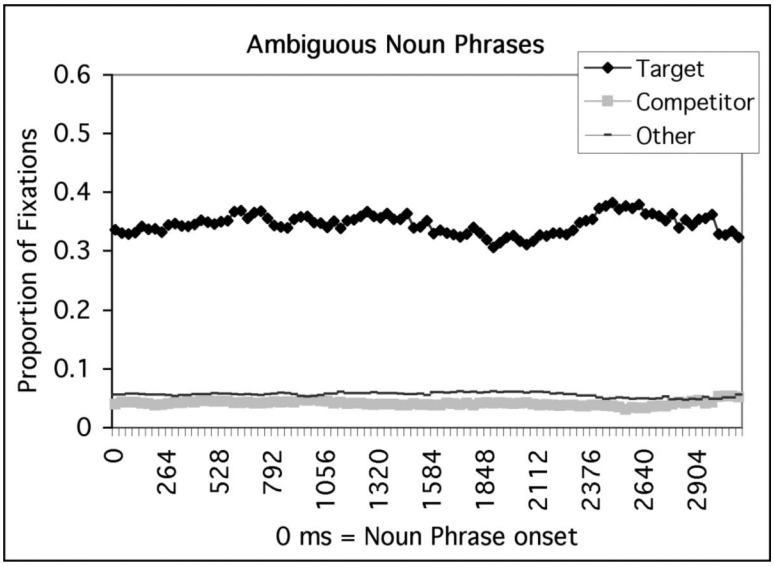


Fig. 10.

Ambiguous noun phrases. Proportion of fixations to target, competitor and other blocks. 0 ms = noun phrase onset.

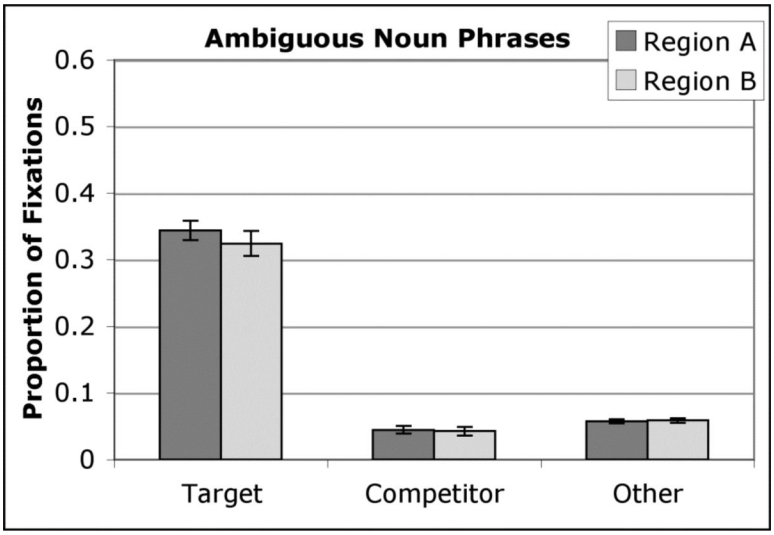


Fig. 11.

Ambiguous noun phrases. Mean proportion of fixations to target, competitor and other blocks in regions before (region A) and after (region B) the average point-of-disambiguation for disambiguated expressions. Error bars indicate standard error of the mean (SEM).

Table 1

Results of three hierarchical linear models for the relationship between ambiguity[a], task, proximity and recency scores, and competitor fixations.

Model[b]	Fixed Effects	coefficient	SE	t-ratio	df[c]	p-value
#1	task	−.011851	.017426	−0.680	11	=.510
	proximity	−.025193	.006550	−3.847	11	<.01
	recency	.000044	.000045	0.972	11	=.352
	ambiguity	.046458	.018959	2.450	11	<.05
#2	task	.061205	.061159	1.001	11	=.339
	ambiguity	.127344	.045326	2.810	11	<.05
#3	task	−.039824	.014661	−2.716	11	<.05
	ambiguity	.022439	.024970	0.899	11	=.388

[a] Ambiguity was coded as 0=ambiguous; 1=disambiguated; Task was coded as 0=predicted by the task; 1=not predicted by the task.

[b] Model 1 included each eye-tracking trial (n=751). Model 2 included trials (n=130) on which there was a competitor block adjacent to the last mentioned block. Model 3 included trials (n=621) on which there was not a competitor block adjacent to the last mentioned block.

[c] (Approximate).