Incremental Referential Domain Circumscription during Processing of Natural and Synthesized Speech

Mary D. Swift (mswift@ling.rochester.edu)
Department of Linguistics, University of Rochester
Rochester, NY 14627

Ellen Campana (ecampana@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627

James F. Allen (james@cs.rochester.edu)
Department of Computer Sciences, University of Rochester
Rochester, NY 14627

Michael K. Tanenhaus (mtan@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627

Abstract

We present experimental evidence from a study in which we monitor eye movements as people respond to pre-recorded instructions generated by a human speaker and by two text-to-speech synthesizers. We replicate findings demonstrating that people process spoken language incrementally, making partial commitments as the instruction unfolds. Specifically, they establish different referential domains on the fly depending on whether a definite or indefinite article is used. Importantly, incremental understanding is observed for **both** natural speech instructions and synthesized text-to-speech instructions. These results, including some suggestive differences in responses with the two text-to-speech systems, establish the potential for using eye-tracking as a new method for fine-grained evaluation of dialogue systems and for using dialogue systems as a theoretical and experimental tool for psycholinguistic experimentation.

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 (e.g., Allen, Byron, Dzikovska, Ferguson, Galescu & Stent, 2001; Zue, Seneff, Glass, Polifroni, Pao, Hazen & Hetherington, 2000). As the sophistication of such systems increases, we can expect applications to more open-ended domains with larger vocabularies and more varied utterance types. The feasibility of such systems raises both applied and theoretical issues for work on natural language processing that crosses disciplinary boundaries. We focus on two issues here. The first, a computational issue, addresses the need for developing better evaluation tools for dialogue systems, especially tools that can evaluate comprehension on an utterance-by-utterance and within-utterance basis. The second, a psycholinguistic issue, is the possibility that in the near future implemented dialogue systems could serve as a powerful tool for developing and testing psycholinguistic models by allowing stimuli to be generated 'on the fly,' conditioned on the current state of the discourse.

A necessary prerequisite for enabling both of these goals is that people respond to synthesized speech in much the same way as they do to natural speech. We present experimental evidence from a study in which we monitor eye movements as people respond to pre-recorded instructions generated by a human speaker and by two text-to-speech synthesizers. We replicate findings demonstrating that people process spoken language incrementally, making partial commitments as the instruction unfolds. More specifically, listeners establish referential domains on the fly depending on whether a definite or indefinite article is used.

Eye movements as an evaluation tool

Spoken utterances unfold over time, resulting in a stream of temporary ambiguities. For example, as the instruction *Click on the beaker* unfolds, the word *beaker* is briefly consistent with multiple candidates, including *beetle*, *beeper*, and *speaker*. Numerous psycholinguistic studies demonstrate that people comprehend utterances continuously, entertaining multiple lexical candidates (e.g., Marslen-Wilson, 1987), making provisional commitments at points of syntactic ambiguity, and resolving reference incrementally (e.g., Altmann, 1998; Tanenhaus & Trueswell, 1995). Recent studies using eye movements to a task-relevant object in a visual workspace as people follow spoken instructions provide striking evidence for both incremental understanding and rapid integration of multiple constraints (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995; 1996; Tanenhaus, Magnuson & Chambers, forthcoming). For example, if the instruction *Click on a beaker* is presented in a context in which there are two icons of beakers and two icons of beetles, then reference will be delayed until the word *beaker* is disambiguated phonetically

(Allopenna, Magnuson & Tanenhaus, 1998). However, if there is only a single beetle, then reference will be speeded because the indefinite article *a* implies that there should be multiple referents – a condition met by the beakers but not by the beetle. Similarly, reference resolution for *Click on the beetle* will be facilitated because the beetle is the only icon that is unique. However, if the alternatives do not satisfy the uniqueness conditions associated with the article, e.g., if there is only one beaker and two beetles, and the instruction is *Click on a beaker*, then listeners are temporarily confused, looking first at the beetles before clicking on the beaker (Hanna, 2001).

If such incremental behavior carries over into recognition of synthesized utterances, then it should be possible to develop evaluation measures that can track the temporary commitments listeners make as they are processing utterances. This could establish a new evaluation methodology for speech synthesis and dialogue systems that could provide much more fine-grained information than is possible with existing techniques. This is particularly important for evaluating the quality of speech synthesis because crucial information about potential reference resolution, such as the form of an article, is carried by monosyllabic unstressed words that exhibit considerable variability with local phonetic context, as well as the overall prosodic environment of an utterance. Eye-tracking measures are good potential candidates for such an evaluation methodology because they can be incorporated into natural tasks and are well suited for any application in which the user is working within a visual workspace. The present study explores the feasibility of using eye tracking for this purpose by examining (a) whether processing is incremental when instructions are generated from a text-to-speech system and (b) whether such investigations might reveal subtle problems with synthesized speech that could impair real-time performance in natural tasks involving reference resolution.

