REFERENCE RESOLUTION BY HUMAN PARTNERS IN A
NATURAL INTERACTIVE PROBLEM-SOLVING TASK

Ellen Campana, Sarah Brown-Schmidt, and Michael K. Tanenhaus
Department of Brain and Cognitive Sciences
University of Rochester
ecampana@bcs.rochester.edu, sschmidt@bcs.rochester.edu, mtan@bcs.rochester.edu

ABSTRACT
We examined how listeners circumscribe referential domains for referring expressions by monitoring participant’s eye movements as they engaged in a natural interactive problem-solving task with another native participant. This research had two goals: (1) determine whether existing psycholinguistic methodologies for studying online processing can be extended to interactive conversation, and (2) assess whether studying language in this manner can provide insight into the theoretical limitations of various strategies for reference resolution commonly employed in computational spoken dialogue systems.

1. BACKGROUND
Rapid increases in the accuracy and speed of automatic speech recognition, and the increased availability of off-the-shelf text-to-speech systems, has fueled great interest in spoken dialogue systems. To date, the most successful applications of such systems have been in domains requiring a more or less one-sided discourse interaction, such as tutoring and command-control environments. However, the common goal of these systems and many others is to move toward a more balanced mode of collaborative interaction [1] with incremental comprehension and generation. Progress with respect to this goal would be greatly facilitated by the availability of detailed information about how humans process and generate language in real-time during collaborative interaction. Moreover, the same methods used to collect these data could also be used for system evaluation.

Until recently, the need for fine-grained reaction time measures and the dominant theoretical commitment to language as a modular process restricted psycholinguistic methodologies to those involving de-contextualized language presented according to a pre-defined script. These studies tended to ignore factors known to be common in natural interactive communication such as disfluencies, false starts and repairs, and phrasal utterances.

Although there is an alternative psycholinguistic tradition that focuses on natural interactive conversation, it has not focused on real-time processing. For example, researchers have investigated how conversational partners collaborate to define referential terms[4], create ‘conceptual pacts[3]’ and ground specific referential domains during the course of a conversation. One problem with studies conducted in natural interactive settings is that no methods currently exist for extracting the sort of temporally fine-grained information that is necessary to develop explicit models of language processing of the sort that might be useful for improving spoken dialogue systems.

Recently, the advent of light-weight head-mounted eye-tracking systems has made it possible to investigate real-time comprehension in more natural tasks, such as tasks where participants follow spoken instructions to manipulate objects in a task-relevant “visual world” [10]. Fixations to task-relevant objects are closely time-locked to the unfolding utterance, providing a continuous real-time measure of comprehension processes at a temporal grain fine enough to track the earliest moments of lexical access, parsing, and reference resolution [9]. While an improvement over previous work, these studies are still not an ideal model for spoken language systems, as they focus on scripted, one-sided language in an impoverished task context.

The goal of the present study was to explore the feasibility of examining real-time comprehension processes during natural, unscripted, interactive conversation. We focused on the comprehension of definite referring expressions, such as “the red block” and “the cloud”. Definite reference provides an ideal domain for a first investigation for several reasons. First, definite reference is one of the most ubiquitous and central components of natural language. Second, use of definite reference assumes that a referent will be uniquely identifiable within a circumscribed referential domain. Much of the strongest evidence for the collaborative model of language processing comes from demonstrations that people collaborate to define referential domains. Third, work with restricted utterances has established two clear empirical results that allow one to track the time course of reference resolution: lexical competitor (cohort) effects and ‘point of disambiguation’ effects. We view the current work as important for both theoretical and methodological reasons. On the methodological side, if it is possible to evaluate fine-grained comprehension and production processes in natural interactive conversation, then these same methods could be adapted as tools for evaluating dialogue systems. On the theoretical side, information about how people generate and interpret referential expressions in natural interactive conversation could help inform design and architectural decisions in dialogue systems.

