What's There to Talk About?
A Multi-Modal Model of Referring Behavior in the Presence of Shared Visual Information

Darren Gergle
Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburg, PA USA
dgergle+cs.cmu.edu

Abstract

This paper describes the development of a rule-based computational model that describes how a feature-based representation of shared visual information combines with linguistic cues to enable effective reference resolution. This work explores a language-only model, a visual-only model, and an integrated model of reference resolution and applies them to a corpus of transcribed task-oriented spoken dialogues. Preliminary results from a corpus-based analysis suggest that integrating information from a shared visual environment can improve the performance and quality of existing discourse-based models of reference resolution.

1 Introduction

In this paper, we present work in progress towards the development of a rule-based computational model to describe how various forms of shared visual information combine with linguistic cues to enable effective reference resolution during task-oriented collaboration.

A number of recent studies have demonstrated that linguistic patterns shift depending on the speaker's situational context. Patterns of proximity markers (e.g., *this/here* vs. *that/there*) change according to whether speakers perceive themselves to be physically co-present or remote from their partner (Byron & Stoia, 2005; Fussell *et al.*, 2004; Levelt, 1989). The use of particular forms of definite referring expressions (e.g., *personal pronouns* vs. *demonstrative pronouns* vs. *demonstrative descriptions*) varies depending on the local visual context in which they are constructed (Byron *et al.*, 2005a). And people are found to use shorter and syntactically simpler language (Oviatt, 1997) and different surface realizations (Cassell & Stone, 2000) when gestures accompany their spoken language.

More specifically, work examining dialogue patterns in collaborative environments has demonstrated that pairs adapt their linguistic patterns based on what they believe their partner can see (Brennan, 2005; Clark & Krych, 2004; Gergle *et al.*, 2004; Kraut *et al.*, 2003). For example, when a speaker knows their partner can see their actions but will incur a small delay before doing so, they increase the proportion of full NPs used (Gergle et al., 2004). Similar work by Byron and colleagues (2005b) demonstrates that the forms of referring expressions vary according to a partner's proximity to visual objects of interest.

Together this work suggests that the interlocutors' shared visual context has a major impact on their patterns of referring behavior. Yet, a number of discourse-based models of reference primarily rely on linguistic information without regard to the surrounding visual environment (e.g., see Brennan *et al.*, 1987; Hobbs, 1978; Poesio *et al.*, 2004; Strube, 1998; Tetreault, 2005). Recently, multi-modal models have emerged that integrate visual information into the resolution process. However, many of these models are restricted by their simplifying assumption of communication via a command language. Thus, their approaches apply to explicit interaction techniques but do not necessarily support more general communication in the presence of shared visual information (e.g., see Chai *et al.*, 2005; Huls *et al.*, 1995; Kehler, 2000).

It is the goal of the work presented in this paper to explore the performance of language-based models of reference resolution in contexts where speakers share a common visual space. In particular, we examine three basic hypotheses regarding the likely impact of linguistic and visual salience on referring behavior. The first hypothesis suggests that visual information is disregarded and that linguistic context provides sufficient information to describe referring behavior. The second hypothesis suggests that visual salience overrides any linguistic salience in governing referring behavior. Finally, the third hypothesis posits that a balance of linguistic and visual salience is needed in order to account for patterns of referring behavior.

In the remainder of this paper, we begin by presenting a brief discussion of the motivation for this work. We then describe three computational models of referring behavior used to explore the hypotheses described above, and the corpus on which they have been evaluated. We conclude by presenting preliminary results and discussing future modeling plans.

2 Motivation

There are several motivating factors for developing a computational model of referring behavior in shared visual contexts. First, a model of referring behavior that integrates a component of shared visual information can be used to increase the robustness of interactive agents that converse with humans in real-world situated environments. Second, such a model can be applied to the development of a range of technologies to support distributed group collaboration and mediated communication. Finally, such a model can be used to provide a deeper theoretical understanding of how humans make use of various forms of shared visual information in their everyday communication.