Dialogue systems as a psycholinguistic tool

There is a growing awareness in the psycholinguistic community of the importance of examining real-time processing in natural tasks using conversational language. The advent of head-mounted eye tracking has begun to make such investigations possible. However, the field is currently facing both a theoretical and a methodological challenge. The long-term theoretical challenge is that we need theories of discourse processing that can incorporate the notion of a rich, dynamic context that characterizes the type of knowledge that listeners and speakers bring to bear on real-time interactive conversation. We suggest that practical dialogue systems, that is, dialogue systems in which participants focus on a specific task such as tutoring or problem solving, are the right grain to provide such models, if they are modified to address real-time generation and understanding. The shorter-term methodological challenge is that we need methods of generating utterances on the fly based on the current state of the discourse, in order to allow testing of alternative hypotheses, by presenting trials on which, for example, an inappropriate referential expression is used. Such trials cannot be plausibly generated by a confederate speaker, nor is it feasible to use pre-recorded instructions in any but the simplest experiments. We believe that it will soon be possible to use practical dialogue systems for this purpose. However, a crucial precondition is to determine whether listeners do indeed process synthesized utterances incrementally.

Experiment

The current experiment was intended as an initial investigation of the utility of using eye movements to evaluate spoken dialogue systems and using text-to-speech utterances in psycholinguistic experiments. We addressed the following question: Would listeners use the presence of an indefinite article compared to a definite article to differentially circumscribe potential referents as an expression unfolds? We addressed this question by examining eye movements within displays containing a pair of identical shapes and two unique shapes, using instructions such as *Click on the/a square*. Previous research with experimenter-generated instructions demonstrates that listeners assume that a definite article introduces a uniquely describable referent, whereas an indefinite article assumes that more than one referent meets the referential description (Chambers, Tanenhaus, Eberhard, Filip & Carlson, in press; Hanna, 2001).

Method

Fifteen members of the University of Rochester community were paid for their participation in this study. All participants were native speakers of English and had normal or corrected-to-normal vision. In the experimental trials, participants saw a visual display (described below) and heard an auditory instruction (one of three voice conditions) directing them to click on one of the objects on the screen. We used a within design, so each participant heard all three voice conditions (synthesizer 1, synthesizer 2 and the human voice), which were counterbalanced across experimental trials. The auditory stimuli were generated using two commercially available text-to-speech synthesizers and a digitally recorded human voice. For the human voice auditory stimuli, each instruction sentence was read aloud by an adult male volunteer and recorded with a TASCAM portable DAT recorder. The recorded voice instructions were then digitized using the SoundEdit 16 program. All auditory stimuli were minimally adjusted digitally so that the critical noun phrases were comparable in length for all three voices.

Eye movements were monitored using a lightweight head-mounted pupil/corneal reflection tracking system (ISCAN, model RK-726PCI). Calibration was monitored throughout each trial, and adjustments were made between trials if necessary. The experimental materials were presented with the PsyScope 1.0 program on a Power

Macintosh 7100/66 with a 15” color monitor.

During the experimental session, participants were seated at a comfortable distance from the computer monitor. For each trial a grid (Figure 1) appeared on the screen. The participant then clicked on the bull’s eye in the center of the grid to hear the auditory instruction, e.g., *Click on the heart*, which began playing 2000 ms after the mouse click. When the participant clicked on the target object the grid was replaced by a white screen with the printed instruction *Click here for the next trial* in a random location on the screen.