When listeners are instructed to pick up or move an object, such as a racket, fixations to the target object begin as early as 200 ms after the onset of the noun [2]. Eye-movements launched at this point in the speech stream are equally likely to be directed to the eventual referent and other objects with names that are also consistent with the speech signal, such as a raccoon. However looks to these competitors, hereafter
“cohort” competitors are reduced or eliminated when contextual and conversational factors make a cohort an implausible referent. Thus we can use cohort effects to infer the degree to which these factors restrict initial referential domains.

One of the most striking sources of evidence for rapid restriction of referential domains comes from point of disambiguation effects. For example, Eberhard, Spivey-Knowlton, Sedivy and Tanenhaus presented subjects with displays containing a variety of differently colored shapes, as subjects listened to instructions such as Click on the red triangle [5]. In a subset of trials, the color of the target item was not shared with any other items in the referential domain. In these trials the referentially disambiguating information was the color, which was conveyed in the prenominal adjective. In the remaining trials, the target item was the same color as another item in the referential domain. For example, the display accompanying the instruction Click on the red triangle might contain a red circle and a red triangle. In these trials the referentially disambiguating information came at the noun. Eye movements to the target were again closely time-locked to the speech. Looks to the target increased dramatically immediately following the point of disambiguation (POD), whether it came at the adjective or the noun.

In the present experiment we monitored eye-movements as pairs of participants, separated by a curtain, worked together to arrange blocks in matching configurations and confirm those configurations. The characteristics of the blocks afforded comparison with findings from scripted experiments investigating language-driven eye-movements, specifically those demonstrating cohort effects and incremental reference resolution. We investigated: (1) whether these effects could be observed in a more complex domain during unrestricted conversation, and (2) under what conditions the effects would be eliminated, indicating that factors outside the speech itself might be operating to circumscribe the referential domain.

2. METHOD

We tested 4 pairs of participants from the University of Rochester, who were paid for their participation in the study. The discourse partners each had an array of blocks and a board on which to place them. The boards were partially covered, creating 5 distinct sub-areas. Initially, sub-areas contained 56 stickers representing blocks. The task was to replace each sticker with a matching block. While partners’ boards were identical with respect to sub-areas, partners’ stickers differed: Every place that one partner had a sticker, the other partner had an empty spot. Pairs were instructed to tell each other where to put blocks so that in the end their boards would match. No other restrictions were placed on the interaction. The entire experimental study lasted approximately 2.5 hours. For each pair we recorded the eye movements of one partner, and the speech of both partners.

The initial configuration of the stickers was such that the color, size, and orientations of the blocks would encourage the use of complex noun phrases and grounding constructions. 19 of the stickers (and the corresponding blocks) contained pictures similar to those used in the study described above[2]. We used a full-color version [6] of a large corpus of normed pictures, balanced for their linguistic codability [7]. We selected pairs of these pictures that referred to objects which had initially acoustically consistent names (cohort competitors). Half of the cohort competitor stickers were placed such that both cohort competitor blocks would be placed in the same sub-area of the board, and half of the cohort competitor stickers were placed such that the cohort competitor blocks would be placed in different sub-areas of the board. All of the cohort competitor pairs were separated by approximately 3.5 inches. We examined the eye movements of one discourse partner with respect to the speech generated by the other discourse partner.

3. RESULTS

The conversations for each of the 4 pairs were transcribed. We present eye-tracking analyses for 2 of the pairs, as we are still analyzing eye-movements from the latter two pairs. The non-eye-tracked partner of each pair generated approximately 100-150 definite references to blocks during the course of the conversation. While the length of the conversation prevented us from initially analyzing more than 4 pairs, the large number of ‘trials’ generated by each pair gave us enough statistical power to circumvent this problem.

An ISCAN eyetracking visor was used [12]. The image of the eye-tracked partner’s board, and their superimposed eye position, along with the entirety of the conversation (both participants’ voices) were recorded using a frame-accurate digital video recorder (a SONY DSR-30). Eye-movements were analyzed at the onset of the definite reference, and continued 2000ms after the NP was completed. There was a high degree of variability in the length of utterances, especially those to color blocks.