The development of an integrated multi-modal model of referring behavior can improve the performance of state-of-the-art computational models of communication currently used to support conversational interactions with an intelligent agent (Allen *et al.*, 2005; Devault *et al.*, 2005; Gorniak & Roy, 2004). Many of these models rely on discourse state and prior linguistic contributions to successfully resolve references in a given utterance. However, recent technological advances have created opportunities for human-human and human-agent interactions in a wide variety of contexts that include visual objects of interest. Such systems may benefit from a data-driven model of how collaborative pairs adapt their language in the presence (or absence) of shared visual information. A successful computational model of referring behavior in the presence of visual information could enable agents to emulate many elements of more natural and realistic human conversational behavior.

A computational model may also make valuable contributions to research in the area of computer-mediated communication. Video-mediated communication systems, shared media spaces, and collaborative virtual environments are technologies developed to support joint activities between geographically distributed groups. However, the visual information provided in each of these technologies can vary drastically. The shared field of view can vary, views may be misaligned between speaking partners, and delays of the sort generated by network congestion may unintentionally disrupt critical information required for successful communication (Brennan, 2005; Gergle et al., 2004). Our proposed model could be used along with a detailed task analysis to inform the design and development of such technologies. For instance, the model could inform designers about the times when particular visual elements need to be made more salient in order to support effective communication. A computational model that can account for visual salience and understand its impact on conversational coherence could inform the construction of shared displays or dynamically restructure the environment as the discourse unfolds.

A final motivation for this work is to further our theoretical understanding of the role shared visual information plays during communication. A number of behavioral studies have demonstrated the need for a more detailed theoretical understanding of human referring behavior in the presence of shared visual information. They suggest that shared visual information of the task objects and surrounding workspace can significantly impact collaborative task performance and communication efficiency in task-oriented interactions (Kraut et al., 2003; Monk & Watts, 2000; Nardi *et al.*, 1993; Whittaker, 2003). For example, viewing a partner's actions facilitates monitoring of comprehension and enables efficient object reference (Daly-Jones *et al.*, 1998), changing the amount of available visual information impacts information gathering and recovery from ambiguous help requests (Karsenty, 1999), and varying the field of view that a remote helper has of a co-worker's environment influences performance and shapes communication patterns in directed physical tasks (Fussell *et al.*, 2003). Having a computational description of these processes can provide insight into why they occur, can expose implicit and possibly inadequate simplifying assumptions underlying existing theoretical models, and can serve as a guide for future empirical research.

3 Background and Related Work

A review of the computational linguistics literature reveals a number of discourse models that describe referring behaviors in written, and to a lesser extent, spoken discourse (for a recent review see Tetreault, 2005). These include models based primarily on world knowledge (e.g., Hobbs *et al.*, 1993), syntax-based methods (Hobbs, 1978), and those that integrate a combination of syntax, semantics and discourse structure (e.g., Grosz *et al.*, 1995; Strube, 1998; Tetreault, 2001). The majority of these models are salience-based approaches where entities are ranked according to their grammatical function, number of prior mentions, prosodic markers, etc.

In typical language-based models of reference resolution, the licensed referents are introduced through utterances in the prior linguistic context. Consider the following example drawn from the PUZZLE CORPUS[1] whereby a "Helper" describes to a "Worker" how to construct an arrangement of colored blocks so they match a solution only the Helper has visual access to:

(1) Helper: Take the dark red piece.
Helper: Overlap it over the orange halfway.

In excerpt (1), the first utterance uses the definite-NP "the dark red piece," to introduce a new discourse entity. This phrase specifies an actual puzzle piece that has a color attribute of dark red and that the Helper wants the Worker to position in their workspace. Assuming the Worker has correctly heard the utterance, the Helper can now expect that entity to be a shared element as established by prior linguistic context. As such, this piece can subsequently be referred to using a pronoun. In this case, most models correctly license the observed behavior as the Helper specifies the piece using "it" in the second utterance.

3.1 A Drawback to Language-Only Models

However, as described in Section 2, several behavioral studies of task-oriented collaboration have suggested that visual context plays a critical role in determining which objects are salient parts of a conversation. The following example from the same PUZZLE CORPUS—in this case from a task condition in which the pairs share a visual space—demonstrates that it is not only the linguistic context that determines the potential antecedents for a pronoun, but also the physical context as well:

(2) Helper: Alright, take the dark orange block.
Worker: OK.
Worker: *[moved an incorrect piece]*
Helper: Oh, that's not it.