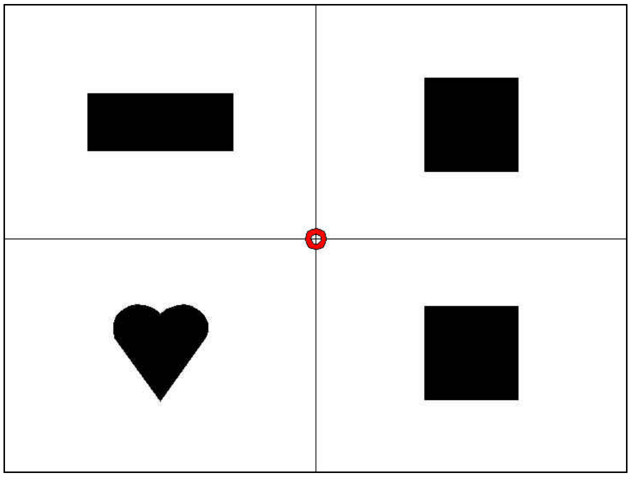

Figure 1: Sample trial screen for the experiment

Each person participated in 24 such trials, 8 in each speech condition. The voice conditions were counterbalanced across experimental trials in three lists. The order of the voice condition was pseudo-randomized so that the same voice would occur in no more than two consecutive trials. Half of the trials involved a definite noun phrase, e.g., *Click on the heart*, and half involved an indefinite noun phrase, e.g., *Click on a square.* The definite and indefinite articles were never used infelicitously – shapes referred to with an indefinite article were always duplicated and shapes referred to with a definite article were always unique.

Results

Participants clearly made use of information about definiteness when comprehending both human and synthesized speech. Recall that the display contained four objects: two were identical (duplicated) and two were unique. For a definite instruction one of the unique objects was the target. For an indefinite instruction, the participant could select either of the duplicated shapes. For instructions with definite articles (e.g., *Click on the heart*) participants were more likely to look at the unique distractor than either of the duplicated distractors (F(2)=310.38, MSE=7.34, p<.01), and there was no interaction with voice type. For instructions with indefinite articles (e.g., *Click on a square*) participants were more likely to look at either of the duplicated items than at the definite distractors (F(1)=117.52, MSE=10.29, p<.01). Again, there was no interaction with voice type.

Let us first consider the trials with instructions containing definite articles. Figure 2 shows the proportion of looks over time to the target, the unique unrelated item, and the two duplicate unrelated items in the trials with a definite article for the human voice condition. The zero point on the x-axis corresponds to the onset of the noun phrase, e.g., *the heart*. Participants clearly use the definiteness information carried by the article because looks to the duplicate unrelated items subside approximately 100 milliseconds before the target is distinguished from the unique unrelated item. Thus the items that are consistent with the definite article are first disambiguated from the items that are not consistent with the definite article, and then the target is disambiguated from the unrelated item.

The data from the two synthesized voice conditions follow the same general pattern. Specifically, the disambiguation between unique and duplicated items (i.e., definite vs. indefinite) occurs approximately 100 milliseconds before the two unique items (i.e., definite target vs. definite unrelated) are disambiguated in each of the synthesized voice conditions.

There is, however, an important difference between the human and the synthesized voice conditions – the disambiguation points occur later in the synthesized voice conditions than in the human voice condition (Figure 3). This difference is not due to differences in the length of the articles in the three voice conditions – we have verified that these did not differ.

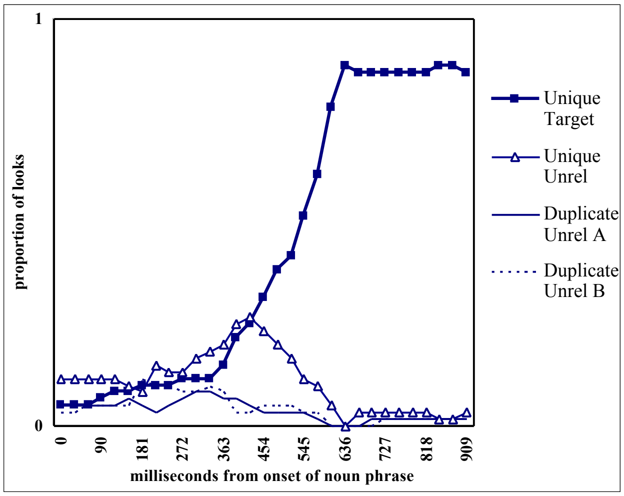

Figure 2: Results from the human voice condition: Proportion of looks to all items for definite instructions

One explanation for this difference could be that participants have greater difficulty understanding the synthesized voices. During debriefing, all participants reported that they had heard at least two distinct voices during the experiment, and at least one of these voices was readily identifiable as synthetic.