References to blocks which had cohort competitors (approximately 75 references per pair) were expected to reveal similar cohort effects as observed in more restricted experimental paradigms. To our surprise, we observed only one look to a cohort competitor during both 2 1/2 hour conversations we have analyzed thus far. We do not think this null result is due to poor stimulus design, as we did observe looks to cohort competitors under special circumstances. Periodically, subjects needed to remove the eye-tracker to take a break. When we put the tracker back on and re-calibrated, we tested the calibration by asking the subjects to look at different items on the board. Under these circumstances the referential domain is not restricted by conversational constraints. Here we saw clear cases of subjects initially looking at the cohort competitor before looking at the intended referent.

Each pair provided us with approximately 75 trials of data for eye-movements elicited by definite references to colored blocks. Two researchers coded the definite NPs for their point-of-disambiguation (taking into account only the NP itself and the visual context), and resolved any coding differences. The POD was the point at which the NP uniquely identified a referent, given the visual context at the time. Average POD was 858ms (26 frames) following NP onset. Eye-movement analyses for NPs which did have a unique linguistic point of disambiguation (50.3%) were analyzed separately from those in which the NP itself did not uniquely specify a referent, given the visual context. The eye-tracking analysis was restricted to cases where at least one competitor block was present. As a result, the number of ambiguous trials (contained no POD)
used for the analysis was 77, while there were only 24 for the disambiguated trials.

Eye-movement analyses (planned comparisons) were performed on 800ms epochs for both ambiguous and disambiguated NPs (1 RM ANOVA at each epoch, followed by Bonferroni tests). Eye movements elicited by disambiguated references showed clear point of disambiguation effects; within 200ms after the disambiguation point, looks converged on the Target block (see Figure 1), F(2,20)=64.03, p<.0001. Bonferroni tests reveal significantly more looks to Target than both Competitor and Other (p’s <.0001). Before the disambiguation point, subjects were looking equally at the Target and Competitor block(s), but a significant effect of condition is due to more looks to Target than Other, F(2,19)=6.765, p=.01. Bonferroni test for Target vs. Other, p=.001. Finally, in the baseline region (-1122 to -600ms before the average POD), we see no effect of condition, as subjects looked equally at the different blocks F<.8, p>.45. This replicates the point-of-disambiguation effect seen by Eberhard, et al. (1995), demonstrating we were successful in using a more natural task to investigate on-line language processing.

![Figure 1: Proportion of fixations to Targets, Competitors, and Other blocks by time in ms for Disambiguated NPs. Graph is centered by item with 0 frames = POD onset.](image1)

Most remarkably, ambiguous utterances elicited significantly more looks to the Target than unambiguous utterances (see Figure 2).

Moreover, fixations were primarily restricted to the referent shortly after onset of the definite reference; we observed significantly more looks to targets than competitors within the first 200 ms of NP onset, F(2,53)= 18.37, p<.0001, Bonferroni tests reveal significantly more looks to Target than both Competitor and Target, p’s<.0001).

In comparison, a subsequent analysis of the disambiguated utterances, centered at NP onset, didn’t reveal significant differences between target and competitor until 1000ms following NP onset (we did see more looks to Target than Other at this window). These results suggest that (1) speakers systematically use less specific utterances when the referential domain has been pragmatically constrained; (2) the attentional states of speakers and addresses become closely tuned; and (3) utterances are interpreted with respect to referential domains circumscribed by pragmatic and/or discourse constraints.

![Figure 2. Proportion of fixations to Targets, Competitors, and Other blocks by time in ms for Ambiguous NPs. Zero ms = NP onset.](image2)

In order to further examine what information listeners were using to understand these ambiguous references, we performed a detailed analysis of ambiguous references. We identified four different factors that contributed to the intelligibility of these ambiguous references: 1) Recency of mention of the target referent; 2) Proximity of target referent to the last mentioned block; 3) Task-based constraints (such as limitations on block placement due to the size and shape of the board); 4) Collaboratively-defined terms. Of the 77 ambiguous NPs used in this analysis, approximately 29% were influenced by Recency, 19% by Proximity, 32% by Task-based constraints, and 8% by Collaborative terms.