In excerpt (2), both the linguistic and visual information provide entities that could be co-specified by a subsequent referent. In this excerpt, the first pronoun "that," refers to the "*[incorrect piece]*" that was physically moved into the shared visual workspace but was not previously mentioned. While the second pronoun, "it," has as its antecedent the object co-specified by the definite-NP "the dark orange block." This example demonstrates that during task-oriented collaborations both the linguistic and visual contexts play central roles in enabling the conversational pairs to make efficient use of communication tactics such as pronominalization.

3.2 Towards an Integrated Model

While most computational models of reference resolution accurately resolve the pronoun in excerpt (1), many fail at resolving one or more of the pronouns in excerpt (2). In this rather trivial case, if no method is available to generate potential discourse entities from the shared visual environment, then the model cannot correctly resolve pronouns that have those objects as their antecedents.

This problem is compounded in real-world and computer-mediated environments since the visual information can take many forms. For instance, pairs of interlocutors may have different perspectives which result in different objects being occluded for the speaker and for the listener. In geographically distributed collaborations a conversational partner may only see a subset of the visual space due to a limited field of view provided by a camera. Similarly, the speed of the visual update may be slowed by network congestion.

Byron and colleagues recently performed a preliminary investigation of the role of shared visual information in a task-oriented, human-to-human collaborative virtual environment (Byron et al., 2005b). They compared the results of a language-only model with a visual-only model, and developed a visual salience algorithm to rank the visual objects according to recency, exposure time, and visual uniqueness. In a hand-processed evaluation, they found that a visual-only model accounted for 31.3% of the referring expressions, and that adding semantic restrictions (e.g., "open

[1] The details of the PUZZLE CORPUS are described in §.4.

that" could only match objects that could be opened, such as a door) increased performance to 52.2%. These values can be compared with a language-only model with semantic constraints that accounted for 58.2% of the referring expressions.

While Byron's visual-only model uses semantic selection restrictions to limit the number of visible entities that can be referenced, her model differs from the work reported here in that it does not make simultaneous use of linguistic salience information based on the discourse content. So, for example, referring expressions cannot be resolved to entities that have been mentioned but which are not visible. Furthermore, all other things equal, it will not correctly resolve references to objects that are most salient based on the linguistic context over the visual context. Therefore, in addition to language-only and visual-only models, we explore the development of an integrated model that uses both linguistic and visual salience to support reference resolution. We also extend these models to a new task domain that can elaborate on referential patterns in the presence of various forms of shared visual information. Finally, we make use of a corpus gathered from laboratory studies that allow us to decompose the various features of shared visual information in order to better understand their independent effects on referring behaviors.

The following section provides an overview of the task paradigm used to collect the data for our corpus evaluation. We describe the basic experimental paradigm and detail how it can be used to examine the impact of various features of a shared visual space on communication.

4 The Puzzle Task Corpus

The corpus data used for the development of the models in this paper come from a subset of data collected over the past few years using a referential communication task called the puzzle study (Gergle et al., 2004).

In this task, pairs of participants are randomly assigned to play the role of "Helper" or "Worker." It is the goal of the task for the Helper to successfully describe a configuration of pieces to the Worker, and for the Worker to correctly arrange the pieces in their workspace. The puzzle solutions, which are only provided to the Helper, consist of four blocks selected from a larger set of eight. The goal is to have the Worker correctly place the four solution pieces in the proper configuration as quickly as possible so that they match the target solution the Helper is viewing.

Each participant was seated in a separate room in front of a computer with a 21-inch display. The pairs communicated over a high-quality, full-duplex audio link with no delay. The experimental displays for the Worker and Helper are illustrated in Figure 1.

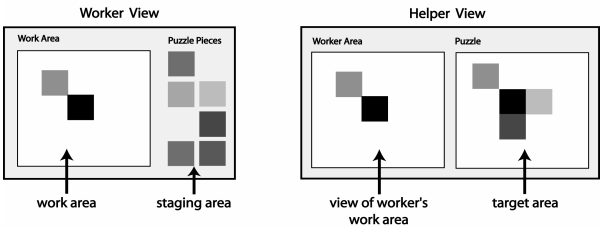

Figure 1. The Worker's view (left) and the Helper's view (right).