We evaluated this hypothesis more formally by conducting a simple voice judgment survey. A new group of 16 participants listened to the auditory stimuli without the visual context and wrote down what they thought they heard for each trial. We compared their responses to the intended speech and found no differences in accuracy between the voice conditions for the definite instructions – in fact performance was at ceiling for all definite instruction auditory stimuli except one. In contrast, we observed large differences in accuracy for the indefinite instructions (F = 6.43, p < .01). The average accuracy for the human voice was 80%, while the accuracy scores for the synthesized voices were 65% for synthesizer 1 and 31% for synthesizer 2. This suggests that the delay in reference resolution for synthesized definite instructions may be due to distributional characteristics of the voices over the course of the interaction. We will return to this issue after examining the results for the indefinite instruction trials.

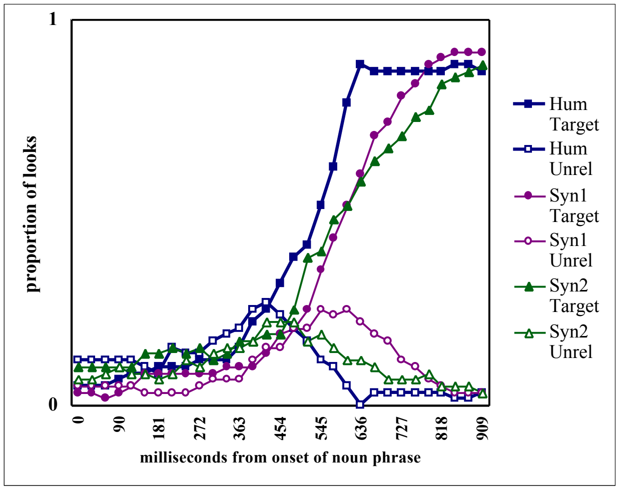

Figure 3: Proportion of looks to target (unique) and unique unrelated items for definite instructions

Now let us consider the trials with instructions containing indefinite articles. Figure 4 shows the average proportion of looks over time to the target (duplicate) items and the unrelated (unique) items for each of the three voice conditions during the trials with indefinite instructions, e.g., *Click on a square*. Note that in the indefinite condition, either of the duplicated items is an appropriate target in response to the spoken instruction. For clarity of presentation, looks to either of the indefinite targets are summed together and represented as a single line for each of the voice conditions in Figure 3. Similarly, looks to either of the unrelated (unique) items are summed together in a single line for each voice. Again, the zero point on the x-axis corresponds to the onset of the noun phrase, e.g., *a square*.

For all voice conditions, looks to the duplicated items diverge from looks to the unique items at roughly the same point. We cannot tell from this data whether these eye movements are due to processing of the indefinite article or whether they are due to processing of the noun. It is surprising that we do not see differences in the time course of looks between the voice conditions, given the differences in accuracy for the voice judgement survey, but an examination of looks to the two duplicated items may provide an explanation.

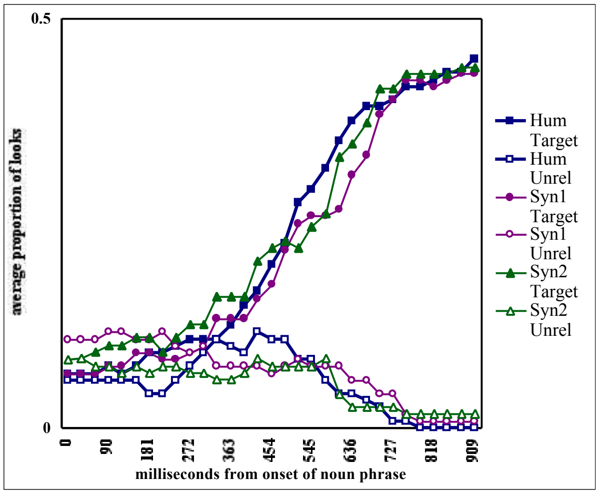

Figure 4: Average proportion of looks to target (duplicate) and unrelated (unique) items for indefinite instructions

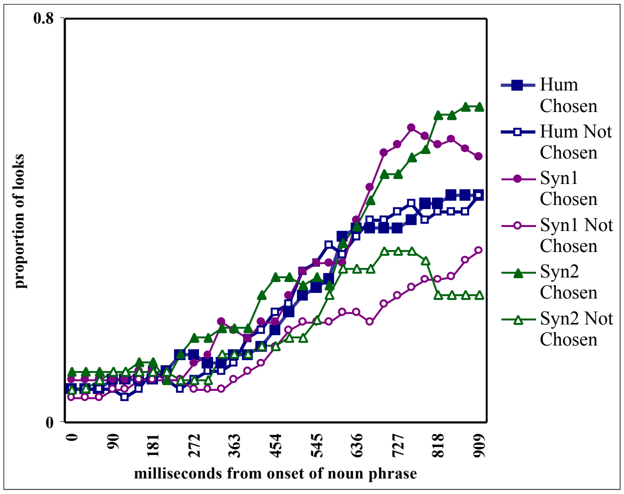

Figure 5: Proportion of looks to the target (duplicate) items chosen and not chosen for indefinite instructions