As we analyze more of the data it will be possible to investigate the role that these constraints play in the comprehension of even those definite NPs that uniquely specify a referent given the visual context. We can do this by extracting those NPs for which different constraints predict points of disambiguation and noting which of the constraints best predicts the pattern of eye-movements that we observe during comprehension. Unfortunately, however, the data is currently too sparse for this analysis.

### 4. CONCLUSIONS

In this experiment we investigated 1) whether existing psycholinguistic methodologies for studying online processing can be extended to interactive conversation, and 2) whether studying language in this manner can provide insight into the theoretical limitations of various strategies for reference resolution commonly employed in computational spoken dialogue systems. We were successful on both counts. We did observe incremental reference resolution in certain contexts, and in the contexts in which it was not observed we were able to identify a number of constraints that seemed to be operating to circumscribe the referential domain. The relative frequency of underspecified referential expressions suggests that spoken dialogue systems would have problems with this data. The findings also suggest ways in which reference resolution algorithms might be improved.

#### 4.1 Implications for spoken dialogue systems

The simplest and most straightforward strategy for reference resolution that one could employ would be to
consider only the NP itself and the entities in the context that are available for reference. If a given NP does not uniquely specify a referent on this basis, then the system would be forced to request more information from the user in order to identify the referent. We chose this as baseline because it is the simplest theoretically possible strategy, even though in practice it is rarely used in any but the most restricted of domains. A reference resolution strategy based on only the immediate context and the NP itself could achieve a maximum of 50.3% accuracy on our data.

A more sophisticated approach, one which is commonly employed in spoken dialogue systems, is to use discourse factors such as recency and subjecthood in addition to the immediate context and the NP itself to resolve referents. In our corpus this strategy could potentially account for an additional 32% of the cases, but critically 17.7% would remain unaccounted for.

It is clear that pragmatic factors must also be considered in order to achieve human-like performance. A resolution strategy that incorporated top-down task structure and physical proximity information on top of language and discourse information would theoretically be able to account for 15.5% more of the cases in this corpus. An additional strategy of interpreting collaboratively defined referring expressions and relations on-the-fly would account for 8% of the total data.

The results from this experiment suggest that as dialogue systems become more complex with respect to the types of information that they rely upon for reference resolution, it will be increasingly important to consider the relative weights of the different factors. For example, in 14.2% of the definite NPs that we looked at, linguistic recency predicted the “wrong” referent, while task constraints predicted the “right” referent, as in the following example:

2: k, and then there’s a penguin, it’s to the right of the upper dark green one, but there’s a space in between and he’s two above
1: yup
2: ok, and then three over, is the dark green one —- from the penguin

In this example the participants are verifying the location of blocks they have already placed on the board (figure 3).

![Figure 3. Configuration of blocks present during comprehension of “the dark green one” in the transcript excerpt. The block labeled “P” has a picture of a penguin on it, and dark and light shaded blocks are dark and light green.](image)

Eye movements during comprehension of “the dark green one” (bolded) indicate that the human interlocutor was not confused by these conflicting factors, even though most computational strategies of reference resolution would fail on this example. Crucially, the pattern of data we see in examples like this suggest that using discourse factors as a first-pass in reference resolution will result in systems exhibiting behavior that is quite unlike that of humans. Additionally, the fact that 49.7% of analyzed NPs were ambiguous yet easily understood, is a critical observation for work in generation. A system that always produced fully specified NPs might confuse human users, in much the same way that use of a full NP when a pronoun is expected can introduce processing difficulty. In order to achieve more human-like performance in spoken dialogue systems it will be useful to have data concerning how the factors constraining human reference resolution interact during natural interactive problem-solving tasks. The methodology introduced in this paper is promising because it allows researchers to investigate questions related to this issue.

5. REFERENCES


6. ACKNOWLEDGEMENTS

This research was supported by NIH HD 27206. We thank our colleague James Allen for helpful comments.