The Worker's screen (left) consists of a staging area on the right hand side where the puzzle pieces are held, and a work area on the left hand side where the puzzle is constructed. The Helper's screen (right) shows the target solution on the right, and a view of the Worker's work area in the left hand panel. The advantage of this setup is that it allows exploration of a number of different arrangements of the shared visual space. For instance, we have varied the proportion of the workspace that is visually shared with the Helper in order to examine the impact of a limited field-of-view. We have offset the spatial alignment between the two displays to simulate settings of various video systems. And we have added delays to the speed with which the Helper receives visual feedback of the Worker's actions in order to simulate network congestion.

Together, the data collected using the puzzle paradigm currently contains 64,430 words in the form of 10,640 contributions collected from over 100 different pairs. Preliminary estimates suggest that these data include a rich collection of over 5,500 referring expressions that were generated across a wide range of visual settings. In this paper, we examine a small portion of the data in order to assess the feasibility and potential contribution of the corpus for model development.

4.1 Preliminary Corpus Overview

The data collected using this paradigm includes an audio capture of the spoken conversation surrounding the task, written transcriptions of the spoken utterances, and a time-stamped record of all the piece movements and their representative state in the shared workspace (e.g., whether they are visible to both the Helper and Worker). From these various streams of data we can parse and extract the units for inclusion in our models.

For initial model development, we focus on modeling two primary conditions from the PUZZLE CORPUS. The first is the "*No Shared Visual Information*" condition where the Helper could not see the Worker's workspace at all. In this condition, the pair needs to successfully complete the tasks using only linguistic information. The second is the "*Shared Visual Information*" condition, where the Helper receives immediate visual feedback about the state of the Worker's work area. In this case, the pairs can make use of both linguistic information and shared visual information in order to successfully complete the task.

As Table 1 demonstrates, we use a small random selection of data consisting of 10 dialogues from each of the *Shared Visual Information* and *No Shared Visual Information* conditions. Each of these dialogues was collected from a unique participant pair. For this evaluation, we focused primarily on pronoun usage since this has been suggested to be one of the major linguistic efficiencies gained when pairs have access to a shared visual space (Kraut et al., 2003).

Task Condition	**Corpus Statistics**			
	Dialogues	*Contri-butions*	*Words*	*Pro-nouns*
No Shared Visual Information	10	218	1181	30
Shared Visual Information	10	174	938	39
Total	20	392	2119	69

Table 1. Overview of the data used.

5 Preliminary Model Overviews

The models evaluated in this paper are based on Centering Theory (Grosz et al., 1995; Grosz & Sidner, 1986) and the algorithms devised by Brennan and colleagues (1987) and adapted by Tetreault (2001). We examine a language-only model based on Tetreault's Left-Right Centering (LRC) model, a visual-only model that uses a measure of visual salience to rank the objects in the visual field as possible referential anchors, and an integrated model that balances the visual information along with the linguistic information to generate a ranked list of possible anchors.

5.1 The Language-Only Model

We chose the LRC algorithm (Tetreault, 2001) to serve as the basis for our language-only model. It has been shown to fare well on task-oriented spoken dialogues (Tetreault, 2005) and was easily adapted to the PUZZLE CORPUS data.

LRC uses grammatical function as a central mechanism for resolving the antecedents of anaphoric references. It resolves referents by first searching in a left-to-right fashion within the current utterance for possible antecedents. It then makes co-specification links when it finds an antecedent that adheres to the selectional restrictions based on verb argument structure and agreement in terms of number and gender. If a match is not found the algorithm then searches the lists of possible antecedents in prior utterances in a similar fashion.

The primary structure employed in the language-only model is a ranked entity list sorted by linguistic salience. To conserve space we do not reproduce the LRC algorithm in this paper and instead refer readers to Tetreault's original formulation (2001). We determined order based on the following precedence ranking:

$$\textit{Subject} \prec \textit{Direct Object} \prec \textit{Indirect Object}$$

Any remaining ties (e.g., an utterance with two direct objects) were resolved according to a left-to-right breadth-first traversal of the parse tree.

5.2 The Visual-Only Model

As the Worker moves pieces into their workspace, depending on whether or not the workspace is shared with the Helper, the objects become available for the Helper to see. The visual-only model utilized an approach based on visual salience. This method captures the relevant visual objects in the puzzle task and ranks them according to the recency with which they were *active* (as described below).