Figure 5 shows the proportion of looks over time to the two possible target items. For each voice condition the item identified as "chosen" is the duplicated item that the participant eventually clicked on and the item identified as "not chosen" is the other duplicated item.

For instructions in the human voice condition, participants considered each of the duplicated items before clicking on one, reflecting the expected circumscription of referential domain in the indefinite condition. For instructions in the synthesized voice conditions, participants tended to click on the first of the duplicated items that came to their attention, reflecting a more restricted referential domain than expected – one more appropriate to a definite article interpretation.

These results demonstrate that participants make different assumptions about the felicity of the use of the article for the synthesized speech instructions than for the natural speech instructions, due to global differences in how well the indefinite articles could be understood in the three voice conditions. This could also explain the delays in disambiguation for the definite article instructions.

Implications

People process spoken language continuously, even though continuous recognition entails resolving numerous temporary ambiguities on the fly. We have shown that this mode of recognition carries over to speech that is clearly identifiable as computer-generated artificial speech. The results suggest that this paradigm can be used to provide a fine-grained evaluation of comprehension during human-computer dialogue. Specifically, during reference resolution, listeners use cues such as definiteness that are often carried by short unstressed words that are difficult to synthesize. Lack of clarity in synthesizing these words may interfere with reference resolution. While perhaps not a problem in such simple tasks as these, we can expect it to be more problematic in more complex applications, and as the global characteristics of the speech cause interactions with additional error sources, such as unnatural prosodic cues. The important point here, however, is that the eye-tracking technique can reveal even subtle comprehension problems at a fine degree of temporal resolution. This suggests that the same technique could be used to evaluate components such as those affecting lexical choice, sentence structure, intonation and even higher-level discourse intentions. In addition, the eye-tracking paradigm may provide a valuable new method of comprehension evaluation in multimodal language applications using visual displays.

Moreover, our finding that people naturally process synthesized speech incrementally means that computational dialogue-based systems have the potential to be a psycholinguistic tool, especially for experimental questions where it is important to be able to generate utterances on the fly. By using such systems, we could generate more complex stimuli than is possible using a confederate or pre-recorded speech. While there remains much to be done to make this a reality, the range of experiments it would enable is great.

Acknowledgments

This research was supported by NIH HD-27206 to MKT.

References

Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001). Towards conversational human-computer interaction. *AI Magazine*, *22*, 27-35.

Chambers, C.G., Tanenhaus, M.K, Eberhard, K.M., Filip, H. & Carlson, G.N. (in press). Circumscribing referential domains in real-time sentence comprehension. *Journal of Memory and Language*.

Allopenna, P., Magnuson, J., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye-movements: evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419-439.

Altmann, G. (1998). Ambiguity in sentence processing. *Trends in Cognitive Sciences*, *2(4)*, 146-152.

Hanna, J. E. (2001). *The effects of linguistic form, common ground, and perspective on domains of referential interpretation*. Doctoral Dissertation, Department of Brain and Cognitive Sciences, University of Rochester.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition*, *25*, 71-102.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition, 71*, 109-147.

Tanenhaus, M. K., Magnuson, J., & Chambers, C. (forthcoming). Eye-movements and spoken language comprehension: Bridging the language as action and language as product tradition. *Trends in Cognitive Science*.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science*, *268*, 1632-1634.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1996). Using eye-movements to study spoken language comprehension: Evidence for visually-mediated incremental interpretation. In T. Inui & J. McClelland (Eds.), *Attention & Performance XVI: Integration in Perception and Communication*. Cambridge: MIT Press.

Tanenhaus, M. K., & Trueswell, J. (1995). Sentence comprehension. In J. Miller & P. Eimas (Eds.), *Handbook of Perception and Cognition*. San Diego: Academic Press.

Zue, V., Seneff, J., Glass, J., Polifroni, J., Pao, C., Hazen, T., & Hetherington, L. (2000). Jupiter: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, *8(1)*.