Given the highly controlled visual environment that makes up the PUZZLE CORPUS, we have complete access to the visual pieces and exact timing information about when they become visible, are moved, or are removed from the shared workspace. In the visual-only model, we maintain an ordered list of entities that comprise the shared visual space. The entities are included in the list if they are currently visible to both the Helper and Worker, and then ranked according to the recency of their activation.[2]

[2] This allows for objects to be dynamically rearranged depending on when they were last 'touched' by the Worker.

5.3 The Integrated Model

We used the salience list generated from the language-only model and integrated it with the one from the visual-only model. The method of ordering the integrated list resulted from general perceptual psychology principles that suggest that highly active visual objects attract an individual's attentional processes (Scholl, 2001).

In this preliminary implementation, we defined *active objects* as those objects that had recently moved within the shared workspace. These objects are added to the top of the linguistic-salience list which essentially rendered them as the focus of the joint activity. However, people's attention to static objects has a tendency to fade away over time. Following prior work that demonstrated the utility of a visual decay function (Byron et al., 2005b; Huls et al., 1995), we implemented a three second threshold on the lifespan of a visual entity. From the time since the object was last active, it remained on the list for three seconds. After the time expired, the object was removed and the list returned to its prior state. This mechanism was intended to capture the notion that active objects are at the center of shared attention in a collaborative task for a short period of time. After that the interlocutors revert to their recent linguistic history for the context of an interaction.

It should be noted that this is work in progress and a major avenue for future work is the development of a more theoretically grounded method for integrating linguistic salience information with visual salience information.

5.4 Evaluation Plan

Together, the models described above allow us to test three basic hypotheses regarding the likely impact of linguistic and visual salience:

Purely linguistic context. One hypothesis is that the visual information is completely disregarded and the entities are salient purely based on linguistic information. While our prior work has suggested this should not be the case, several existing computational models function only at this level.

Purely visual context. A second possibility is that the visual information completely overrides linguistic salience. Thus, visual information dominates the discourse structure when it is available and relegates linguistic information to a subordinate role. This too should be unlikely given the fact that not all discourse deals with external elements from the surrounding world.

A balance of syntactic and visual context. A third hypothesis is that both linguistic entities and visual entities are required in order to accurately and perspicuously account for patterns of observed referring behavior. Salient discourse entities result from some balance of linguistic salience and visual salience.

6 Preliminary Results

In order to investigate the hypotheses described above, we examined the performance of the models using hand-processed evaluations of the PUZZLE CORPUS data. The following presents the results of the three different models on 10 trials of the PUZZLE CORPUS in which the pairs had no shared visual space, and 10 trials from when the pairs had access to shared visual information representing the workspace. Two experts performed qualitative coding of the referential anchors for each pronoun in the corpus with an overall agreement of 88% (the remaining anomalies were resolved after discussion).

As demonstrated in Table 2, the language-only model correctly resolved 70% of the referring expressions when applied to the set of dialogues where only language could be used to solve the task (i.e., the no shared visual information condition). However, when the same model was applied to the dialogues from the task conditions where shared visual information was available, it only resolved 41% of the referring expressions correctly. This difference was significant, $\chi^2(1, N=69) = 5.72, p = .02$.

	No Shared Visual Information	**Shared Visual Information**
Language Model	70.0% (21 / 30)	41.0% (16 / 39)
Visual Model	n/a	66.7% (26 / 39)
Integrated Model	70.0% (21 / 30)	69.2% (27 / 39)

Table 2. Results for all pronouns in the subset of the PUZZLE CORPUS evaluated.

In contrast, when the visual-only model was applied to the same data derived from the task conditions in which the shared visual information was available, the algorithm correctly resolved 66.7% of the referring expressions. In comparison to the 41% produced by the language-only model. This difference was also significant, $\chi^2(1, N=78) = 5.16, p = .02$. However, we did not find evidence of a difference between the performance of the visual-only model on the visual task conditions and the language-only model on the language task conditions, $\chi^2(1, N=69) = .087$, p = .77 (*n.s.*).

The integrated model with the decay function also performed reasonably well. When the integrated model was evaluated on the data where only language could be used it effectively reverts back to a language-only model, therefore achieving the same 70% performance. Yet, when it was applied to the data from the cases when the pairs had access to the shared visual information it correctly resolved 69.2% of the referring expressions. This was also better than the 41% exhibited by the language-only model, $\chi^2(1, N=78) = 6.27$, p = .012; however, it did not statistically outperform the visual-only model on the same data, $\chi^2(1, N=78) = .059$, p = .81 (*n.s.*).

In general, we found that the language-only model performed reasonably well on the dialogues in which the pairs had no access to shared visual information. However, when the same model was applied to the dialogues collected from task conditions where the pairs had access to shared visual information the performance of the language-only model was significantly reduced. However, both the visual-only model and the integrated model significantly increased performance. The goal of our current work is to find a better integrated model that can achieve significantly better performance than the visual-only model. As a starting point for this investigation, we present an error analysis below.

6.1 Error Analysis

In order to inform further development of the model, we examined a number of failure cases with the existing data. The first thing to note was that a number of the pronouns used by the pairs referred to larger visible structures in the workspace. For example, the Worker would sometimes state, “like this?”, and ask the Helper to comment on the overall configuration of the puzzle. Table 3 presents the performance results of the models after removing all expressions that did not refer to pieces of the puzzle.

	No Shared Visual Information	**Shared Visual Information**
Language Model	77.7% (21 / 27)	47.0% (16 / 34)
Visual Model	n/a	76.4% (26 / 34)
Integrated Model	77.7% (21 / 27)	79.4% (27 / 34)

Table 3. Model performance results when restricted to piece referents.

In the errors that remained, the language-only model had a tendency to suffer from a number of higher-order referents such as events and actions. In addition, there were several errors that resulted from chaining errors where the initial referent was misidentified. As a result, all subsequent chains of referents were incorrect.

The visual-only model and the integrated model had a tendency to suffer from timing issues. For instance, the pairs occasionally introduced a new visual entity with, “this one?” However, the piece did not appear in the workspace until a short time *after* the utterance was made. In such cases, the object was not available as a referent on the object list. In the future we plan to investigate the temporal alignment between the visual and linguistic streams.

In other cases, problems simply resulted from the unique behaviors present when exploring human activities. Take the following example,

(3) Helper: There is an orange red that obscures half of it and it is to the left of it

In this excerpt, all of our models had trouble correctly resolving the pronouns in the utterance. However, while this counts as a strike against the model performance, the model actually presented a true account of human behavior. While the model was confused, so was the Worker. In this case, it took three more contributions from the Helper to unravel what was actually intended.

7 Future Work

In the future, we plan to extend this work in several ways. First, we plan future studies to help expand our notion of visual salience. Each of the visual entities has an associated number of domain-dependent features. For example, they may have appearance features that contribute to overall salience, become activated multiple times in a short window of time, or be more or less salient depending on nearby visual objects. We intend to explore these parameters in detail.

Second, we plan to appreciably enhance the integrated model. It appears from both our initial data analysis, as well as our qualitative examination of the data, that the pairs make tradeoffs between relying on the linguistic context and the visual context. Our current instantiation of the integrated model could be enhanced by taking a more theoretical approach to integrating the information from multiple streams.

Finally, we plan to perform a large-scale computational evaluation of the entire PUZZLE CORPUS in order to examine a much wider range of visual features such as limited field-of-views, delays in providing the shared visual information, and various asymmetries in the interlocutors' visual information. In addition to this we plan to extend our model to a wider range of task domains in order to explore the generality of its predictions.

Acknowledgments

This research was funded in by an IBM Ph.D. Fellowship. I would like to thank Carolyn Rosé and Bob Kraut for their support.

References

Allen, J., Ferguson, G., Swift, M., Stent, A., Stoness, S., Galescu, L., et al. (2005). Two diverse systems built using generic components for spoken dialogue. In Proceedings of *Association for Computational Linguistics, Companion Vol.*, pp. 85-88.

Brennan, S. E. (2005). How conversation is shaped by visual and spoken evidence. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 95-129). Cambridge, MA: MIT Press.

Brennan, S. E., Friedman, M. W., & Pollard, C. J. (1987). A centering approach to pronouns. In Proceedings of *25th Annual Meeting of the Association for Computational Linguistics*, pp. 155-162.

Byron, D. K., Dalwani, A., Gerritsen, R., Keck, M., Mampilly, T., Sharma, V., et al. (2005a). Natural noun phrase variation for interactive characters. In Proceedings of *1st Annual Artificial Intelligence and Interactive Digital Entertainment Conference*, pp. 15-20. AAAI.

Byron, D. K., Mampilly, T., Sharma, V., & Xu, T. (2005b). Utilizing visual attention for cross-modal coreference interpretation. In Proceedings of *Fifth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-05)*, pp.

Byron, D. K., & Stoia, L. (2005). An analysis of proximity markers in collaborative dialog. In Proceedings of *41st annual meeting of the Chicago Linguistic Society*, pp. Chicago Linguistic Society.

Cassell, J., & Stone, M. (2000). Coordination and context-dependence in the generation of embodied conversation. In Proceedings of *International Natural Language Generation Conference*, pp. 171-178.

Chai, J. Y., Prasov, Z., Blaim, J., & Jin, R. (2005). Linguistic theories in efficient multimodal reference resolution: An empirical investigation. In Proceedings of *Intelligent User Interfaces*, pp. 43-50. NY: ACM Press.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory & Language, 50*(1), 62-81.

Daly-Jones, O., Monk, A., & Watts, L. (1998). Some advantages of video conferencing over high-quality audio conferencing: Fluency and awareness of attentional focus. *International Journal of Human-Computer Studies, 49*, 21-58.

Devault, D., Kariaeva, N., Kothari, A., Oved, I., & Stone, M. (2005). An information-state approach to collaborative reference. In Proceedings of *Association for Computational Linguistics, Companion Vol.*, pp.

Fussell, S. R., Setlock, L. D., & Kraut, R. E. (2003). Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. In Proceedings of *Human Factors in Computing Systems (CHI '03)*, pp. 513-520. ACM Press.

Fussell, S. R., Setlock, L. D., Yang, J., Ou, J., Mauer, E. M., & Kramer, A. (2004). Gestures over video streams to support remote collaboration on physical tasks. *Human-Computer Interaction, 19*, 273-309.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2004). Language efficiency and visual technology: Minimizing collaborative effort with visual information. *Journal of Language & Social Psychology, 23*(4), 491-517.

Gorniak, P., & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research, 21*, 429-470.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics, 21*(2), 203-225.

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions and the structure of discourse. *Computational Linguistics, 12*(3), 175-204.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua, 44*, 311-338.

Hobbs, J. R., Stickel, M. E., Appelt, D. E., & Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence, 63*, 69-142.

Huls, C., Bos, E., & Claassen, W. (1995). Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics, 21*(1), 59-79.

Karsenty, L. (1999). Cooperative work and shared context: An empirical study of comprehension problems in side by side and remote help dialogues. *Human-Computer Interaction, 14*(3), 283-315.

Kehler, A. (2000). Cognitive status and form of reference in multimodal human-computer interaction. In Proceedings of *American Association for Artificial Intelligence (AAAI 2000)*, pp. 685-689.

Kraut, R. E., Fussell, S. R., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human Computer Interaction, 18*, 13-49.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Monk, A., & Watts, L. (2000). Peripheral participation in video-mediated communication. *International Journal of Human-Computer Studies, 52*(5), 933-958.

Nardi, B., Schwartz, H., Kuchinsky, A., Leichner, R., Whittaker, S., & Sclabassi, R. T. (1993). Turning away from talking heads: The use of video-as-data in neurosurgery. In Proceedings of *Interchi '93*, pp. 327-334.

Oviatt, S. L. (1997). Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction, 12*, 93-129.

Poesio, M., Stevenson, R., Di Eugenio, B., & Hitzeman, J. (2004). Centering: A parametric theory and its instantiations. *Computational Linguistics, 30*(3), 309-363.

Scholl, B. J. (2001). Objects and attention: the state of the art. *Cognition, 80*, 1-46.

Strube, M. (1998). Never look back: An alternative to centering. In Proceedings of *36th Annual Meeting of the Association for Computational Linguistics*, pp. 1251-1257.

Tetreault, J. R. (2001). A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics, 27*(4), 507-520.

Tetreault, J. R. (2005). *Empirical evaluations of pronoun resolution.* Unpublished doctoral thesis, University of Rochester, Rochester, NY.

Whittaker, S. (2003). Things to talk about when talking about things. *Human-Computer Interaction, 18*, 149-170.